Natural Language Processing Methods for Attitudinal Near-Synonymy

Mary Elizebeth Gardiner
Bachelor of Science (Honours), Macquarie University, Sydney
Bachelor of Arts, University of Sydney

This dissertation is presented for the degree of

Doctor of Philosophy

at

March 2013

Declaration

The research presented in this thesis is the original work of the author except where otherwise indicated. This work has not been submitted for a degree or any other qualification to any other university or institution. All verbatim extracts have been distinguished by quotations, and all sources of information have been specifically acknowledged.

Some parts of this thesis include early or revised versions of published papers:

- Gardiner, Mary and Mark Dras (2007a). Corpus statistics approaches to discriminating among near-synonyms. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007)*, pages 31–39. Melbourne, Australia
- Gardiner, Mary and Mark Dras (2007b). Exploring approaches to discriminating among near-synonyms. In *Proceedings of the Australasian Language Technology Workshop 2007*, pages 31–39. Melbourne, Australia. URL http://www.aclweb.org/anthology/U07-1007
- Gardiner, Mary and Mark Dras (2012). Valence shifting: Is it a valid task? In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 42–51. Dunedin, New Zealand. URL http://www.aclweb.org/anthology/U/U12/U12-1007

Additional papers co-authored by the author of this thesis are referred to but not incorporated into the text of this thesis:

- Hawker, Tobias; Mary Gardiner; and Andrew Bennetts (2007). Practical queries of a massive n-gram database. In *Proceedings of the Australasian Language Technology Workshop 2007*, pages 40–48. Melbourne, Australia. URL http://www.aclweb.org/anthology/U07-1008
- Dras, Mark; Debbie Richards; Meredith Taylor; and Mary Gardiner (2010a). Deceptive agents and language. In *Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2010)*. Toronto, Canada
- Dras, Mark; Debbie Richards; Meredith Taylor; and Mary Gardiner (2010b). Deceptive agents and language. In *Proceedings of the International Workshop on Interacting with ECAs as Virtual Characters*. Toronto, Canada

The research presented in Section 5.3 of this thesis, using human subjects, was approved by the Macquarie University Ethics Review Committee as per Appendix G.

Mary Elizebeth Gardiner

Date:

Abstract

When either a human author or a computer natural language generation system tries to express an idea, there is usually more than one way to say it. This is a problem both for systems that process language, such as systems that recognise textual entailment, which must detect when two surface forms express the same idea; and for systems that generate language, which must choose the most appropriate way to express an idea from a potentially large number of surface forms.

For a natural language generation system, for a given meaning there may be multiple words that could be chosen to express it, or multiple phrases that express the same idea. However, it has also been argued that there are no true synonyms, that even words that have very similar meanings cannot be substituted for each other in all circumstances. Automatic natural language generation systems therefore have a use for modules which make effective word and phrase choices among closely related alternatives.

In this thesis we consider the specific problem of choosing an appropriate word or phrase where the alternatives are closely related in meaning but differ in sentiment or attitude. One example is *stingy* and *frugal*, one of which is critical of what it describes and the other of which is complimentary.

The thesis will address three aspects of the problem. The first question is whether existing methods to predict word choice among closely related words are sufficient for choosing between words that differ in sentiment. There are several methods in the literature for this, relying on statistical models of words in context. The early, relatively poor performance of these methods had been used to argue that statistical methods are not suitable for this task, but later successes with statistical approaches suggest that sufficient amounts of data make it approachable. Using a comprehensive set of data for this thesis, we show that sets of words that differ in sentiment behave in a distinct fashion, suggesting that they are particularly amenable to statistical approaches.

The second aspect of our research into choosing between related words or phrases that differ in sentiment is investigating whether or not including some global information about the entire text is useful in predicting word choice. We hypothesise that information about the sentiment of a document as a whole (for example, if the document is a movie review, whether it is favourable or not) will assist in choosing between closely related words that differ in sentiment. We demonstrate several models improving prediction of the correct word in context, incorporating information from the entire document, the most successful of which are metrics which account for distance from the target word.

The third aspect is an investigation into human perceptions of word choice in a particular generation task — valence shifting — with the goal of changing an existing text so that it is similar in meaning, but more negative in tone. Existing work, which includes using hand-crafted vocabularies annotated with sentiment data, and corpus-derived cues, has found this to be a difficult problem. This work investigates both the success of establishing a more negative tone, and the resulting fluency of the text by asking human judges to evaluate both aspects of the text then explores possible metrics that can predict negativity for use in valence shifting.

To Andrew Bennetts, with thanks for his love and patience.

Acknowledgements

This work was supported by the Australian Research Council's Discovery Grant DP0558852 from 2006–2008 and an Australian Postgraduate Award from 2006–2009.

My supervisor Mark Dras is responsible for much of the success of my PhD project. His friendship and guidance to me over the last ten years of my membership of the Centre for Language Technology at Macquarie University has been very important to me. I will miss working with him and I wish him many happy (and ideally speedier) PhD students in his future.

Thank you to my examiners Lawrence Cavedon, Diana Inkpen and Ali Knott for their pain-staking review of this thesis and their many suggestions for making it a better and more thorough piece of work. I am also grateful to other researchers who I had useful conversations with during this project, including James Curran, Robert Dale, Dominique Estival, Mark Johnson and Rada Mihalcea. I am grateful also to the people and research culture of the CLT as a whole.

A crucial part of doing a PhD in the Centre for Language Technology, especially with Mark, is the support of one's fellows. Thank you to my many postgraduate "siblings" under Mark's supervision over the years: Elena Akhmatova, Suzy Howlett, Teresa Lynn, Pawel Mazur, Yasaman Motazedi, Jean-Philippe Prost, Stephen Wan, Jojo Wong and Simon Zwarts together with our CLT kin Benjamin Börschinger, Andrew Lampert, Chris Rauchle and Jette Viethen. The more distant fellowship of our long-lost cousins at the University of Sydney—Jeremy Fletcher, Toby Hawker, Matt Honnibal, and Tara McIntosh—was also appreciated.

My personal support team's size is only exceeded by my gratitude to them. Thank you to my parents Jill and David Gardiner, who request special mention of the number of software installs and occasional hardware configurations of theirs that I have either "broken" or "improved" (near-synonyms in this case) over the course of about 25 years. Julia Gardiner, Stephanie Gardiner and Joel Connolly contributed proof-reading, babysitting and husband-sitting at various times. The staff and 2011–2013 boards of directors of the Ada Initiative—Valerie Aurora, Rachel Chalmers, Sue Gardner, Denise Paolucci, Caroline Simard and Matt Zimmerman—contributed greatly to the final year of my work with ongoing support and generosity with leave. My kind Dreamwidth cheerleaders got me through April/May 2012 and February/March 2013, thank you to those not mentioned elsewhere: Anna, Chally, Dorothea, Raven, Terri, Tiferet and Tim.

It is impossible to overstate how grateful I am to my husband Andrew Bennetts for his continual emotional and other support while I undertake degrees, found businesses and otherwise contribute to our household's drama quotient. Our son Vincent Gardiner (who arrived in the course of Chapter 4) in turn contributes cuddles, laundry piles and distraction to the whole enterprise.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The problem of Natural Language Generation (NLG) in general is the process of automatically producing natural (human) language via a computational process for some purpose. There are many examples of NLG system goals including:

communicating a piece of information such as the bilingual weather reporting system of Goldberg et al. (1994), who generate forecast descriptions from computational models;

determining a users' needs and then communicating information such as a back-and-forth conversation between the user and the system about a medical diagnosis, as in the virtual nurse project described by Bickmore et al. (2009), allowing patients with low health literacy to receive detailed personalised information;

assisting a user in producing text such as by having them input their communicative goal or determining it from their existing speech or writing, and providing improved phrasings or word choices, such as the help with synonym choice and simile choice that the system of Liu et al. (2011) gives to love letter authors; and

producing creative or artistic works such as humour or poetry with little or no human input, for example the pun generation system of Binsted et al. (1997).

The goal of the *creator* of an NLG system may vary from a straightforward need to have an agent perform the system's goal, e.g. to have an automatic system provide medical information for human resourcing or accuracy reasons; to a desire to investigate the linguistic phenomenon in question, for example, to analyse the basis of humour by

attempting to identify patterns in humour and then see if subjects respond similarly to automatically generated humour as they do to human-created humour.

There are in turn a multitude of sub-tasks involved in natural language generation, including determining the goals of the system, determining an appropriate representation of the system's state, determining how to translate that state into natural language with appropriate syntax, semantics and pragmatics (and, in the case of speech, phonetics and prosody).

One such task is the task of Lexical Choice, that is, choosing the individual words that the system will output. This in turn has several subcomponents:

selecting words with correct meanings where the system wants to refer to a place where goods are purchased, the system probably chooses *market* or *supermarket* and not *bank* or *camel*;

selecting words with correct implications where the system wants to, for example, refer to a law enforcement official in a neutral context, it should choose *police officer* rather than the derogatory *pig*; and

selecting words that fit the surrounding text where there is a strong tendency towards certain words appearing or not appearing together, the system may want to, for example, choose *difficult task* or *difficult job* over *difficult duty* all other things being equal.

In this thesis, we investigate this problem of lexical choice, particularly with reference to the sentiment of the word. Consider the following example sentence, from a movie review:

(1.1) Thomas Bo Larses is particularly good as the younger son overflowing with toxic *anger*.

Consider how the meaning of example (1.1) changes if it is rewritten as any of the following:

(1.2) Thomas Bo Larses is particularly good as the younger son overflowing with toxic *rage*.

(1.3) Thomas Bo Larses is particularly good as the younger son overflowing with toxic *irritation*.

(1.4) Thomas Bo Larses is particularly good as the younger son overflowing with toxic *annoyance*.

All of these choices have certain effects on the sentence. There is the issue of emotional strength in which *anger* and *rage* describe strong negative emotions, *irritation* a lesser one, and *annoyance* a mild one, there are also issues of fluency, particularly the collocation of these words with *toxic*, which is itself strongly negative. A lexical choice system needs to consider all these aspects of meaning when making lexical choices.

1.1 Choosing among near synonyms

The first specific lexical choice problem addressed by this thesis is that of the NEAR SYNONYM CHOICE PROBLEM. NEAR SYNONYMS are words which are very close in meaning, although they may differ somewhat in their denotations or connotations. For example, in the case of *wood* and *forest*, both refer to a grouping of trees, but the former refers to a smaller grouping of trees than the latter. In the case of *slim* and *skinny* the latter is more negative or derogatory in tone. In the case of *attorney* and *solicitor* the difference is regional, something like the situation of words that are translations of each other in different natural languages.

It has been argued that in fact there is no such thing as true synonyms, where the two words are completely substitutable for each other in any context (at least in a given sense), that instead any two words will differ in at the very least their selectional preferences. In any event, in this work we are interested in near synonyms with a clear difference in one aspect of meaning, that of sentiment (sometimes polarity), as in *slim* and *skinny* above. In communicating, the speaker will usually find it important to convey the appropriate sentiment at the appropriate strength, that is, to avoid being negative where a positive sentiment was meant, and also to avoid being lukewarm where a strong sentiment was intended.

This problem is clearly applicable to natural language generation, especially to applications that generate polarised text or which wish to present an accurate summary of such and in the case of dialogue agents, an agent expressing a sentiment of inappropriate polarity or strength to the context may prove to be actively destructive to its conversational goal.

Thus, the first problem considered by this thesis is that of choosing the correct near synonym in the event where near synonyms carry, and differ in, polarity.

Specifically, we investigate the FILL IN THE BLANKS (FITB) task, where a system attempts to re-predict an author's original choice of word, for example, choosing between *error*, *mistake* and *oversight* in this example given by Edmonds (1997):

(1.5) However, such a move also would run the risk of cutting deeply into U.S. economic growth, which is why some economists think it would be a big {*error* | *mistake* | *oversight*}.

This serves as one possible proxy for best near synonym choice, in aggregate, since in general the word human authors choose most often in any given context ought to be the word that systems choose too.

The near-synonym choice problem and the FITB task has a comparatively long history. Initial indications were that it was not particularly amenable to statistical techniques (Edmonds, 1997, 1999) that attempt to predict near synonyms primarily based on using their context as features, however, in the last five years there has been markedly increasing success with various statistical techniques.

However, we observe that in measuring this success, little account has been taken of whether statistical techniques are performing equally well when taking into account what *kind* of near synonyms they do best on. That is, near synonyms can differ in denotational meaning, formality, sentiment or sentiment strength, and similar axes. But the evaluation of statistical approaches to the FITB task have largely assumed that performance on any set of near synonyms is independent of these axes.

Our specific intuition is that, in the case of near synonyms with affective meaning, the choice of near synonym may not rely solely on local context as many statistical approaches have found best, but may be able to be better predicted by relying on cues from the entire document. Our reasoning is first that, at least in many opinionated documents, especially ones intended to summarise a point of view or to be persuasive, maintaining a consistent sentiment is one of the foremost principles of coherence, likely to affect near-synonym choice as well as other linguistic choices; and that work in the domain of sentiment analysis has found that features from the entire document have proved useful in correctly predicting the sentiment of that document.

We thus concentrate on the axis of sentiment difference, and provide evidence that near synonyms that differ in sentiment do behave differently in the FITB task when approached using statistical techniques. Specifically we evaluate the comparative performance of near-synonym sets that differ in sentiment and those that do not, both on existing approaches to FITB developed by Edmonds (1997) and Inkpen (2007b), and on a new approach utilising a Support Vector Machine based method which uses word presence features weighted by their distance from the near synonym they are being used to predict.

1.2 Valence shifting

VALENCE SHIFTING is the problem of transforming a text into one that expresses a different sentiment, either in strength or in polarity, while otherwise preserving the meaning of the text as far as is possible. This can be considered as a special case of PARAPHRASE GENERATION, the problem of creating two texts with different surface expressions but close or identical meanings.

Automatic valence shifting has applications in natural language generation and in assisted writing tools. For example, an automatic thesaurus such as that suggested by Inkpen and Hirst (2006); Inkpen (2007b) might suggest a word choice that allows a text to keep more closely to the tone of the surrounding document, or a summarisation tool may select a sentence to incorporate into its summary but alter its expression in order to align its sentiment strength more closely with the summary rather than the original document.

This task should be possible, but existing approaches to valence shifting have had very mixed results when presented to human judges who are tasked specifically with choosing whether the original or the shifted sentence has greater sentiment in the hypothesised direction. Existing approaches have ranged from lexical substitution like that discussed here to full paraphrasing: given the existing approaches have to date not reported strong evaluations we return to lexical substitution as a proof of concept.

Thus, this thesis thus first asks the question: is a simple version of valence shifting at all effective? Intuitively, it should be the case, at least allowing for difficulties like negation, that replacing *evil* with *bad* renders a sentence such as examples (1.6) and (1.7) somewhat less negative:

(1.6) Movies, like literature, have always been fascinated with twins—especially when one sibling is good and the other is *evil*.

(1.7) Movies, like literature, have always been fascinated with twins—especially when one sibling is good and the other is *bad*.

Thus, our first hypothesis is that, as above, in general replacing a word in a sentence with a word that has less negativity renders the sentence less negative. This is confirmed by an experiment with human subjects.

The second problem considered by this thesis is how to predict the ability of a lexical substitution to alter the perceived sentiment strength of the resulting sentence. We hypothesise that a metric to predict the ability of words to shift the valence of a sen-

tence could be derived from their distribution among documents with different sentiment. We provide one such measure which appears to be useful for building a model of human judgements.

1.3 Contributions of this thesis

Work described in this thesis makes several contributions to the research into lexical choice, particularly with reference to sentiment.

1. This thesis shows that the statistical techniques used to choose among near synonyms respond noticeably differently when there are sentiment differences between the near synonyms, in a way that is not typically considered when evaluating such systems (Chapter 3).

2. This thesis introduces new annotated datasets for the evaluation of the FITB task on near synonyms differing in sentiment (Chapter 3, Chapter 4, Appendix A, Appendix B).

3. This thesis investigates several different feature sets which may improve performance on the FITB task described above (Chapter 4).

4. This thesis develops a new statistical approach to the FITB task, the first such which achieves any success with features other than those immediately surrounding the missing near synonym (Chapter 4).

5. This thesis shows that the new statistical approach responds comparatively well when there are sentiment differences among the near synonyms (Chapter 4).

6. This thesis shows, based on human subject judgements, that the choice of a single lexical item can make a significant difference in the perceived negativity of a sentence, providing support for lexical choice as a crucial aspect of valence shifting (Chapter 5).

7. This thesis proposes techniques for automatically determining the valence-shifting capabilities of a near synonym based on that near synonym's distribution among sentiment charged texts (Chapter 5).

1.4 Thesis outline

This thesis is divided into six chapters. The six chapters following this one first explore further the scope of the problem and then investigate three aspects of it.

In Chapter 2 of this thesis we describe existing literature on the near-synonymity choice problem, with related work in identifying and producing text with similar meanings including paraphrasing, entailment, and word sense disambiguation; we describe existing literature on the both detecting the sentiment of text (known as SENTIMENT ANALYSIS) and of changing it (known as VALENCE SHIFTING); and we discuss existing approaches to the FILL IN THE BLANKS (FITB) task.

In Chapter 3 we discuss existing techniques for correctly predicting the author's choice of near synonyms, specifically those described by Edmonds (1997) and Inkpen (2007b), and consider their success in choosing between near synonyms where there are sentiment differences between the synonyms, showing that there is some difference in performance.

In Chapter 4 we describe several new approaches to the FITB task, relying primarily on unigram word presence features, augmented with a distance weighting measure which allows features from the entire document to be considered, and analysed specifically in terms of improving on correctly predicting the author's choice of near synonym in the specific case where there are sentiment differences between the synonyms.

In Chapter 5 we investigate valence shifting by lexical substitution, both demonstrating with human subjects that this works as expected, and demonstrating that a metric based on Kullback–Leibler divergence measures may be able to capture the ability of a lexical choice to affect the perceived negativity of the sentence.

In Chapter 6 we review the overall contributions of this thesis and suggest avenues of further research.

Chapter 2

Related work

In this chapter, we will first review the conceptual background to this project. We begin by discussing the theory of meaning and synonymy in Section 2.1, as this provides us with the key task definition for this work. We then briefly discuss central computational approaches to semantics in general in Section 2.2, together with three semantic tasks that are related to words and phrases with similar meanings: paraphrase, textual entailment, and word sense disambiguation. Finally we discuss the broad task requiring the correct choice of near-synonym, which is the problem of Natural Language Generation (NLG) in Section 2.3 with particular reference to the task of REALISATION, which is choosing the surface form of language output.

We then review the specific task which we will explore throughout this thesis, that of computational approaches to near-synonym choice. We discuss the existing approaches to the specific FILL IN THE BLANKS (FITB) task in Section 2.4 together with their present performance. We then review the popular research avenue of sentiment analysis, which we use to justify certain approaches to near-synonymy, in Section 2.5. Finally, we review the task of VALENCE SHIFTING, one of the uses to which near-synonym choice can be put, in Section 2.6. Finally we conclude in Section 2.7 with some avenues of investigation suggested by the literature, which inform the rest of this thesis.

2.1 Meaning and synonymy

In this section, we describe the characterisation of the meaning of words, when they are regarded as synonyms, and the historical tendency to model the relationship between synonyms and near-synonyms differently from the relationship between other words.

Loosely, synonymy is the relation where two words mean the same thing. "Mean[ing] the same thing" of course does an enormous amount of work here, and the philosophical literature on what constitutes meaning, and then the same meaning, is very large. As a very small example of the kind of problems that arise Lyons (1977) discusses the sentence given by Frege (1892):

(2.1) The Morning Star is the Evening Star

The complexity of the idea of *the Morning Star* and *the Evening Star* "meaning the same thing" becomes clear when one considers that, historically, it was not always known that the Morning and Evening Stars are actually both the planet Venus. So example (2.1) is not a simple tautology informing us that Venus is identical with itself, it contains actual information. Furthermore, the two phrases *the Morning Star* and *the Evening Star* are not substitutable for each other despite referring to the planet Venus, nor can every occurrence of *the Morning Star* and *the Evening Star* in English be substituted by *[the planet] Venus*, for example, in literary or poetic usages.

There are thus many complexities in the notion of meaning, which have led to much philosophical discussion. In this review we do not provide an extensive summary of the philosophical literature: for influential texts summarising the key philosophical arguments by linguistic semanticists see Lyons (1977) and Cruse (1986). Rather, in this section, we introduce the key concepts of SYNONYMY, PARAPHRASE, REFERENCE, CONNOTATION and DENOTATION as used in work on the computational linguistic problems of word choice, paraphrase and textual entailment.

2.1.1 Sense, reference, connotation and denotation

The terms DENOTATION, CONNOTATION, SENSE and REFERENCE in the philosophical literature may refer to several possible relationships between utterances and their meaning.

One such is a set-theoretic framework for semantics which is called MODEL-THEORETIC or TARSKIAN MODEL-THEORETIC, for the philosophical treatment of truth of Tarski (1936). In model-theoretic semantics there is a distinction between the denotation of a word and its connotation, and between its sense and its reference. The term DENOTATION refers to the reference-related parts of a word's meaning. Lyons (1977) gives the distinction between reference and denotation as follows: reference is a semantic property of a particular utterance, and describes that utterance's relationship with the world insofar as the utterance makes statements about truths. Denotation is rather a property of lexemes independent of particular utterances. For example, the denotation of *cow* as a lexeme is

not a particular cow, but rather the set of its possible referents: all cows. The denotation of *red* is not a particular red object, but the set of all red objects.

The distinction between REFERENCE and SENSE is due to Frege (1892) and motivated by example (2.1), which shows that two phrases may pick out the same object in the world but nevertheless not be identical in all respects. A word or phrase REFERS to an object that it is picking out of the world, or rather a model world maintained by the parties to the discourse. The SENSE of the word is the meaning independent of referent, so for example independently of picking out the planet Venus, the phrase *the Morning Star* has the sense of a small bright point in the sky associated with the early part of the day.

The traditional split between the linguistic study of SEMANTICS and PRAGMATICS follows the denotation/connotation binary.

In the philosophical literature, the term DENOTATION refers to the reference-related parts of a word's meaning. Lyons (1977) gives the distinction between reference and denotation as follows: reference is a semantic property of a particular utterance, and describes that utterance's relationship with the world insofar as the utterance makes statements about truths. Denotation is rather a property of lexemes independent of particular utterances. For example, the denotation of *cow* as a lexeme is not a particular cow, but rather the set of its possible referents: all cows. The denotation of *red* is not a particular red object, but the set of all red objects.

In this view LEXICAL SEMANTICS is the study of this relationship, the study of the relationship between words and their intrinsic meaning outside (or rather, generalised over) their use in specific utterances.

Lyons observes that the term CONNOTATION is more problematic. In the philosophical literature, following Mill (1843), it may be used to refer to a property in a word's meaning (so that *white* denotes the set of white things, but connotes the property of whiteness itself), or used to refer to the sense of the word, which is a related but slightly different concept (recall the sense of *the Morning Star*, which picks out particular properties of Venus by which to refer to it, but the use of connotation by Mill relates only to the properties themselves, not to the relationship between the properties and an intended referent).

Finally Lyons gives a non-philosophical use of CONNOTATION, which actually corresponds most closely to its use in the rest of this thesis, which is the associations of word: "the connotation of a word is thought of as an emotive or affective component additional to its central meaning." Emotive and affective are used in a narrower sense in this thesis, but as we will see below the term CONNOTATION has tended to be used in the computa-

tional linguistics literature we discuss to mean implications of a word in addition to its central meaning, rather than as it is used in philosophy.

While we will use the model-theoretic account of semantics in this thesis,| alternative models of semantics are possible. For example, cognitive linguists—who refer to the model-theoretic account as the DICTIONARY MODEL—contrast their ENCYCLOPEDIC MODEL, in which the meaning of words are part of a cognitive network of knowledge, in which the denotation of words is not privileged over connotation. (Evans and Green, 2006, pp. 207–210)

2.1.2 Modelling lexical relationships above the near-synonym level

In natural language, a very typical model of the semantics of words is that of a *taxonomy* or *ontology* of words, showing their denotation as either a superset or a subset of that of other words. Typical relationships are SYNONYMY; HYPERNYMY, that is, having a meaning superordinate to another lexical item such as the relationship between *thing* and *truck*); and HYPONYMY, the inverse of hypernymy.

Ontologies are often built manually by domain experts or lexicographers, but may also be built semi-automatically (eg Kang and Lee (2001)) or automatically (eg Hearst (1992, 1998); Mititelu (2006)). They may be domain-specific, such as the insurance company ontology of Kietz et al. (2000) or they may be general purpose, of which the best known example in computational linguistics is WordNet, discussed in Section 2.1.2.1.

2.1.2.1 WordNet

We here describe the general purpose lexical ontology WordNet (Fellbaum, 1998)[1] in more detail, since it will be used as a source of data in this thesis in Chapter 3.

WordNet is an English language lexical database, which arranges words — or rather word senses — into SYNSETS: sets of synonyms. Each synset is given a definition. For example, consider two of the WordNet synsets containing *evil* (the first as a noun sense and the second as an adjective sense):

(2.2) *evil*, *immorality*, *wickedness*, *iniquity*, defined as "morally objectionable behavior"

(2.3) *malefic*, *malevolent*, *malign*, *evil*, defined as "having or exerting a malignant influence"

The primary relationships between synsets are those of hypernymy and hyponymy, thus the concept of a *human* (also a *person*, *individual*, or *soul*) is a hyponym of the less

[1] Available from http://wordnet.princeton.edu/

specific concept *organism* and is a hypernym of the more specific denotations of *adult*, *female person* and *inhabitant* among many others. Other relations provided are those of MERONYMY (part-whole) and ANTONYMY (opposed or contrasted meanings, such as *wet* and *dry*).

WordNet has been influential in having no distinction between items in its synsets, so that no difference is made between *human* (also a *person*, *individual*, or *soul* in its dataset.

WordNet 2.0 is used throughout this thesis.

2.1.3 Defining near-synonymy and near-synonymous relationships

The key principle of synonymy is a persistent distinction made between the relationship between any two lexical items which are often modelled in an ontology as outlined in Section 2.1.2, and the relationship between lexical items identified as (near-)synonyms, for which there is a strong recent tendency to use unhierachical models. In this section, we discuss how (near-)synonyms are defined, and how differences in their connotation and denotation are detailed.

As Stubbs (2001) puts it, "[it] is often said that it is difficult to find examples [of synonyms] which are entirely convincing." However, as Edmonds (1999) observed "absolute synonymy is somewhat of a red herring for us, for if all of the options in a given context are completely identical there is obviously no need for non-random lexical choice." This is as true for our purposes as it was for his. We thus come to the concept of NEAR-SYNONYMY, the term used in the FILL IN THE BLANKS (FITB) literature to signify, essentially, "words that mean similar but not quite the same things" (Edmonds, 1999; Edmonds and Hirst, 2002; Inkpen, 2004; Inkpen and Hirst, 2006; Inkpen, 2007b; Islam and Inkpen, 2010; Wang and Hirst, 2010). As a more precise definition of near-synonym, Edmonds (1999) defines near-synonyms as words that differ in meaning only in a sufficiently fine-grained way. "Sufficiently fine-grained" in this case has two possible, specific, meanings, either:

- the level where the differences in meaning stop corresponding between languages; or
- at the level intuitively defined by lexicographers in the production of dictionaries, particularly dictionaries of word choice.

There are several possible ways of dividing the space of "meaning similar things". Cruse (1986) gives a criterion of "low degree of implicit contrastiveness" for synonyms, excluding, for example the dog breeds *spaniel* and *alsatian* because part of the use of those words is to exclude other dogs from their reference. Overall, Cruse gives the following sub-categorisation of synonyms:

cognitive synonyms which, when substituted for one another, preserve the truth conditions of a sentence but may change other properties of it such as style; and

plesionyms which, when substituted for one another, alter the truth conditions of a sentence but preserve a certain amount of semantic similarity.

As an example of cognitive synonymy, he gives *violin* and *fiddle* as in his example context here:

(2.4) He plays the *violin* very well.

(2.5) He plays the *fiddle* very well.

He permits cognitive synonyms to differ in expressive meaning (for example *father* and *daddy*), it is propositional content he is concerned with in making this distinction. The explicit test then for a plesionym is that it is possible to assert one while denying the other, for which he gives examples including the following:

(2.6) It wasn't *foggy* last Friday — just *misty*.

(2.7) You did not *thrash* us at badminton — but I admit you *beat* us.

Cruse does not absolutely distinguish where plesionymity becomes non-synonymity, instead identifying a notion of SEMANTIC DISTANCE. In previous work on the FILL IN THE BLANKS (FITB) problem in particular, this is formalised by Edmonds (1999); Edmonds and Hirst (2002) who have an ontology of concepts that becomes clustered, rather than hierarchical, at the level of plesionyms (henceforth "near-synonyms", the term more widely used in the FITB literature).

In the computational linguistics literature, the divisions given by Cruse are broadly maintained in what comes to be called near-synonyms that differ denotationally (roughly, cognitive synonyms) and connotationally (roughly, plesionyms). Initially DiMarco et al. (1993) follow Cruse closely and uses the reference/sense distinction dating to Frege, give two main axes for lexical choice at the near-synonym level:

denotational choices where the choice is between two words that have slightly different meanings, for example *mist* and *fog*; and

connotative choices where the choice is between two words that have the same meaning but have different usages, such as the different interpersonal contexts in which the speaker might choose to use *police officer*, *cop* and *pig*.

However, once the focus becomes choosing the correct word from within a set of near-synonyms, the distinctions made among them grow. Edmonds (1999) uses 35 distinctions in total, collected into four groups by Edmonds and Hirst (2002):

stylistic along such dimensions as the formality of the word;

expressive, including attitudinal along such dimensions as whether the word implies a negative or positive judgement towards the object of the description; or

denotational along such dimensions as whether the word implies certain other events, facts or judgements.

structural along such dimensions as collocation, and syntactic variations

Inkpen and Hirst (2006) uses a similar set of groups, removing structural entirely and replacing expressive variations only with attitudinal variations. Ultimately, they group stylistic and attitudinal distinctions together into a class they term ATTITUDE-STYLE DISTINCTIONS, which loosely corresponds to the CONNOTATIONAL distinctions of DiMarco et al. (1993).

2.1.4 Modeling near-synonymy

Given the tendency described in Section 2.1.2 to model meaning in a hierarchical structure, it is the obvious representation of near-synonyms. However, Edmonds and Hirst (2002) argue against using an ontology such as WordNet to model the differences between near synonyms for the following reasons:

- Ontologies privilege hypernymy as the primary relationship between concepts, whereas there is no obvious reason that, among near synonyms, a hypernymy relationship between two words is more important than that of, say, a difference in attitude (such as a pejorative word choice rather than a neutral one).
- The ontology will be very shallow: most near-synonyms will be directly under their parent concept, which is not a very useful representation for choosing between them.
- The ontology will become highly language specific: differences between near synonyms for an identical concept vary across languages.

Hirst (1995) differentiates between two models of near synonyms: a prototype theory approach and a Saussurean approach. In the prototype theory approach, knowledge is modelled on the basis of *prototypes*: the user of the prototypes understands the word *tree*

by analogy with the prototype of a tree that they have in their knowledge base. Near synonyms would be modelled as either individual prototypes or variants on a prototype theme.

In contrast, in the Saussurean approach to modelling near synonyms, near synonyms are associated with each other solely by their differences. Hirst (1995) argues that this is much more computationally feasible than the prototype theory approach. Edmonds and Hirst (2002) and Inkpen and Hirst (2006) both describe formalised versions of the Saussurean approach, which they call a CLUSTER model, in which there is a coarse-grained ontology which models concepts which are cross-linguistic and where, attached to each node in the ontology, there is a cluster of near-synonyms, annotated with their differences from each other. DiMarco et al. (1993) had earlier proposed a similar idea, termed FORMAL USAGE NOTES, based on entries in word usage guides.

Consider the diagram from Edmonds and Hirst (2002) shown in Figure 2.1. We see that the ontology extends from the denotation of THING[2] to that of ERROR, and under error there are language specific clusters, including the English words *error*, *mistake*, *blunder* etc., with lines between them representing the non-hierarchical differences in their meaning.

The motivation for the statistical approach to lexical choice by Edmonds (1997) (see Section 2.4) as discussed in Edmonds (1999) seems to have been to further motivate work on these lexical semantic models by demonstrating the limitations of the narrow-context statistical approaches. The increasing success of the statistical models alone has led inquiry into the FITB task away from knowledge-based approaches to date.

2.2 Computation and meaning

In this section, we discuss several computational linguistic approaches to meaning. We briefly introduce the area of computational semantics in Section 2.2.1, and then discuss three specific problems which are related to the near-synonym choice problem: paraphrase in Section 2.2.2; textual entailment in Section 2.2.3; and word sense disambiguation in Section 2.2.4. We introduce these in order to provide background on some of the approaches to meaning in computational linguistics more broadly.

[2]Items in the hierarchy represent multiple terms, we denote all words in the near-synonym set with *thing* as THING.

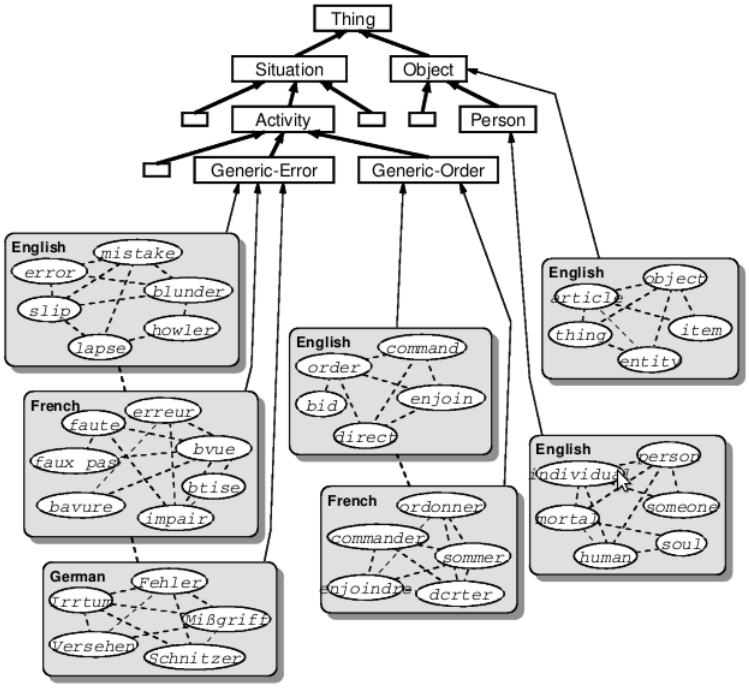


Figure 2.1: A clustered model of lexical knowledge (Edmonds and Hirst, 2002, Figure 6)

2.2.1 Overview of computational semantics

Computational semantics is the integration of formal semantics, and of automated reasoning, into computational linguistics, with a goal of translating symbolic representations of meaning such as logical formulae into human language and vice versa (Bos, 2011).[3]

One major application of computational semantics in natural language processing is that once one has acquired a formal representation of knowledge, one can use automatic inferencing to draw conclusions about the world in a way that may assist an application with its goal. As an example, Gardent and Webber (2001) formulate referring expression disambiguation as a formal problem for reasoners to address. In the question answering domain in particular, even without extensive reasoning, semantic representations may either reduce ambiguity in searching for answers, or allow higher recall in an initial search by allowing searches for semantically related text that do not share a similar surface form (eg Ahn et al. (2005); Furbach et al. (2010)).

Computational semantics relies on databases of background knowledge of various degrees of structure. Use of the WordNet database of hypernymy, hyponymy and other

[3]For full introduction to computational semantics, and a survey of present research, directions, the reader is respectively referred to Blackburn and Bos (2005) and to Bos (2011).

semantic relations (Fellbaum, 1998) is widespread both as a data source and as a gold standard: a very incomplete list includes Lin (1998); McCarthy et al. (2004); Budanitsky and Hirst (2006); Snow et al. (2006). It is used extensively as one of the central gold standards in word sense disambiguation. Other important resources are FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005) and VerbNet (Kipper et al., 2008) (all as cited in Bos (2011)).

Clearly, as we discussed in Section 2.1, the field of lexical semantics has closely concerned itself with synonymy and near-synonymy, and thus computational semantic approaches are one natural approach to the problem. The lexical knowledge base approaches to near-synonymy of Edmonds (1999); Edmonds and Hirst (2002); Inkpen (2004); Inkpen and Hirst (2006) can be characterised as computational semantic approaches, contrasting with later corpus statistics approaches which rely on less formal features, and which are discussed further in Section 2.4.

2.2.2 Paraphrase

The definition of paraphrase is similar to that of synonymy: phrases (from the level of words upwards) that mean the same thing, with the same difficult considerations with regards to "meaning the same thing" discussed in Section 2.1. In computational linguistic usage, the definition tends to be less strong than that of synonyms. For example, the definition given by Dras (1999, definitions 4.1.1 and 4.1.3) is still widely cited:

> 1. A PARAPHRASE PAIR is a pair of units of text deemed to be interchangeable[…]
> 2. A PARAPHRASE of a unit of text is the alternative to that unit of text given in some paraphrase pair.

Likewise, Madnani and Dorr (2010) give examples where words and phrases which are substitutable, rather than having identical semantic content, are considered paraphrases, such as substituting the hypernym *say* for the more specific word *reply*. Hirst (2003) described paraphrasing as "talk[ing] about the same situation in a different way" where the different ways include variations on similar axes to the near-synonym axes: denotation, connotation, style and so on.

There are two broad parts to the paraphrase problem, the first is identifying paraphrase pairs, and the other is substituting them for one another. The paraphrase task has several applications, including query expansion for information retrieval and question answering

where the surface form of a query may not by itself resemble the context of the answer closely; assessing the output of machine translation systems by allowing paraphrases of the human-authored gold standard output; and improving machine translation by searching for translations of not only a phrase from the original text but also translations of its paraphrases. (Madnani and Dorr, 2010)

Knowledge based approaches to paraphrase problems have been widely discussed (among the many examples Madnani and Dorr list are McKeown (1979); Dras (1999); Fujita et al. (2004)), corpus-based approaches have also been widely investigated more recently. Some of the corpus-based approaches to paraphrase acquisition discussed by Madnani and Dorr include:

paraphrase acquisition from a single corpus in which phrases with similar meanings are usually identified by appearing in similar contexts (known as DISTRIBUTIONAL SIMILARITY) (eg Lin (1998) for words, and Lin and Pantel (2001); Paşca and Dienes (2005) for multi-word paraphrases); and

paraphrasing using parallel corpora in a single language, for example, multiple translations of a single text as used by Barzilay and McKeown (2001).

Some researchers have additionally investigated generating paraphrases with specific qualities, for example Zhao et al. (2009) investigated generating paraphrases tuned for specific desires such as shorter or simpler sentences; and as discussed in Section 2.3.1 Keshtkar (2011) investigated acquiring and generating paraphrases that shared an emotional property, such as fear, anger or happiness.

Evaluation is typically manual, with authors such as Barzilay and McKeown presenting human judges with paraphrase pairs and asking them to assess the correctness. Callison-Burch (2007, Section 4.1) argued for substitution-based evaluation instead, in which one of a paraphrase pair replaces the other in a context and judges are asked questions similar to those suggested by Linguistic Data Consortium (2005) for machine translation. Callison-Burch thus asked subjects:

- How much of the meaning of the original phrase is expressed in the paraphrase?
- How do you judge the fluency of the sentence?

As is typical with Natural Language Generation evaluation, to be discussed in Section 2.3.2, Madnani and Dorr report that the evaluation of paraphrase is very often

Text	Hypothesis	Entailment
Norway's most famous painting, "The Scream" by Edvard Munch, was recovered Saturday, almost three months after it was stolen from an Oslo museum.	Edvard Munch painted "The Scream".	True
The Republic of Yemen is an Arab, Islamic and independent sovereign state whose integrity is inviolable, and no part of which may be ceded.	The national language of Yemen is Arabic.	True
Bush returned to the White House late Saturday while his running mate was off campaigning in the West.	Bush left the White House.	False

Table 2.1: Examples of Text-Hypothesis pairs from Dagan et al. (2006, Table 1)

application-driven, since the generation of paraphrases is seldom pursued as an application in and of itself.

2.2.3 Textual entailment

Textual entailment is the problem of recognising or generating pairs of statements such that the truth of one statement entails the truth of the other. These are referred to as the text (T) and hypothesis (H). Examples of three text/hypothesis pairs from Dagan et al. (2006) are shown in Table 2.1.

The entailment problem has ancient roots, with logical inference explored and catalogued extensively by Aristotle (350 BCE). Computational linguistic work on inference from one statement to another also has a long history (comparatively speaking), with examples of early work including Charniak (1979) discussing how to infer the correct domain of discussion (he gives the example of inferring that the phrase "[t]he woman waved while the man on the stage sawed her in half" is in the domain of magic performances) and Joshi et al. (1984) exploring how to avoid having a system make statements from which a hearer might draw false inferences.

The Recognising Textual Entailment (RTE) challenge has been posed in the natural language processing community annually for several years, beginning with the 2005 challenge described by Dagan et al. (2006) and most recently with RTE-7, held in 2011 (Bentivogli et al., 2011).

The survey of Androutsopoulos and Malakasiotis (2010) lists seven textual entailment and paraphrase recognition approaches: using logical representations to check for logi-

cal entailment; using vector space models of semantics to check for "sufficient closeness" in meaning; using the surface form of T and H to evaluate similarity; using models of syntactic structure to evaluate similarity; using semantic role-labelling resources such as FrameNet (Baker et al., 1998) to evaluate similarity; using machine learning techniques with features such as context of the phrases and certain features within them including negation; and the template-based method of searching for similar enough contexts for the two phrases.

2.2.4 Word sense disambiguation

The problem of word sense disambiguation (WSD) is the problem of distinguishing between words that have identical surface forms but different meanings. Navigli (2009) gives examples (2.8) and (2.9), in which *bass* has the sense of *low-pitched* and *a type of fish* respectively:

(2.8) I can hear *bass* sounds.

(2.9) They liked grilled *bass*.

Navigli describes WSD as an intermediate task, ideally solved in applications like machine translation, since the two senses of *bass* in English are likely to have different translations in a target language. However, he describes the contribution of WSD to applications as still undemonstrated, despite the essence of the description of the WSD task dates back to early machine translation literature (Weaver, 1955).

Approaches to WSD make considerable use of structured knowledge bases including thesauri, machine readable dictionaries and ontologies, together with corpora both sense-annotated and not. Navigli characterises the computational approaches as largely based on machine learning, with examples including Yarowsky (1992) using decision lists; Mooney (1996) using several approaches including neural nets and decision trees; and Lee and Ng (2002) using Support Vector Machines.

2.3 Word choice in natural language generation

Natural Language Generation (NLG) is the problem of the generation, rather than analysis, of natural language. Generating natural language requires concrete decisions with regard to the meaning of words, because ultimately the system must produce meaningful output to be successful. In this section, we focus on NLG treatments of affect, and discuss

Module	Content task	Structure task
Document planning	Content determination	Document structuring
Microplanning	Lexicalisation	Aggregation
	Referring expression generation	
Realisation	Linguistic realisation	Structure realisation

Table 2.2: Natural Language Generation modules and tasks (Reiter and Dale, 2000, Figure 3.1)

the existing evaluation of its success in order to inform our evaluation of the valence-shifting task.

NLG can be divided into at least three sub-tasks, in the widely accepted division given by Reiter and Dale (2000):

document planning deciding what to say, that is given a communicative goal (such as "describe tomorrow's weather"), output a discourse plan to achieve the goal, at about sentence-level granularity

microplanning optimising the discourse plan by, for example, combining or dividing sentences, removing repetition, and other manipulations to improve the clarity of the communication at the sentence level

realisation transforming the discourse plan into natural language output, that is, a string of meaningful words

Reiter and Dale (2000) provide a more detailed breakdown shown in Table 2.2.

Here we see the lexicalisation task occurring at the microplanning phrase of NLG, they later note controversy about which phase of NLG lexicalisation belongs to, with suggestions that it could occur at any time from document planning on.

In any case, NLG ultimately requires that the lexicalisation problem be solved. In this section, we outline several natural language generation approaches to lexicalisation. In Section 2.3.1 we discuss systems which integrate affective lexical choice mechanisms; and in Section 2.3.2 we discuss the present state of evaluating the success of NLG systems with particular reference to realisation.

For a general review of NLG, the reader is referred to Reiter and Dale (2000) and Krahmer and Theune (2010).

2.3.1 NLG systems which integrate affective lexical choice mechanisms

NLG systems already require subtle word and phrase choice decisions to be made for some applications. Generating affective lexical items in particular has been addressed as part of designing a CONVERSATIONAL AGENT: a dialogue agent that communicates with human partners in a natural style. Such agents will often have use-cases in which the agent must communicate in an emotionally charged environment Closely related to and partially overlapping with NLG systems are authoring tools, which could make suggestions about appropriate word choice.

Fleischman and Hovy (2002) investigated word choices for a conversational agent by modelling the task in terms of agents and their dispositions towards—plans and goals surrounding—or attitudes to the subject of their speech, where attitude was simplified to a single integer value from -5 to 5. Values are manually assigned: in their example of discussing the driver of a military vehicle that had an accident, using the driver's name is assigned a disposition of 5, *the driver* a disposition of 0 and *a private* a disposition of -2. A realisation fitting attitudes to different sentential constituents was chosen using a simple distance metric which correlated significantly with human judgements of the attitude of the resulting sentences. The metric used is as follows:

1. the speaker's attitude to the object x, is manually assigned, is denoted by attitude(x)
2. the distance between the speaker's attitude to x attitude(x) and the sentiment strength of a particular lexical reference to x shade(x) is computed, looking up shades in a manually compiled database:

$$\text{Dist}(x) = |\text{attitude}(x) - \text{shade}(x)| \tag{2.1}$$

3. the emotional score of the entire sentence is computed:

$$\text{EmotScore}(x) = \text{Dist}(\text{verb}) - \sum \text{Dist}(\text{constituent}) \tag{2.2}$$

Mairesse and Walker (2010) describe a complex conversational agent system called PERSONAGE designed to be variable along the so-called "Big Five" personality traits: extraversion, emotional stability, agreeableness, conscientiousness and openness to experience, tested in the domain of restaurant recommendation. The PERSONAGE system integrates parametrised personality traits into every step of the standard NLG pipeline: their

Emotion	Paraphrases
Disgust	*getting of evil*; *been rather sick*
Anger	*am royally pissed*; *see me angry*
Happiness	*the love of*; *good feeling*

Table 2.3: Examples of emotionally laden paraphrases acquired by Keshtkar (2011, Table 4.7)

many parameters include variables ranging from the tendency of the personality to express multiple ideas in a sentence or not and its tendency to use expletives. At the level of lexical choice they make use of verb near-synonymy, giving the example of *appreciate*, *like* and *love*, using the STRONGER-THAN relationship in the VERBOCEAN database (Chklovski and Pantel, 2004).

Researchers addressing the FILL IN THE BLANKS (FITB) task have often cited an INTELLIGENT THESAURUS as a possible application of their work. Edmonds and Hirst (2002) describe such an intelligent thesaurus which moves beyond simple lists or descriptions of near-synonyms:

> [The thesaurus] would actively help a person to find and choose the right word in any context. Rather than merely list possibilities, it would rank them according to the context and to parameters supplied by the user and would also explain potential effects of any choice, which would be especially useful in computer-assisted second-language instruction.

One such intelligent thesaurus system is described by Inkpen (2007a). Several authors who have addressed near-synonym lexical choice have subsequently also integrated their systems as a module in an NLG or conversational agent system (Inkpen and Hirst, 2006; Inkpen, 2007b; Islam, 2011).

In Dras et al. (2010a,b) we developed a conversational agent designed to train border agents to detect deception, for example, entering a country on false premises. The agent varies its use of language depending on whether it is being deceptive or not by, among other techniques, having the deceptive agent make more negative statements and complaints.

Keshtkar (2011) acquired emotion-inflected paraphrases from a blog corpus using a bootstrapping method based on seed words (*glad* and *cheerful* were two seed words for happiness, for example, and *anxiously* and *distrust* for fear). From the context surrounding these seeds, he discovers contexts which may yield paraphrases for these seeds. Example paraphrases are shown in Table 2.3.

Keshtkar then used a modified version of SimpleNLG (Gatt and Reiter, 2009) to develop a template-based NLG system in which the NLG system substitutes elements into largely pre-prepared texts, with the goal of producing an authoring system in which the authors can supply most of the text and the NLG system can supply appropriately emotive paraphrases. Keshtkar also explored the generation of affective paraphrases identified using features from the emotionally laden blog corpus. Keshtkar is largely interesting in *retaining* the emotion of the original sentence, but expressing it differently, thus, given input of "I am so *incredibly angry* right now!" his system replaced it with "I am so *unbelievably mad* right now!"(Keshtkar, 2011, Table 6.13).

2.3.2 Evaluating natural language generation and word choice

In order to measure our success on the word choice task, we are obviously in need of an evaluation strategy. Given word w as a system's chosen word for a particular meaning, we need to evaluate the fittingness of that choice. In this section we discuss NLG evaluation with an emphasis on realisation, with reference to the applicability of realisation metrics for evaluating word choice.

Reiter and Belz (2009) give an overview of possible evaluation techniques for NLG:

task based evaluations directly measuring the impact of generated texts on end users, for example measuring how well users carry out instructions or how health decisions change amongst users of the tool;

evaluations based on human ratings and judgements human judges are asked to rate various features of text such as coherence, writing style and correctness, typically on an n-point rating scale; or

evaluation on automated metrics where an automatically calculated metric can be found that correlates with either task-based evaluations or human ratings, evaluation can be performed using the value of the metric.

Reiter and Belz also review reasons to use each type of evaluation: in particular, automated evaluation is usually cheaper than human ratings which are in turn typically cheaper than task-based evaluations. However, other considerations may occur, such as that it may not be ethical to use task-based evaluations or even human ratings until the system has reached a certain level of effectiveness. It is also sometimes the case that one type of evaluation is unsuitable, for example Reiter and Belz aren't certain what a task

based evaluation of the output of a humour generation system (Binsted et al., 1997) would involve.

The scope of evaluation is also unclear. Reiter (2011) draws a distinction between CONTROLLED evaluations where systems' outputs are compared against a baseline output with as little confounding as possible, and ECOLOGICALLY VALID evaluations where systems' outputs are evaluated by users who match the target audience of the product, and notes that his own opinion, in task-based applications such as medical advice systems, is still in flux.

A methodology for automated evaluation of NLG systems is far from being settled on for all common tasks. The *Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation: Position Papers* (Dale and White, 2007) shows the range of discussion, with, for example, Paris et al. arguing that then-present NLG evaluation relied too strongly on comparisons to reference output rather than on the needs of system stakeholders; Scott and Moore listing no fewer than eight reasons to be cautious of shared-task metric-based evaluation systems including the difficulty of creating gold standards that are meaningful, and the likely prematurity of specifying what the input into any one phase of the NLG process should look like; and Belz arguing that only by comparative evaluation of core tasks can the NLG progress collectively on core tasks. Before and since that discussion, some tasks have begun shared evaluations, including regular expression generation, surface realisation and text correction (eg Gatt and Belz; Belz et al.; Gatt et al. in 2008, Janarthanam and Lemon; Belz et al.; Dale and Kilgarriff; Striegnitz et al. in 2011).

Moreover, a significant proportion of realisation literature concentrates on the REFERRING EXPRESSION GENERATION (REG) problem, that is, "how we produce a description of an entity that enables the hearer to identify that entity in a given context" (Reiter and Dale, 2000, p. 55). This problem overlaps with the word choice problem in some respects, including the near-synonym choice problem: to give an example, in attempting to pick out a young man for further discussion, is the system better off referring to the *teen* or the *youth*? But choosing the appropriate word or phrase to uniquely identify a discourse entity is not the entire word choice problem: word choice also affects style, perception, ease-of-use and other aspects of NLG output. Therefore we do not treat REG evaluation extensively in this review, and the reader is referred to Krahmer and van Deemter (2012, pp. 199–203) for a extensive survey of REG evaluation.

In brief, evaluation of referring expression *realisation* is performed by comparing the expressions (almost always multi-word phrases like "the red ball") with referring expres-

sions solicited from human subjects either in the course of a complex task such as giving directions (Anderson et al., 1991) or when specifically asked to uniquely identify objects in a scene (Gorniak and Roy, 2004; Viethen and Dale, 2006). The comparison scores are produced by various distance metrics such as the Dice (1945) or Jaccard (1901) measures. This does not provide us with a guide to evaluating individual word choice, since these metrics are designed to measure the distance between multi-word phrases.

A concrete task for evaluating the appropriate choice of near-synonym was thus developed by Edmonds (1997), the Fill In the Blanks (FITB) task. This task is central to this thesis, and is therefore discussed in detail in Section 2.4.

2.4 The Fill In the Blanks (FITB) task and related tasks

In this section we discuss the Fill In the Blanks (FITB) task initially posed by Edmonds (1997) and the varying approaches of Edmonds (1997, 1999); Inkpen (2007b); Islam and Inkpen (2010); Wang and Hirst (2010); Yu et al. (2010); Islam (2011). A summary of the task as posed by Edmonds is given in Section 2.4.1, a detailed overview of each of their approaches is given in Section 2.4.2. A comparative summary of the performance of each approach is given in Section 2.4.3. In Section 2.4.4 we describe approaches taken to other similar lexical choice problems.

We take a full section to discuss this as we use this as an evaluation method in the thesis to investigate the behaviour of affective and non-affective synonyms.

2.4.1 Overview of the FITB task

As discussed above, the FITB task has been directly addressed by several authors. It was introduced by Edmonds (1997), who posed it as follows:

> [An] important sub-problem [of lexical choice is] that of determining the near-synonym that is most typical, or expected, if any, in a given context. Although weaker than full lexical choice, because it doesn't choose the 'best' word, we believe that it is a necessary first step, because it would allow one to determine the effects of choosing a non-typical word in place of the typical word…
>
> For example, our implemented lexical choice program selects *mistake* as most typical for the 'gap' in sentence [2.10], and *error* in [2.11].

(2.10) However, such a move also would run the risk of cutting deeply into U.S. economic growth, which is why some economists think it would be a big {*error* | *mistake* | *oversight*}.

(2.11) The {*error* | *mistake* | *oversight*} was magnified when the Army failed to charge the standard percentage rate for packing and handling.

2.4.2 Approaches to the FITB task

In this section we outline the several existing approaches to the FITB task. In Section 2.4.2.1 we describe the baseline method proposed by Edmonds (1997, 1999) and used thereafter; in Section 2.4.2.2 we describe the first approach to FITB as developed by Edmonds; in Section 2.4.2.3 an anti-collocation method developed by Inkpen (2007b); in Section 2.4.2.4 we describe an unsupervised approach to FITB also developed by Inkpen; in Section 2.4.2.5 we describe a supervised approach to FITB also developed by Inkpen; in Section 2.4.2.6 we describe a language model approach developed by Islam and Inkpen (2010); in Section 2.4.2.7 an approach using features identified by Latent Semantic Analysis developed by Wang and Hirst (2010); in Section 2.4.2.8 an approach using cosine similarity measures adjusted for errors developed by Yu et al. (2010); and in Section 2.4.2.9 we describe a two-phrase frequency method with a fall-back by Islam (2011).

Results for each approach will be presented in Section 2.4.3.

2.4.2.1 The most frequent baseline approach

The baseline approach suggested by Edmonds (1997, 1999) is that of choosing the most frequent near-synonym to fill every gap. In the case of example (2.10), the system would count the occurrence of *error*, *mistake* or *oversight* in a corpus, and predict the one with the highest frequency count every time the problem was posed for that near-synonym set.

2.4.2.2 Lexical co-occurrence network (Edmonds-Collocate)

Edmonds (1997, 1999) describes an approach he terms a LEXICAL CO-OCCURRENCE NETWORK, henceforth called EDMONDS-COLLOCATE in this thesis. EDMONDS-COLLOCATE relies on collocation statistics, so that the choice of a near-synonym is based on the words that closely surround it.

EDMONDS-COLLOCATE predicts near-synonym choice in part-of-speech tagged text, so that the system uses part-of-speech tagged tokens such as *(JJ arduous)* or *(NN fight)* both as candidate words and as words surrounding the gap.

In order to choose between candidate tokens $c_1 \dots c_n$ to fill a gap g in a sentence S, each candidate c_i is assigned a score score(c_i, g) measuring its appropriateness for the gap g. The score is a function of the collocational score, co-occur, of each individual word w_j in the sentence S_g which contains gap g.

$$\text{score}(c_i, S_g) = \sum_{w_j \in S} \text{co-occur}(c_i, w_j) \tag{2.3}$$

The candidate c_i which has the largest value of score(c_i, S_g) is chosen as the word fitting the lexical gap g in sentence S_g. Where there is more than one candidate c_i with an equal maximum value of score(c_i, S), or where no candidate has a non-zero score, we regard the Edmonds's method as unable to make a prediction.

Edmonds computed the score coccur(c_i, w_j) by connecting words in a COLLOCATION NETWORK. The principle of his collocation network is that if token w_x has a co-occurrence score above a set threshold in a training corpus with token w_y which in turn co-occurs above the threshold with token w_z, then the presence of w_x should weakly predict the appearance of w_z even if they do not co-occur together above the threshold in the training corpus. That is, he assumes that if, for example, *task* co-occurs above the threshold with *difficult*, and *difficult* co-occurs with *learn*, then *task* and *learn* should weakly predict each other's presence.

Edmonds proposes extending this technique to co-occurrence networks with prediction chains of arbitrary length, but his experimental results suggest that in practice two connections approaches the limit of the usefulness of the technique. Therefore, to compute co-occur(c_i, w_j) we take the shortest path of collocation between the tokens c_i and w_j, which is either c_i, w_j where c_i and w_j co-occur, or c_i, w_x, w_k where c_i and w_j both co-occur with a third word, w_x.

Where tokens c_i and w_j co-occur, their co-occurrence score is their t-score (Church et al., 1991):

$$\text{co-occur}(c_i, w_j) = t(c_i, w_j) \tag{2.4}$$

Where tokens c_i and w_j both significantly co-occur with token w_x, their significance score is a combination of their t-scores, with a bias factor devised by Edmonds to account for their weaker connection.

$$\text{sig}(c_i, w_j) = \frac{1}{8}(t(c_i, w_x) + \frac{t(w_x, w_j)}{2}) \tag{2.5}$$

If there is more than one candidate word w_x co-occurring significantly with both c_i and w_j, the word w_x is chosen so that the value of $\text{sig}(c_i, w_j)$ in equation 2.5 is maximised.

In the above, we have used "co-occur" and "co-occur above a threshold" interchangeably and without definition. Quantitative thresholds for these are given by Edmonds (1999): any two words w_x and w_y co-occur above threshold[4] if their t-scores are greater than 2.0 and their mutual information score is greater than 3.0. This was suggested to Edmonds by the observation of Church et al. (1991) that t-scores and mutual information scores emphasise different kinds of co-occurrence: co-occurrence with relatively common and uncommon words respectively. He thus uses both scores to attempt to capture both types of co-occurrence.

2.4.2.3 Anti-collocations method

The anti-collocations method of near-synonym choice is briefly described by Inkpen (2007b). In this method, Inkpen draws on a database of near-synonym collocations developed by Inkpen and Hirst (2002) and compute collocations using the 100 million word British National Corpus[5] in order to have collocations with different styles and topics.

The intent of the database is to identify preferential collocations, that is, words that collocate with one near-synonym from a cluster but not as much with another. They initially use several separate measures of collocation; point-wise mutual information (PMI) (Church and Hanks, 1990); the Dice co-efficient (Dice, 1945); Pearson's Chi-square; Log-likelihood ratios; and Fisher's exact test (last three as calculated in Pedersen (1996)). This identifies candidate collocations with presumably high recall but possibly not precision. These candidates are then filtered using a differential t-test on the results of web searches to eliminate some collocations.

Table 2.4 shows sample collocations for the members of the *task*, *job*, *assignment* etc. near-synonym set with three adjectives, and illustrates how *daunting* might, for example, be used to predict the use of *task*.

Inkpen (2007b) uses this database to predict near-synonym usage for the FITB task, simply predicting the near-synonym with the strong collocation scores (and the most frequent one, in the event of ties).

[4] In Edmonds' terminology w_x and w_y SIGNIFICANTLY CO-OCCUR, we have chosen an alternative phrasing to avoid confusion with the unrelated concept of statistical significance.

[5] http://www.hcu.ox.ac.uk/BNC/

Near synonyms	*daunting*	*particular*	*tough*
task	√	√	√
job	?	√	√
assignment	*	√	√
chore	*	?	*
duty	*	√	*
stint	*	*	*
hitch	*	*	*

√ indicates a collocation, ? a less-preferred collocation and * an anti-collocation

Table 2.4: Example collocations produced by the method of Inkpen and Hirst (2002, Table 3)

2.4.2.4 Point-wise mutual information (Inkpen-PMI)

Although the results of Edmonds (1997) were not so promising, the increasing use of the World Wide Web as a corpus (e.g. as early as Grefenstette (1999) for machine translation, and Turney (2001) for statistical association-based searches for synonyms) encouraged Inkpen (2007b) to turn to the Web for the FITB task.

The unsupervised method developed by Inkpen, hereafter INKPEN-PMI, is that of using the point-wise mutual information (PMI) of each candidate near-synonym c with the words in its immediate context. PMI is computed as in equation (2.6) (Church and Hanks, 1989):

$$\text{PMI}(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \tag{2.6}$$

In INKPEN-PMI, the suitability of candidate c for a given gap is approximated differently from EDMONDS-COLLOCATE: the entire sentence is not used to measure the suitability of the word. Instead, a certain sized window of k words either side of the gap is used. For example, if $k = 3$, the word missing from the sentence in example (2.12)) is predicted using only the six content words shown in example (2.13).

(2.12) Visitors to Istanbul often sense a second, _____ layer beneath the city's tangible beauty.

(2.13) Istanbul [often] sense [a] second, _____ layer [beneath] [the] city's tangible

Given a text fragment f consisting of $2k$ words, k words either side of a gap g $(w_1, w_2, \ldots, w_k, g, w_{k+1}, \ldots, w_{2k})$, the suitability $s(c, g)$ of any given candidate word c to fill the gap g is given by:

$$s(c,g) = \sum_{j=1}^{k} \text{PMI}(c, w_j) + \sum_{j=k+1}^{2k} \text{PMI}(w_j, c) \tag{2.7}$$

INKPEN-PMI estimates the token counts for $PMI(x, y)$ by issuing queries to the Waterloo MultiText System (Clarke and Terra, 2003) for occurrences of x and y separately and within a QUERY FRAME of length q within a corpus: that is, a count for the number of times x and y occur within distance q of each other.

2.4.2.5 Supervised methods using presence and PMI scoring

Inkpen (2007b) also proposed using supervised methods for predicting near-synonyms usage. In this case, the feature sets considered were:

- the PMI scores of the left and right context g of the missing near-synonym, as computed by Equation 2.7 above
- the presence of words in those context windows (with binary scores of 0 for absence from the context or 1 for presence), limited to the 500 most common words found in the contexts for each set of near-synonyms
- a combination of both features.

Various machine learning approaches were tried, including decision trees, naive Bayes, and k-nearest neighbour.

2.4.2.6 Language model

Islam and Inkpen (2010) use a language model built using data from the Google Web 1T corpus (Brants and Franz, 2006). Web 1T contains n-gram frequency counts, up to and including 5-grams, as they occur in a trillion words of World Wide Web text. In a language model, consider the problem of choose between candidate tokens $c_1 \ldots c_m$ to fill a gap g in a sentence S, containing words $w_1, \ldots w_j$ before the gap g and $w_{j+2}, \ldots, w_n$ after the gap.

The language model approach to the FITB problem is then estimating the comparative probabilities of each of the sentences, one for each candidate c_i:

$w_1, \ldots, w_j, c_1, w_{j+2}, \ldots, w_n$

$w_1, \ldots, w_j, c_2, w_{j+2}, \ldots, w_n$

...

$w_1, \ldots, w_j, c_m, w_{j+2}, \ldots, w_n$

The probability of a sentence S comprising words $w_1, \ldots\ w_n$ is given by equation 2.8 (Islam and Inkpen, 2010, eq 1):

$$P(S) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\ldots P(w_n|w_1\ldots w_{n-1}) \tag{2.8}$$

Probability counts from a corpus are used to approximate the probabilities needed: as Web 1T only has 5-grams, only at most four previous words can be used to estimate any given word. If the counts are missing the probability of the unseen sequence $w_{j-4}, \ldots, w_j$ must be estimated rather than set to zero, lest the probability result in Equation 2.8 be zero. Such counts require a factor α to be introduced estimating the likelihood of unseen data in the probability distribution. Islam and Inkpen fall back to estimating $P(w_j|w_{j-4}\ldots w_{j-1})$ using shorter n-grams with the probability smoothed using the one-count smoothing method (Chen and Goodman, 1996) and values of α estimated by Yuret (2007).

2.4.2.7 Latent Semantic Analysis

The approach of Wang and Hirst (2010) uses Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997), in which dimensionality reduction is performed on a feature space originally composed of vectors representing co-occurring lexical items, to produce a new, smaller, set of features that preserve the most important information in the original, larger, matrix of features. Drawing on the success of Rapp (2008)—on a different task, that of identifying potential synonyms rather than choosing the usage of them—with using a small window of context to derive the initial vectors, they compare large contexts, that is, the document as context, with smaller ones.

Typically the application of the matrix of features resulting from LSA is unsupervised classification, that is, a near-synonym would be chosen for a gap if that near-synonym's feature matrix was closest in cosine similarity to the context around the gap. Wang and Hirst investigate a supervised approach, using the latent semantic space features to train Support Vector Machines in near-synonym choice.

2.4.2.8 Discriminative approach using Web 1T

Yu et al. (2010) propose another discriminative approach to the FITB problem, in addition to those of Inkpen and Wang and Hirst discussed in Sections 2.4.2.5 and 2.4.2.7 respectively. Yu et al. use feature values are derived from the Google Web 1T corpus (Brants

and Franz, 2006) as suggested by this author (Gardiner and Dras, 2007b)[6]. Each individual feature value in their approach is the weight of a word w with respect to collocating with a near-synonym s, where the weight is a function of $c(w|j)$, the number of times that w occurs in a n-gram with s, and $c(w)$, the number of times that w is observed in the corpus in total:

$$\text{weight}(w) = \frac{c(w|s)}{c(w)} \tag{2.9}$$

The initial error rate using cosine similarity scores between gap contexts and near-synonym usages with this weighting is found and then Yu et al. use discriminative training to adjust the waiting so as to diminish the error rate, using the generalised probabilistic descent (GPD) algorithm (Katagiri et al., 1998).

2.4.2.9 Two-phase method

Islam (2011) developed an FITB approach he terms the TWO-PHASE METHOD. Islam categorises n-grams based on the position of the candidate word to fill in the blank, c_j, within the n-gram. For 3-grams, the candidate could be at the end, the middle, or the start of the n-gram, respectively Islam refers to these as having type (k) of 1, 2 and 3.

Islam defines a measure called the NORMALISED FREQUENCY VALUE, where the normalised value F of each candidate c_i is given in terms of the frequency of the n-gram containing c_i and the n-gram frequency f_j of all other candidates $c_j \in c_1, \ldots, c_n$:

$$F(c_i) = \frac{f_i}{\max(f_1, \ldots, f_n)} \tag{2.10}$$

Initially Islam attempts to use the Web 1T data to choose the appropriate candidate by maximising the value of equation (2.10).

2.4.3 Performance of different approaches to FITB

Comparison of the existing approaches is to date always performed using seven near-synonyms chosen by Edmonds (1997) and shown in Table 2.5. Edmonds states that these words were chosen because they have low polysemy and similar frequencies to other words within their near-synonym set.

Edmonds (1999) goes on to perform a more comprehensive experiment, testing his system's performance on the 2103 synsets in WordNet 1.5, and his results are shown in

[6]The work described in Gardiner and Dras (2007b) is incorporated into this thesis in Chapter 3.

Set	POS	Synonyms (with training corpus frequency)[a]
1	JJ	difficult (352), hard (348), tough (230)
2	NN	error (64), mistake (61), oversight (37)
3	NN	job (418), task (123), duty (48)
4	NN	responsibility (142), commitment (122), obligation (96), burden (81)
5	NN	material (177), stuff (79), substance (45)
6	VB	give (624), provide (501), offer (302)
7	VB	settle (126), resolve (79)

[a] [Present author's note] The training corpus referred to is the 1989 part-of-speech tagged *Wall Street Journal* corpus

Table 2.5: Test words used for the FITB task (Edmonds, 1997, Table 1)

Baseline	73·3%
Recall	67·9%
Precision	74·5%

Table 2.6: Performance of EDMONDS-COLLOCATE on 2103 WordNet synsets (Edmonds, 1999, p 93)

Table 2.6, however subsequent authors have returned to using the seven word sets shown in Table 2.5 in order to compare their system's performances.

Comparative performances of all systems discussed above are shown in Table 2.7 as reported by their respective authors. Where more than one parameter setting is described, the best performing value is given. Performance on this test set is unavailable for the anti-collocations method of Inkpen (2007b) described in section 2.4.2.3 and for the discriminative training method of Yu et al. (2010) described in section 2.4.2.8.

The anti-collations method was tested on the British National Corpus, and the discriminative training method on n-grams from Web 1T. Values for the latter are shown in Table 2.7 but are not directly comparable because of the different test set. Values for the former are reported in Inkpen (2007b, Table III) and are intermediate between the baseline and the unsupervised method described in section 2.4.2.4.

2.4.4 Related tasks

The task of finding an appropriate near-synonym for a piece of text has been posed in alternative ways to the FILL IN THE BLANKS (FITB) task, for example: a context-free identification task, where given a word its synonym must be selected from a list; and a replacement task instead where given sentence S, a system must find another word w_r that is a suitable replacement for an original word w_o in S.

The problem of solving synonym multiple choice problems of the style posed in stan-

Test set	No. cases	Accuracy as a percentage							
		Baseline §2.4.2.1	Edmonds §2.4.2.2	Inkpen (PMI) §2.4.2.4	Inkpen (SVM) §2.4.2.5	Language model §2.4.2.6	LSA §2.4.2.7	*DT*[a] §2.4.2.8	Two phase §2.4.2.9
difficult, *hard*, *tough*	6630	41·7	47·9	59·1	57·3	63·2	61·7	*63·1*	**69·4**
error, *mistake*, *oversight*	1052	30·9	48·9	61·5	70·8	78·7	82·5	*79·2*	**83·3**
job, *task*, *duty*	5506	70·2	68·9	73·3	**86·7**	78·2	82·4	*75·7*	85·5
responsibility, *burden*, *obligation*, *commitment*	3115	38·0	45·3	66·0	66·7	72·2	63·5	*69·6*	**77·9**
material, *stuff*, *substance*	1715	59·5	64·6	72·2	71·0	70·4	**78·5**	*75·1*	77·3
give, *provide*, *offer*	11504	36·7	48·6	52·7	56·1	55·8	**75·4**	*68·3*	60·1
settle, *resolve*	1594	37·0	69·5	76·9	75·8	70·8	**77·9**	*84·1*[b]	74·5
Average over all sentences	31116	44·9	53·5	61·7	65·2	65·3	—	—	**70·8**
Average over each group	31116	44·8	55·7	66·0	65·2	69·9	74·5	—	**75·4**

[a] Values in this column are not directly comparable with the others due to a different test set.
[b] If this value was directly comparable, it would be the highest in that row.
Entries in bold face are the highest in that row.
Unavailable figures are reported as —

Table 2.7: Performance of the FITB systems on the Wall Street Journal test corpus

dardised exams has been investigated many times.[7] This problem is that of, given a word and a small set of candidate synonyms, without any context, and only one of which is correct, to identify that correct synonym. One example is to identify a synonym for *hidden* among *laughable*, *veiled*, *ancient* and *revealed* (Turney et al., 2003). Presently the highest scoring system is that of Turney et al. (2003) combining weighted scores from several systems including one using Latent Semantic Analysis, one using point-wise mutual information scores and one using a thesaurus. The system as a whole achieved accuracy of 97.50% on a test set from several real tests of English language skill administered to second language speakers. Other high scoring systems include a Latent Semantic Analysis-based system with a score of 92.50% (Rapp, 2003) and a Generalised Latent Semantic Analysis-based system with a score of 86.25% (Matveeva et al., 2005).

The FITB task is also closely related to the lexical substitution task addressed at SemEval-2007 (McCarthy and Navigli, 2007). In this task, instead of being presented with a gap and being required to correctly replicate the author's original word choice from among several alternatives, the systems are given the author's original word choice and ask to choose a suitable alternative. Further, unlike in the FITB task, the set of alternative words is not fixed. The results of this harder task are not directly comparable to the FITB task, and the ability of systems to choose the very best alternative word had precision and recall of not greater than 13%. Baseline systems derived from walking the WordNet hierarchy around a target word performed at about 10% precision and recall whereas a distributional similarity baseline performed at under 9% precision and recall.

The methods used by the two best performing lexical substitution systems on the task requiring the best single substitute, as ranked by McCarthy and Navigli, were: a language model trained on 10^{12} words selecting from words drawn from all WordNet synsets of the target word and all neighbouring synsets (Yuret, 2007); and a voting system between several selection methods including a language model, a machine translation test and a Latent Semantic Analysis measure selecting words drawn from several sources including WordNet and Microsoft Encarta (Hassan et al., 2007). The major explanation for the considerably lower performance reported by the best systems on the lexical substitution task as opposed to the FITB task is presumably the number of substitutes that the lexical substitution systems consider, none of which is guaranteed to be the correct replacement. In the FITB task, typically systems are considering between 3 and 7 possible substitutes, and the correct replacement is guaranteed to be among them.

[7] There is a standard test set available for this problem. The relative performance of many published approaches to the multiple-choice synonym problem is tracked at http://aclweb.org/aclwiki/index.php?title=TOEFL_Synonym_Questions_%28State_of_the_art%29

2.5 Sentiment and subjectivity analysis

As seen above in Section 2.4, the statistical approaches to the FILL IN THE BLANKS (FITB) task do not typically investigate their performance on different near synonym meaning types discussed in Section 2.1.3. In this thesis we will examine one such distinction in particular, that of affect or sentiment differences between near-synonyms.

There is now a large body of research on the problems of SUBJECTIVITY ANALYSIS and SENTIMENT ANALYSIS which supports such an approach by applying natural language processing techniques to opinionated or affective documents for various tasks. These approaches both suggest that affective text requires specific approaches, a finding which may apply to the FITB task, and suggest some such approaches. We therefore discuss sentiment and subjectivity analysis techniques here. The reader is referred to the surveys of Pang and Lee (2008) and Liu (2012) for comprehensive overviews of sentiment and subjectivity analysis, now a very large field.

2.5.1 The sentiment and subjectivity analysis tasks

2.5.1.1 Overview of sentiment analysis

Sentiment analysis is the process of determining the sentiment expressed by a piece of writing towards its subject, for example, determining whether a movie review recommends or pans the reviewed movie. Sentiment analysis can be performed at different levels of granularity, for example individual words have been classified into 'positive' and 'negative' groups, and entire documents have been classified into positive and negative groups. The classification is also not necessarily polar, with some systems grouping documents into degrees of sentiment (Pang and Lee, 2005).

Turney and Littman (2003) divide the sentiment analysis field into three broad problems:

1. that of identifying sentimental features of individual words or phrases;
2. that of classifying documents into groups based on their positivity or negativity towards their subject; and
3. that of identifying the parts of any text that are subjective.

Uses proposed for sentiment analysis include: providing useful search results and summaries for users seeking opinions about travel destinations and other potential purchases (Turney, 2002); and in commercial systems providing feedback on the response to

a company's offerings (Pang et al., 2002). More recently its use in various applications have been proposed and tested. For example, Liu et al. (2007) attempt to detect the usefulness of user product reviews to other users, including using a sentiment polarity tool to contribute features to their classifier, although the polarity features ended up making no contribution. Attempts to predict the performance of products include Yano and Smith (2010), who attempted to predict the comment volume that blog posts would receive, using naïve bag of words models, regression and topic models and achieving F-scores of up to 65.9 in predicting the number of words left in comments and 57.8 in predicting number of comments; and Joshi et al. (2010), who attempted to predict the box-office performance of movies based on their reviews, achieving some improvement over previous approaches using only meta-data such as the movie's title, or reviewer-assigned review scores alone.

Most approaches to sentiment analysis and classification rely on machine learning techniques of various sorts. Sentiment analysis is additionally related to a significant amount of older work and focuses on categorising documents by style; for example: into different genres (Yang and Pedersen, 1997), into works by the same author (Koppel et al., 2003) or into works by authors of the same gender (Koppel et al., 2002). The justification for addressing it as a specific task rather than simply applying text categorisation techniques to sentiment analysis is that Pang et al. (2002) showed that classifying documents by sentiment is not the same problem as classifying them by genre: the most successful machine learning approaches to topic classification — Naive Bayes, maximum entropy classification and support vector machines — performed significantly worse on sentiment classification of movie reviews than they do on topic based classification in general.

Liu (2012) identifies two major new classes of problems: cross-domain sentiment analysis, which is difficult because sentiment prediction features and models are generally quite domain-specific and cross-language sentiment classification requiring models be able to be generalised over multiple languages.

2.5.1.2 Overview of subjectivity analysis

Wiebe (1990) introduced the OBJECTIVE and SUBJECTIVE distinction of Banfield (1982) to computational linguistics: "[o]bjective sentences are those that objectively narrate events… [s]ubjective sentences are those that present the consciousness of an experiencing character within the story." Wiebe (1994) identified subjectivity with PRIVATE STATES (Quirk et al., 1985), which cannot be verified by an observer but only reported by the person experiencing them.

The problem of identifying objectivity and subjectivity is related to, but not identical to

sentiment analysis: subjective sentences aren't necessarily affective and objective sentences aren't necessarily non-affective.

The problem of subjectivity identification has been pursued both for its own sake and in service of applications, for example, Riloff and Wiebe (2003) suggest several applications for being able to identify these parts of a document: allowing factual question answering systems to ignore them; including some of their content in automatically generated summaries or extended answers that are meant to reflect different opinions or perspectives; and identifying some potentially undesirable content in correspondence such as email "flames" or certain kinds of commercial email.

Subjectivity analysis and sentiment analysis often occur in tandem, with Pang and Lee (2004) finding that first identifying subjective parts of a document and then using only those to classify the sentiment of the document assists with classification performance.

2.5.2 Techniques and features used in sentiment and subjectivity analysis

In this section, we describe features that have been identified as useful for either or both the sentiment and subjectivity analysis problems, in order to inform our choice of features when attempting to predict near-synonym choice when the near-synonyms have affective meaning or are used in affective contexts.

2.5.2.1 Techniques and features used in document sentiment analysis

In this section we briefly describe approaches to the problem of classifying documents into groups based on their positivity or negativity towards their subject. The sentiment classification problem is the problem of classifying documents into groups based on sentiment. "Positive" and "negative" is the most common set of classifications, with identifying various degrees of positive and negative being the next most common.

The problem of automatically classifying words or short phrases into groups based on their inherent positivity or negativity is discussed in more detail in Section 2.5.3.

Pang et al. (2002) tested common supervised machine classification methods—naive Bayes, maximum entropy and support vector machines—with a variety of features including unigrams, bigrams and part of speech to classify movie reviews into positive and negative groups. They achieved accuracy of between 72% and 83%, which is well above the baseline performance, but below the performance of the same algorithms and features for topic based classification, which is around 90%. They concluded that sentiment analysis

was amenable to supervised learning approaches but that at the very least new features would need to be developed to make the performance match that of topic classification.

Turney (2002) achieved scores of up to 84% accuracy classifying reviews as having either positive or negative recommendations, based on mutual information between phrases in the review and the terms "excellent" and "poor". Turney discovered that movie reviews are significantly more difficult to classify using this method than bank or automobile reviews and hypothesised that the negative descriptions of movie characters and their actions (as, for example "evil" or "foolish") were being confused with judgements of the review.

Riloff et al. (2006) noticed that Pang et al. (2002) found that unigrams were more successful than many other features, and decided that reducing the number of features to include only the most successful features would achieve better results than including every conceivable feature. They used a feature hierarchy to capture features in a way that represented their overlap and their impact on performance. They discovered that reducing the features using their hierarchy improved support vector machine performance on several sentiment classification datasets by a small but statistically significant amount.

There are some thirty-odd supervised approaches to document sentiment classification cited by Liu (2012). Examples include Kennedy and Inkpen (2006), who considered VALENCE SHIFTERS such as negations and intensifiers, incorporating them into bigrams with other words and finding that they improved over a unigram approach; and Yessenalina et al. (2010), who isolate subjective sentences in order to improve classification at the document level, but who for the first time use document classification performance to evaluate the success of the subjectivity identification. Unsupervised approaches after that of Turney (2002) are less common, but recent examples include Taboada et al. (2011), who built the Semantic Orientation CALculator (SO-CAL) based on dictionaries annotated with sentiment orientation in both polarity and strength, and on identifying contextual clues such as negation. Dasgupta and Ng (2009) developed a semi-supervised technique, first identifying the easy-to-classify documents, and bootstrapping a distriminative learner on those documents and hand-labelled examples to classify more ambiguous documents.

Pang and Lee (2008) report that other features used in sentiment classification have included position information, higher order n-grams, and the amount of contrast between terms. Becker and Aharonson (2010) use psycholinguistic experiments to suggest focussing on the conclusion of the text. Mejova and Srinivasan (2011) experimented with feature definition and selection for sentiment polarity detection at the document level, exploring features including words versus stems, binary versus term frequency weights, n-grams of up to 3 and syntactic phrases. They found that certain more complex features—including

stemming and bigrams—do not improve performance, and that adjectives are an especially important feature to consider.

2.5.2.2 Subjectivity features

Spertus (1997) did some of the earliest commonly cited work on automatically identifying subjective pieces of text: in her case, entire email messages, seeking "flames". Spertus manually developed several rules, including for example rules that found obscenities and statements addressed in the second person like "you suck!" These rules were ordered using a C4.5 decision tree in order to produce weighted rules which were able to classify messages as flames or not flames with reasonable accuracy: it agreed with human judgement of non-flames 98% of the time and with human judgements of flames or probable flames 64% of the time.

Riloff and Wiebe (2003) hypothesised that many markers of subjectivity—including subjective adjectives like "unseemly" and metaphorical phrases like "dealt a blow"—occur relatively infrequently in text and therefore subjectivity classifiers need to be trained on large bodies of text to identify a useful number of subjectivity-related features. They had success bootstrapping a learning system with a number of patterns that mark subjective features and using the subjective expressions learned from those patterns their system found further patterns, giving it more examples to learn from. Using this method they increased the recall of their subjectivity identifier from 33% to 40%, with precision only dropping slightly from 91% to 90%.

Wiebe et al. (2004) explore the kinds of features that mark subjectivity. A particularly interesting finding is that subjective features tend to be low frequency: a token's uniqueness in the corpus is a good indicator of subjectivity. Subjective parts of text also tend to have high density subjective features, that is, they will be in the context of many markers of subjectivity rather than few.

Wilson et al. (2006) combine subjectivity detection and intensity detection: they explore not only finding subjective text but assigning a value to the intensity of opinion expressed in the identified fragments. Wilson et al. used parts of the MPQA Corpus as training and test sets. The work uses previously established subjectivity clues such as the use of verbs of judgement ("commend", "vilify") and polarised adjectives ("appealing", "brilliant") together with a novel set of syntactic clues as input to three machine learning algorithms: support vector regression yielded improvements of 51% over the baseline and boosting yielded improvements in accuracy of up to 96%.

More recent work has considered the value of CHARACTER n-GRAMS rather than lexical

n-grams—for example, consider string *feature* as the character 7-gram $\{f, e, a, t, u, r, e\}$—with Raaijmakers and Kraaij (2008) achieving state-of-the-art performance (accuracy of 85%) on classification of the MPQA corpus. Raaijmakers et al. (2008) further found that while character n-grams are the strongest individual feature, that combining them with features including word-level information and acoustic information improves performance.

2.5.2.3 Gold standard sentiment and subjectivity data

Evaluation of a particular sentiment or subjectivity classifier is typically against a gold standard. The most common example of a gold standard are those in which the author of the review has also assigned the subject an explicit rating; such as the 'star ratings' typically given in movie reviews. The sentiment polarity dataset consisting of movie reviews drawn from the World Wide Web as described in Pang and Lee (2004, 2005) are very widely used, despite, or possibly because of, the finding of Turney (2002) that movie reviews are significantly harder to classify by sentiment than reviews of either banks or automobiles. This corpus is described further in Section 2.5.4.1.

Evaluation of sentiment detection effectiveness can be performed against the Multi-Perspective Question Answering Opinion Corpus (MPQA Corpus) described by Wiebe et al. (2005), which annotates subjective parts of texts with both the sentiment expressed, the intensity of the sentiment, the intensity of the sentiment's expression and the contribution that expression makes to the text as a whole.

2.5.3 Sentiment analysis and lexical items

In addition to the sentiment and subjectivity work at the phrase, sentence and document level discussed in Section 2.5.2, sentiment analysis has also considered the question of the affective parts of the meaning of lexical items.

The feelings a word or phrase tends to convey towards a referent is referred to as that word's or phrase's SEMANTIC ORIENTATION. Most work in the area of semantic orientation identifies two orientations: positive towards the referent or negative towards it.

Typically work in this area aims to acquire the lexical semantics of sentiment: to identify semantic orientation that is fairly central to the meaning of the word or phrase in question. A common goal is to use such items as features in document classification as discussed in Section 2.5.2.1.

There is a considerable body of work on expanding sentiment lexicons from small seed sets of words. Andreevskaia and Bergler (2006) divide techniques used to extract

sentiment-bearing words into two broad categories: those that rely on co-occurrence in free-form text in order to derive similarity in sentiment, and those that rely on structured lexicons, typically WordNet. The best known unsupervised approach to this is that of Turney (2002) and Turney and Littman (2003), in which positive and negative words are learned through their collocation with other positive words in a document in a bootstrapping cycle. Variations of this approach include that of Gamon and Aue (2005), who incorporate anti-collocations with words of opposite sentiment, at a sentence level.

Zagibalov and Carroll developed extensions of these techniques for Chinese language sentiment analysis first using only a single hand-chosen word and then only a list of common negations and adverbials. Their seed words allow them to classify document sentiment with accuracy as high as an F-measure of 92% (Zagibalov and Carroll, 2008b,a).

Kamps and Marx (2002) describe an approach to a slightly different and prior problem: that of identifying affective or sentimental adjectives in the first place. They use the synonymy relations given in WordNet (Fellbaum, 1998) to construct paths of synonymy between words: for example there is a synonym path of length 2 between *good* and *proper* in WordNet (*good* is a synonym of *right*, and another sense of *right* is a synonym of *proper*). This is unhelpful in measuring the orientation of words, as antonyms tend to be closely associated in WordNet, but they did identify a cluster of 5410 which are associated with the words *good* or *bad* and hypothesise that this set is essentially the affective adjectives contained in WordNet. Kamps et al. (2004) further explore this idea although neither paper directly tests this claim. Kamps et al. do test a sketched idea from Kamps and Marx, which is that words that are closer to *bad* are negative and words that are closer to *good* are positive, and similarly that words that are closer to *strong* than to *weak* are more potent and so on. They achieve accuracy of between 61% for words on the activity axis and 71% for words on the potency axis as measured in the General Inquirer (Stone et al., 1966) corpus.

Hatzivassiloglou and McKeown (1997) identified adjectives with positive or negative nuances by looking for adjectives that occur in conjunctive relationships joined by *and* or *but*. For example, a proposal might be *simple and well-received*, but is unlikely to be *simple but well-received*, since both *simple* and *well-received* are intended to praise the proposal. Since different conjunctions exhibit different likelihoods of coordinating two adjectives with the same orientation, they applied a log linear regression model to rate the chances of any two adjectives having the same semantic orientation and then divided them into two groups: negative and positive. Their method achieved precision of over 90% on parts of their dataset as compared to human annotators separately marking the

adjectives as positive and negative.

Turney (2002) developed a method for determining the semantic orientation of phrases rather than of individual words. He first used the Brill part-of-speech tagger to extract two word phrases from the documents that fell into five particular patterns of part of speech, for example adjectives followed by nouns and two adjectives not followed by a noun. He then calculated the point-wise mutual information scores between the two word phrases and the words *excellent* and *poor* respectively. Words with a greater mutual information score with *excellent* were considered to have a positive orientation and words with a greater mutual information score with *poor* a negative one. Turney does not evaluate the effectiveness of this classification directly but instead uses it as input to document classification.

Takamura et al. (2006) attempt to determine the semantic orientation of two word adjective-noun phrases, noting in particular the orientation of a phrase is not the sum of the orientation of the words in it. A "light laptop-computer" is a positive phrase but *light* and *laptop-computer* are neutral. Further *low mortality* is a positive phrase despite *low* and *mortality* being negative words in isolation. Their method is intended to capture the notion that if, for example, *low risk* is known to be a positive phrase, and *risk* and *mortality* belong to the same cluster, then *low mortality* is also likely to be a positive phrase, despite the negativity of its component words. They introduce latent variables into the models and tested with Japanese noun and adjective pairs, achieving an accuracy rating of nearly 82% against human annotated data.

Wilson et al. (2005) consider another more complex problem: context dependent orientation of phrases. For example, in the phrase "there is no reason at all to believe..." the normally positive word *reason*, being negated, adds negativity to the sentence. In addition to negation, intensifiers and diminishers alter the sentiment and words that might be expected to have an effect may not when located in multi-word expressions or named entity expressions (eg *trust* in *National Environmental Trust*). Using the BoosTexter AdaBoost.HM machine learning algorithm they achieve 73–76% accuracy in deciding whether a phrase is neutral or affective, and 61–66% accuracy in deciding whether a phrase is positively or negatively affective.

Wiebe and Mihalcea (2006) explore yet another problem: the association of sentiment with particular word senses, rather than with words in general. So, for example, they want to capture the difference between *alarm* as in "fear resulting from the perception of danger", which is emotionally charged, and *alarm* as in "warning device", which is emotionally neutral. Given an unknown word w with multiple senses, they compare the words most distributionally similar to w, and then compare these distributionally similar

	[Maas et al.] model Sentiment + Semantic	[Maas et al.] model Semantic only	LSA
melancholy	*bittersweet*	*thoughtful*	*poetic*
	heartbreaking	*warmth*	*lyrical*
	happiness	*layer*	*poetry*
	tenderness	*gentle*	*profound*
	compassionate	*loneliness*	*vivid*
ghastly	*embarrassingly*	*predators*	*hideous*
	trite	*hideous*	*inept*
	laughably	*tube*	*severely*
	atrocious	*baffled*	*grotesque*
	appalling	*smack*	*unsuspecting*

Table 2.8: The five most similar words to *melancholy* and *ghastly* using the vectors learned by Maas et al., compared to Latent Semantic Analysis (LSA) (Maas et al., 2011, Table 1)

words with the distribution of each sense $s_1 \dots s_n$ of w to arrive at a subjectivity score based on the closest matching sense. Their method improves significantly on a baseline assignment of subjectivity to words in a particular context.

Earlier work in the area of classifying individual words or short phrases very much concentrated on polar classification of the phrases into positive and negative phrases, without attempting to further position, say, the positive phrases in relation to each other as may be required by a language generation system trying to choose amongst a number of possible positive descriptions of an object. Quantified representations of sentiment are explored by some more recent work, including Maas et al. (2011) who learn vector-based representations of the sentiment of words and who can then compute the most similar words in either sentiment or semantics or both as shown in Table 2.8; and Yessenalina and Cardie (2011) who model the compositional effects of individual words on phrase-level sentiment, for example, that *very* will increase the intensity of the sentiment of *good* in the phrase *very good*.

There are difficulties with defining the sentiment lexicon task in the first place. Andreevskaia and Bergler (2006) observe that sentiment is a fuzzy category, in which words such as *good* and *bad* are central to the category and definitively sentiment-bearing, whereas for words on the periphery of the category there is room for interpretational differences and high inter-annotator agreement might not be expected. They argue, in particular, that the fact that the General Inquirer and Hatzivassiloglou and McKeown (1997) word sets agree on the correct sentiment tagging of words only 78.7% of the time suggests that there may be legitimate natural variability between annotators.

This notion of centricity lead Andreevskaia and Bergler to create a different evaluation problem than a simple sentiment-bearing or not assessment. Andreevskaia and Bergler

note that depending on the seed words chosen, their system returned widely differing results, and used this to evaluate centricity, with sets being returned many times, particularly with the same sentiment each time, being regarded as central to the category. Thus the overall accuracy of classification when compared to the General Inquirer as a gold standard was 66.5%, but when limited to highly central words, it was 90%.

In our case, we would like to draw on such resources to evaluate the comparative affect of words: is *bad* less or more negative than *evil*, in order to guide word choice. We therefore require word lists that distinguish the strength of sentiment, rather than simply identifying it, or providing information only about its usefulness as a feature in document classification.

2.5.4 Sentiment corpora and knowledge bases

In this section we describe three existing sentiment corpora and knowledge bases that we draw upon in Chapters 3 to 5: the Scale dataset v1.0 movie review data set (Scale 1.0) corpus and the MicroWNOP lexicon. We also discuss the SentiWordNet lexicon, which we consider and decide not to use.

2.5.4.1 Scale 1.0 corpus

The the Scale dataset v1.0 movie review data set (Scale 1.0) (Pang and Lee, 2005)[8] is a set of 5000 short movie reviews by four authors on the World Wide Web. Each movie review is accompanied by both a three and four degree sentiment rating (that is, a rating on a scale of 0 to 2, and on a scale of 0 to 3) together with original rating assigned by the author to their own review on a scale of 0 to 10. Text is segmented and lower-cased in the corpus, and HTML stripped out.

A resulting review is shown in Figure 2.2. This review was rated 2 by its author on the original scale, and automatically is assigned 0 on both the 3 and 4-class transformations of the scale.

2.5.4.2 SentiWordNet

A lexical database of sentiment we considered and ultimately rejected using in this thesis is SentiWordNet (Esuli and Sebastiani, 2006)[9]; a sentiment-annotated set of WordNet synsets. Each synonym set is given three numbers, representing positive, negative and neutral meanings, which sum to 1.0.

[8] Available from http://www.cs.cornell.edu/people/pabo/movie-review-data/
[9] Available from http://sentiwordnet.isti.cnr.it/

```
all right , i'll admit it : i'm guilty of a bias towards writers . perhaps
it's the part of me that identifies with them , or the romantic image of
a guy alone in a room staring down a blank sheet of paper in a battered
manual smith-corona until the paper blinks first . the fact is that when
i feel a project has gone sour , the writer is frequently the last target
of my wrath , and the first one to get the benefit of the doubt . not this
time . i can't think of a director alive who could have done anything with
cops & robbersons , or an actor who could have made it watchabel . to call
this script inept would be to give it more credit than it deserves . it
is exceedingly difficult to imagine what bernie somers was thinking when he
started putting this mess together . there isn't a single original situation
or character to be found , but that can be said of eight scripts out of ten
that come out of hollywood . the more telling point is that somers didn't
seem to have a clue where to go with this story . the predictable resolution
would have been for milquetoast norman to discover an inner reserve of
strength and end up saving the day . but even that simplistic an answer
seems to be beyond his grasp . norman begins and ends the film exactly the
same , and there isn't even a pat answer provided for why there is anything
different in the way his family would treat him . please , my heart can't
take any more surprises . the biggest losers , however , are probably the
two names above the title . still , it would have been nice if he hadn't
walked through the film as though he were still stunned by the cancellation
of his talk show . he's not even as endearingly incompetent as the vacation
series' clark griswold , just deathly boring . jack palance doesn't fare
much better , but he does have an innate gruff charm that transcends the role
as written ; however , that still doesn't explain why the robbersons seem so
instantly drawn to jake despite his abusiveness , nor why jake eventually
reciprocates . even at bargain prices , that's two bucks a laugh . * that's
* a crime .
```

Figure 2.2: An example review from the SCALE 1.0 corpus (Pang and Lee, 2005)

The numbers were originally automatically derived by having eight separate classifiers make a decision about the status of the synset, each classifier contributes a score of 0.125 for its "vote". Thus, a synset where one classifier had classified it positively, one negatively, and six neutrally would have a positivity score of 0.125, a negativity score of 0.125 and a neutrality score of 0.75. The features used to train the classifiers are WordNet relationships such as hypernymy and antonymy. As an example, the *slaphappy*, *silly* and *punch-drunk* synset has a positive score of 0, a negativity score of 0.25 and a neutrality score of 0.75. There are no intra-synset differences in score.

In this thesis we evaluated SentiWordNet 1.0, the 3.0 version has more recently been released (Baccianella et al., 2010) and the database is now periodically hand corrected with improved values. Further discussion of our evaluation of SentiWordNet 1.0 is found in Section 4.1.2, it was also used as a source of data by Whitehead and Cavedon (2010) with indifferent results, as discussed in Section 2.6.1.

2.5.4.3 MicroWNOP

MicroWNOP is a subset of WordNet annotated with polarity information (Cerini et al., 2007)[10]. Unlike SentiWordNet, the scores are assigned by human judges rather than by an automatic process. There are three sub-corpora in MicroWNOP:

the *Common* group with one Positive and one Negative score per synset, as all five annotators worked collaboratively;

the *Group1* group with three Positive and three Negative scores per synset, each assigned by one of three annotators; and

the *Group2* group with two Positive and two Negative scores per synset, each assigned independently by the remaining two annotators.

An example entry from *Group1* is *baseborn*, *humble*, and *lowly*, with all three annotators assigned it 0 for positivity, and the first two assigning 0.5 for negativity. The third annotator assigned 0.25 for negativity. As with SentiWordNet, the entire synset is assigned the score, not any individual words within it.

[10] Available from http://www-3.unipv.it/wnop/

2.5.4.4 Other lexicons

There are a large number of existing affective lexicons[11]. Many do not annotated the *degree* or *strength* of sentiment, only its polarity. Some of the best known, but not used here, are:

- the General Inquirer word lists, which list unscored words in certain categories including positive (1915 words) and negative (2291 words), developed for use in computer-assisted text analysis (Stone et al., 1966);
- WordNet Affect, which adds "A-labels" (affective labels) to WordNet synsets, such as adding a MOOD label to the synsets headed by both *animosity* and *amiable* (Strapparava and Valitutti, 2004); and
- the subjectivity lexicon that is part of the MPQA Opinion Corpus, again assigning terms to categories, in this case positive, negative, both or neutral, but not scoring the strength of their affective meaning, although this corpus does rate their effectiveness as a cue for subjectivity analysis (Wiebe et al., 2005; Wilson et al., 2005).

Su and Markert (2008b) provide a dataset of subjectivity and polarity annotated WordNet senses, per the annotation scheme described by Su and Markert (2008a). This dataset includes some synsets which contain internal subjectivity or sentiment differences, although the internal differences are not annotated. These are three synsets they regard as having both subjective and objective members:

(2.14) *beneficent*, *benevolent*, *eleemosynary*, *philanthropic* – generous in assistance to the poor; "a benevolent contributor"; "eleemosynary relief"; "philanthropic contributions"

(2.15) *need*, *demand* – a condition requiring relief; "she satisfied his need for affection"; "God has no need of men to accomplish His work"; "there is a demand for jobs"

(2.16) *antic*, *joke*, *prank*, *trick*, *caper*, *put-on* – a ludicrous or grotesque act done for fun and amusement

Su and Markert do not provide their word-by-word annotations.

[11] At time of writing a reasonably up-to-date list can be found at http://neuro.imm.dtu.dk/wiki/Text_sentiment_analysis

2.6 Valence-shifting text

In this section we describe work in the problem called VALENCE SHIFTING, that of rewriting a text to preserve much of its meaning but alter its sentiment characteristics. This is a task this thesis will ultimately address in Chapter 5. Guerini et al. (2008) describe valence shifting thus:

> While there is the active [Natural Language Processing (NLP)] field of opinion mining and sentiment analysis… on the other side, given the large amount of available texts, it would be conceivable to exploit NLP techniques to slant original writings toward specific biased orientation, keeping as much as possible the same meaning… as an element of a persuasive system. For instance a strategic planner may decide to intervene on a draft text with the goal of "coloring" it emotionally. When applied to a text, the changes invoked by a strategic level may be uniformly negative or positive; they can smooth all affective peaks; or they can be introduced in combination with deeper rhetorical structure analysis, resulting in different types of changes for key parts of the texts.

In Section 2.6.1 we review existing approaches to the problem of valence shifting; in Section 2.6.2 approaches to other problems of trying to re-write text in order to change the perceived author.

2.6.1 Valence-shifting existing text

Existing approaches to valence shifting most often draw upon lexical knowledge bases of some kind, whether custom-designed for the task or adapted to it. Existing results do not yet suggest a definitively successful approach to the task,

Inkpen et al. (2006) used several lexical knowledge bases, primarily the near-synonym usage guide *Choose the Right Word* (Hayakawa, 1994) and the General Inquirer word lists (Stone et al., 1966) to compile information about attitudinal words in order to shift the valence of text in a particular direction which they referred to as making "more-negative" or "more-positive". They estimated the original valence of the text simply by summing over individual words it contained, and modified it by changing near synonyms in it allowing for certain other constraints, notably collocational ones. Only a very small evaluation was performed involving three paragraphs of changed text, the results of which suggested that agreement between human judges on this task might not be high. They generated more-

(*hideous*, *ugly*, *unnatural*) $\longleftrightarrow$ (*pretty*, *beautiful*, *gorgeous*)
(*disgusting*, *mediocre*, *tasty*, *delicious*, *exquisite*)

Figure 2.3: Examples of Ordered Vectors of Valenced Terms from Guerini et al. (2011)

positive and more-negative versions of paragraphs from the British National corpus and performed a test asking human judges to compare the two paragraphs, with the result that the system's more-positive paragraphs were agreed to be so three times out of nine tests (with a further four found to be equal in positivity), and the more-negative paragraphs found to be so only twice in nine tests (with a further three found to be equal).

The VALENTINO tool (Guerini et al., 2008, 2011) is designed as a pluggable component of a natural language generation system which provides valence shifting. In its initial implementation it employs three strategies, based on strategies employed by human subjects: modifying single wordings; paraphrasing, and deleting or inserting sentiment charged modifiers. VALENTINO's strategies are based on part-of-speech matching and are fairly simple, but the authors are convinced by its performance.

VALENTINO relies on a knowledge base of Ordered Vectors of Valenced Terms (OVVTs), with graduated sentiment within an OVVT, two examples from Guerini et al. (2011) are shown in Figure 2.3. Substitutions in the desired direction are then made from the OVVTs, together with other strategies such as inserting or removing modifiers. Example output given input of example (2.17) is shown in the more positive example (2.18) and the less positive example (2.19):

(2.17) We ate *a very good dish*.

(2.18) We ate *an incredibly delicious dish*.

(2.19) We ate *a good dish*.

Guerini et al. are presenting preliminary results and appear to be relying on inspection for evaluation: certainly figures for the findings of external human judges are not supplied. In addition, some of the examples of output they supply have poor fluency:

(2.20) * Bob *openly admitted* that John is *highly* the *redeemingest signor*.

(2.21) * Bob *admitted* that John is *highly a well-behaved sir*.

Whitehead and Cavedon (2010) reimplement the lexical substitution, as opposed to paraphrasing, ideas in the VALENTINO implementation, noting and attempting to address two problems with it: the use of unconventional or rare words (*beau*), and the use of grammatically incorrect substitutions as above.

Even when introducing grammatical relation-based and several bigram-based measures of acceptability, they found that a large number of unacceptable sentences were generated. Categories of remaining error they discuss are:

large shifts in meaning for example by substituting *sleeper* for *winner*, accounting for 49% of identified errors;

incorrect word sense disambiguation accounting for 27% of identified errors;

incorrect substitution into phrases or metaphors such as *long term* and *stepping stone*, accounting for 20% of identified errors; and

grammatical errors such as those shown in examples (2.22) to (2.24), accounting for 4% of identified errors.

(2.22) Williams was not *interested* (in) girls.

(2.23) Williams was not *concerned* (with) girls.

(2.24) Williams was not *fascinated* (by) girls.

Whitehead and Cavedon also found that their system did not perform well when evaluated. Human judges had low, although significant, agreement with each other about the sentiment of a sentence but not significant agreement with their system's output. They particularly judge that the corpus SentiWordNet (Esuli and Sebastiani, 2006) may be a useful resource in other tasks, but is a poor resource for this task.

This task therefore appears more difficult than researchers originally anticipated, with many factors making assessment difficult, not least the requirement to be successful at a number of different NLG tasks, such as producing grammatical output, in order to properly evaluate success.

2.6.2 Approaches to re-writing text to obscure the author

In addition to previous approaches to the valence shifting problem, there have been previous attempts to automatically re-write text in such a way as to change one or more of its properties, or its category as assigned by a classifier. One prominent example is that of changing the authorship attribution of the text.

Authorship attribution is one of the most prominent applications of stylometry (see Holmes (1998) for an historical overview), and there has been some research on deceiving stylometry, usually from the point of view of preserving an author's anonymity against

unmasking attempts; for example this was explored as a privacy problem by Rao and Rohatgi (2000).

Kacmarcik and Gamon (2006) describe an experiment in changing a document so as to change its perceived author. Their method is an exploratory attempt to present authors with the most identifying features—word choices in their experiment—that they have made. In what they call SHALLOW OBFUSCATION word choices were ranked by a Decision Tree Root classifier and the most discriminative word choices presented for editing. Testing altered feature vectors against a Support Vector Machine (SVM) model trained to identify authors of the Federalist Papers, they achieved a reduction in correct identifications of 74.66% on the most resistant SVM.

Kacmarcik and Gamon also discuss an alternative, DEEP OBFUSCATION, designed to defeat the multiple rounds of classification that Koppel and Schler (2004) use. Koppel and Schler use multiple rounds of classification in order to identify and eliminate the most discriminating features between any two documents, the insight being that two documents by the same author should have relatively few discriminating features and thus cross-validation accuracy should drop quickly.

Brennan and Greenstadt (2009) describe two broad classes of attempts to deceive automatic authorship attribution: OBFUSCATION, in which a document is altered to remove features identifying its original author; and IMITATION, in which a document is altered to include features suggestive of another author. They had success with having human authors intentionally deceive automatic authorship attribution, particularly where there are more than 2 authors in the set to be distinguished.

In summary, the field of author obfuscation is also somewhat immature, and hence there are not a large number of approaches that could be tested for valence shifting.

2.7 Conclusion

In this chapter we have reviewed the general space of lexical semantic analysis of synonymy; discussed the evaluation of NLG systems; and computational approaches to semantics with particular reference to problems of word and phrase meaning in order to provide context for the problems of near-synonym choice and valence shifting. We have also extensively reviewed the FILL IN THE BLANKS (FITB) task, including existing approaches and their performance; reviewed techniques and data used in sentiment and subjectivity analysis; and reviewed existing approaches to the valence-shifting problem.

We particularly observe that there is more scope for fine-grained analysis of the ways in

which different near-synonym sets perform in the FITB task; that the mature research field of sentiment analysis offers several techniques that could possibly be exploited for near-synonym choice and for valence shifting; and that the valence-shifting problem as discussed by the literature to date appears to be a difficult problem which warrants returning to basics in order to empirically identify the approaches that result in clear valence shifts.

Chapter 3

Sentiment differences in near-synonym choice

In Chapter 2 we saw that the problem of choosing the correct lexical item in NLG systems is an ongoing problem, and in particular that correctly predicting a near synonym for a lexical gap, the so-called Fill In the Blanks (FITB) task, is an ongoing proxy for this problem.

However, when reviewing specific approaches to the FITB task in Section 2.4 we saw that while there are in recent years many statistical approaches to this problem, none of them so far attempt to analyse whether their performance differs on the various axes of near-synonym type given by authors including DiMarco et al. (1993); Edmonds and Hirst (2002); Inkpen and Hirst (2006) described in Section 2.1.3. In the specific case of near synonyms which have, or differ in, affective meaning in particular, our intuition is that because many documents cohere in sentiment, the markers for the correct choice of near synonym may be found throughout the document as shown by work on document-based sentiment analysis. We thus hypothesise that choosing among attitudinal near synonyms may be especially responsive to corpus-based methods.

In Section 2.5 we saw that independently the very large space of sentiment analysis has explored the sentiment of lexical items extensively. It is therefore possible that some of these properties could be exploited for the FITB task and other lexical choice tasks. In this chapter, we therefore assess in particular the relative performance of near synonyms that have differences in the sentiment or affect of the near synonyms, compared to those that do not differ in sentiment.

The sentiment-differences or no-sentiment-difference axis between near synonyms ac-

tually crosses from of the boundaries between near-synonym types given in Section 2.1. As we saw there, the difference in meaning between near synonyms still broadly collapses into the sense-reference/connotation-denotation distinction which has its origins with Frege (1892). But if we consider near synonyms which differ in sentiment from one another, we see some differences which could be described as denotational, and some connotational.

Consider the near synonyms given by Edmonds (1997), *error*, *mistake* and *oversight*. These differ in sentiment, in that *oversight* is the least negative of the three, and they do so largely by means of different denotations. *Oversight* makes a truth statement that the perpetrator of the oversight passively made a trivial and minor error, whereas *mistake* makes a statement about the world in which the perpetrator made some more serious and preventable error, perhaps more actively.

Compare *slim* and *skinny* where the difference in sentiment is largely in the matter of connotation: the truth values of many sentences such as those in examples (3.1) and (3.2) may be identical for most individuals[1] so described:

(3.1) Suzy is quite *slim*.

(3.2) Suzy is quite *skinny*.

While there are important sub-classifications of sentiment differences given in the literature, particularly the intersection with the subjectivity and objectivity distinction discussed in Section 2.5.1.2, we therefore do not consider the connotation-denotation distinction in this chapter, instead concentrating on the sentiment differences or no sentiment differences distinction between different groups of near synonyms.

In this chapter[2], our hypothesis is that, when using statistical approaches to the FITB problem, predicting the correct choice of near synonyms when the near synonyms in question have sentiment differences has systemically different success rates than when the same approaches are used to predict the correct near-synonym from among near synonyms that have no such sentiment differences from each other. We show that there is evidence that this difference between near synonyms that differ in sentiment and those that do not so differ does affect the performance of statistical approaches to the FITB task and therefore that this distinction may require that the approaches to FITB consider more carefully the type of difference between the near synonyms in a set, at least whether or not it is a sentiment difference.

[1] Probably not all individuals: *skinny* is at best *much* more likely to be used to describe someone judged to be in poor health or extremely underweight. So some difference in sentiment may still lie in denotation.

[2] Work described in this chapter was previously published in Gardiner and Dras (2007a,b).

In Section 3.1 we review the near-synonym choice problem posed by Edmonds (1997) and two existing approaches to it; in Section 3.2 we describe a test set of near synonyms we developed which distinguish between those with attitudinal meaning and those without; in Section 3.3 we describe the performance of these synonyms on the FITB task using the methods described in Section 3.1; in Section 3.4 we discuss the performance particularly as it relates to our hypothesis; and in Section 3.5 we draw some preliminary conclusions allowing us to continue to investigate our hypothesis in Chapter 4.

3.1 The FITB task, and experimental approaches

In this section, we briefly review the Fill In the Blanks (FITB) task introduced by Edmonds (1997) and two of the existing approaches to it, originally discussed in detail in Section 2.4. These are here reviewed in Section 3.1.1. We then introduce a variant of one of those approaches to allow the Google Web 1T corpus to serve as a data source in Section 3.1.2.

We will use our reimplementation of these statistical approaches to the FITB problem to test the hypothesis outlined above, that predicting the correct choice of near synonyms when the near synonyms in question have sentiment differences has different success rates to when the near synonyms in question have no sentiment differences. In this section we establish that we have acceptably approximated the approaches of Edmonds (1997, 1999) and of Inkpen (2007b) to the FITB task, building towards testing their comparitive performance on near synonyms with sentiment differences in Section 3.3.

3.1.1 The Fill In the Blanks (FITB) near-synonym choice task, and the Edmonds-Collocate and Inkpen-PMI approaches to it

As discussed in Section 2.4.1, the Fill In the Blanks (FITB) task was proposed by Edmonds (1997) and is to fill a gap in a sentence from a given set of near synonyms. He gives this example of asking the system to choose which of *error*, *mistake* or *oversight* fits into the gap in this sentence, originally introduced by Edmonds (1997) and here in Chapter 1:

(3.3) However, such a move would also run the risk of cutting deeply into U.S. economic growth, which is why some economists think it would be a big _____.

Performance on the task is measured by comparing system performance against real word choices: that is, sentences such as example (3.3) are drawn from real text, a word is removed, and the system is asked to choose between that word and all of its near synonyms as candidates to fill the gap.

The baseline approach suggested by Edmonds (1997) and introduced here in Section 2.4.2.1 is that of choosing the most frequent near synonym to fill every gap. In the case of example (3.3), the system would count the occurrence of *error*, *mistake* or *oversight* in a corpus, and predict the one with the highest frequency count every time the problem was posed for that near-synonym set.

The approach to this problem developed by Edmonds (1997, 1999) and here called EDMONDS-COLLOCATE was described in Section 2.4.2.2. In this approach, in order to choose between candidate tokens $c_1, \ldots, c_n$ to fill a gap g in a sentence S, each candidate c_i for the gap is scored not only by co-occurrence metrics with the words found co-occurring with c_i in a training corpus, but also with words in turn co-occurring with *those* words in a co-occurrence network. Where the EDMONDS-COLLOCATE method is unable to make a prediction or the two top candidates are tied, we fall back to the most frequent baseline in order to have predictions for all gaps.

A second, unsupervised, approach to this problem was developed by Inkpen (2007b), and is here called INKPEN-PMI. It was described in Section 2.4.2.4. In this approach, the suitability of candidate c for a given gap is approximated differently from EDMONDS-COLLOCATE: the entire sentence is not used to measure the suitability of the word. Instead, a certain sized window of k words either side of the gap is used. For example, if $k = 3$, the word missing from the sentence in example (2.12) is predicted using only the six words shown in example (2.13). The suitability of a candidate is given by the sum over its point-wise mutual information scores acquired from a surrounding context. INKPEN-PMI outperformed both the baseline and EDMONDS-COLLOCATE by 22 and 10 percentage points respectively on the seven synsets from (Edmonds, 1997) shown in Section 2.4.3.

3.1.2 Web1T-PMI, our variation of Inkpen's approach

WEB1T-PMI, our variation on INKPEN-PMI, is designed to estimate PMI(x, y), the point-wise mutual information of words x and y, using the Web 1T 5-gram corpus Version 1 (Brants and Franz, 2006). An approximation to the method is required without access to the Waterloo Multitext System and the corpus used by Inkpen. We can then substitute this value into Inkpen's scoring mechanism for the suitability s of a candidate c for a gap

n	n-gram	Web 1T count
3	`means official and`	41
5	`Valley National Park 1948 Art`	51

Table 3.1: An example 3-gram and 5-gram from Google Web 1T

g:

$$\text{PMI}(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \tag{2.6}$$

$$s(c, g) = \sum_{j=1}^{k} \text{PMI}(c, w_j) + \sum_{j=k+1}^{2k} \text{PMI}(w_j, c) \tag{2.7}$$

Web 1T contains n-gram frequency counts, up to and including 5-grams, as they occur in a trillion words of World Wide Web text. There is no context information beyond the n-gram boundaries. Examples of a 3-gram and a 5-gram and their respective counts from Web 1T are shown in Table 3.1.

These n-gram counts allow us to estimate the occurrence of x and y within a query frame k by summing the Web 1T counts of k-grams in which words x and y occur and x is followed by y. Counts are computed using the Web 1T processing software GET 1T detailed in Hawker et al. (2007). Queries are matched case-insensitively, but no stemming takes place, and there is no deeper analysis (such as part of speech matching).

This gives us the following methodology for a given lexical gap g and a window of k words either side of the gap:

1. for every candidate near synonym c:
   (a) for every word w_i in the set of words preceding the gap, $w_1, \ldots, w_k$, calculate $\text{PMI}(w_i, c)$, given counts for occurrences of w_i, c and w_i and c within a query frame from Web 1T;[3]
   (b) for every word w_j in the set of words following the gap, $w_{k+1}, \ldots, w_{2k}$, calculate $PMI(c, w_j)$ as above;
   (c) compute the suitability score $s(c, g)$ of candidate c as given by equation (2.7);
2. select the candidate near synonym with the highest suitability score for the gap where a single such candidate exists;

[3]Where the counts are 0 for the purpose of computing $s(c, g)$, we define $PMI(x, y) = 0$ so that it has no influence on the score $s(c, g)$ given by equation (2.7).

3. where there is no single candidate with a highest suitability score, select the most frequent candidate for the gap (that is, fall back to the baseline described in Section 2.4.2.1).[4]

Since Web 1T contains 5-gram counts, we can use query frame sizes from $q = 1$ (words x and y must be adjacent, that is, occur in the 2-gram counts) to $q = 4$.

3.1.2.1 Effectiveness of the Web1T-PMI approximation to Inkpen-PMI

In this section we compare Web1T-PMI as described in Section 3.1.2 with Inkpen-PMI as described in Section 2.4.2.4. This will allow us to determine whether our approximation is effective enough to allow us to compare attitudinal and non-attitudinal near synonyms.

In order to compare the two methods, we use five sets of near synonyms, also used as test sets by both Edmonds and Inkpen, the nouns and adjectives from Table 2.5:

- the adjectives *difficult*, *hard* and *tough*;
- the nouns *error*, *mistake* and *oversight*;
- the nouns *job*, *task* and *duty*;
- the nouns *responsibility*, *burden*, *obligation* and *commitment*; and
- the nouns *material*, *stuff* and *substance*.

We do not use the two sets of verbs as our attitudinal and non-attitudinal data (described in Section 3.2) does not include annotated verbs. We are therefore interested in the predictive power of Web1T-PMI compared to Edmonds-Collocate and Inkpen-PMI on adjectives and nouns.

Table 3.2 shows the performance of Edmonds-Collocate and Inkpen-PMI as given in (Inkpen, 2007b)[5] and Web1T-PMI on each of the test sets described above. Note that Inkpen reports different baseline results from us—we have not been able to reproduce her baselines. For the most part these are small, with the biggest difference for the set *error*, *mistake* and *oversight* at 16%, see columns 2 and 6 of Table 3.2. This may be due to choosing different part of speech tags: we simply used `JJ` for adjectives and `NN` for nouns.

Inkpen-PMI's improvements over the baseline for the test synsets given above were between +3.2% and +30.6%. The performance of Web1T-PMI is roughly comparable,

[4] Typically, in this case, all candidates have scored 0.

[5] Inkpen actually gives two methods, one using PMI estimates from document counts, one using PMI estimates using word counts. Here we are discussing her word count method and use those values in our table.

Set	Inkpen-PMI baseline value %	Edmonds-Collocate increase over baseline %	Inkpen-PMI increase over baseline ($q = 5$) %	No. test sentences we found	Web1T-PMI baseline value %	Web1T-PMI increase over baseline %	
						$q = 2$	$q = 4$
difficult etc.	41·7	+6·2	+17·4	5959	44·3	+15·3	+12·3
error etc.	30·9	+18·9	+30·6	1026	46·8	+25·5	+20·4
job etc.	70·2	−1·3	+3·2	4020	74·2	−14·4	−23·0
responsibility etc.	38·0	+7·3	+28·0	1119	36·7	+31·2	+24·9
material etc.	59·5	+5·1	+12·7	934	57·8	+5·5	−1·1

Table 3.2: Performance of Inkpen's test sentences on Edmonds's method, Inkpen's method and our method ($k = 2$)

with improvements as high as 31.2%. Further, WEB1T-PMI tends to improve especially largely over the baseline where INKPEN-PMI also does so: on the two sets *error* etc and *responsibility* etc..

The major anomaly when compared to INKPEN-PMI's performance is the set *job*, *task* and *duty*, where WEB1T-PMI performs very badly compared to both EDMONDS-COLLOCATE and INKPEN-PMI and the baseline (which perform similarly). WEB1T-PMI also performs under both methods on *material*, *stuff* and *substance*, although not as dramatically.

Overall, the fact that WEB1T-PMI tends to improve over EDMONDS-COLLOCATE where INKPEN-PMI also does so suggests that WEB1T-PMI takes advantage of the same aspects as INKPEN-PMI to gain improvements over EDMONDS-COLLOCATE, and thus that WEB1T-PMI is sufficiently different from EDMONDS-COLLOCATE to make a reasonable comparator and hence is a good candidate for use in our main experiment.

3.2 Near-synonymous test sets

For the FILL IN THE BLANKS (FITB) task, it is necessary to have candidate near synonyms, $w_1 \dots w_n$ where a word w_i chosen by a text's original author is compared with the other candidate words for being the best fit for the context.

In this section we first review the near-synonymous sets used by Edmonds (1997) in Section 3.2.1 and then in Section 3.2.2 sets of near-synonymous words we developed in order to test the performance of the experimental methods EDMONDS-COLLOCATE and WEB1T-PMI on words that differ in sentiment.

3.2.1 Edmonds' original near-synonymous test sets

Recall from Section 2.4.3 that the sets of synonyms Edmonds (1997) and most subsequent authors use to evaluate FITB performance are:

1. *difficult*, *hard* and *tough*
2. *error*, *mistake* and *oversight*
3. *job*, *task* and *duty*
4. *responsibility*, *commitment*, *obligation* and *burden*
5. *material*, *stuff* and *substance*

6. *give*, *provide* and *offer*; and

7. *settle* and *resolve*.

Edmonds observes that these differ in one respect relevant to his experiment: some sets contain a word with multiple senses — *job* for example — which have other senses aside from the sense which is nearly synonymous with other words in the set. Other sets do not contain polysemous words.

However, there are other ways in which the seven sets different from each other: the dimensions along which they differ in meaning. For example, *material*, *stuff* and *substance* differ somewhat in formality — *stuff* is less formal than its near synonyms — but also in denotational aspects, as both *material* and *substance* denote a more homogeneous thing than *stuff* does. *Difficult*, *hard* and *tough* differ less in denotation. *Error*, *mistake* and *oversight* are part of a set of near synonyms discussed extensively by Edmonds (1999) precisely because they differ dramatically in the speaker's attitude towards the agents responsible for the error: *error* implies some degree of blameworthiness, *oversight* less so and *mistake* less so again.

With only seven entries total, this test set is not large enough to establish if there is a difference in performance between affective near-synonym sets, and those without affective meaning.

3.2.2 Near-synonymous test sets that differ in sentiment

In this section, we describe an annotation experiment to develop a test set of near synonyms that differ in sentiment.

The most widely used corpus of near synonyms is WordNet (Fellbaum, 1998), introduced in Section 2.1.2.1. WordNet is a relatively fine-grained corpus: only words very near in meaning are assigned to the same set of synonyms ("synset"). For example, in Edmonds' assessment sets given in Sections 2.4.3 and 3.2.1 *oversight* belongs to a different synset from *error* and *mistake*, and *commitment* and *burden* are not found with *responsibility*, and *obligation*.

Edmonds and Kilgarriff (2002) describe some concerns with using WordNet in word sense disambiguation tasks (see Section 2.2.4): it focuses on similarities rather than differences between words, and Kilgarriff (2001) suggests that it does not provide enough information about sense distinctions to be useful in some word sense disambiguation tasks. Alternative corpora of near synonyms exist: for example, Navigli et al. (2007) prepared a coarse-grained corpus resulting from mapping WordNet onto a dictionary's encoding of

sense-distinctions for their SemEval-2007 task specifically due to the concerns of Edmonds and Kilgarriff; Inkpen and Hirst (2006) developed a corpus of near synonyms and ways in which they differ based on the definitions in Hayakawa (1994): *Choose the Right Word*, a handbook designed to explain (especially for non-native speakers of English) the nuances of near synonym choice.

As the Inkpen and Hirst corpus is unavailable for copyright reasons, in this work we have chosen to draw sets of near synonyms from WordNet.

3.2.2.1 Annotation method

We conducted an annotation experiment to provide a larger test set of near synonyms to test our hypothesis — that near synonyms with sentiment differences behave differently on FITB measures that statistical approaches — against. The annotation guidelines required a decision on whether selected WordNet synsets differed from each other mainly in attitude, or whether they differed in some other way.

The synsets were chosen from among the most frequent synsets found in the 1989 Wall Street Journal corpus. We identified the 300 most frequent WordNet 2.0 (Fellbaum, 1998) synsets in the 1989 Wall Street Journal using this frequency function, where $w_1 \ldots w_n$ are the words in the synset and $count(w_i)$ is the number of occurrences of w_i tagged with the desired part of speech in the 1989 Wall Street Journal:

$$\text{frequency}_{\text{synset}} = \sum_{i=1}^{n} \text{count}(w_i) \tag{3.1}$$

Synsets were then manually excluded from this set by the present author and her supervisor if they:

- contained only one word (for example *commercial* with the meaning "of the kind or quality used in commerce");
- contained a substantial number of words seen in previous, more frequent, synsets (for example the synset consisting of *position* and *place* was eliminated due to the presence of the more frequent synset consisting of *stead*, *position*, *place* and *lieu*);
- only occurred in a frozen idiom (for example *question* and *head* as in "the subject matter at issue"[6]);

[6]WordNet gives example usages of "the *question* of disease merits serious discussion" and "under the *head* of minor Roman poets", thus *head* in the sense of *heading*

- contained words that were extremely lexically similar to each other (for example, the synset consisting of *ad*, *advertisement*, *advertizement*, *advertising*, *advertizing* and *advert*); or
- contained purely dialectical variation (*lawyer* and *attorney*).

The aim of this pruning process is to exclude either synsets where there is no choice to be made (synsets that contain a single word); synsets where the results are likely to be very close to that of another synset (synsets that contain many of the same words); synsets where the words in them have very few contexts in which they are interchangeable (synsets used in frozen idioms) and synsets where there is likely to be only dialectical or house style reasons for choosing one word over another.

This left 124 synsets of the original 300. These synsets were then independently annotated by this author and her supervisor into two distinct sets:

1. synsets that differ primarily in attitude; and
2. synsets that differ primarily in some way other than attitude.

The annotation scheme allowed the annotators to express varying degrees of certainty:

1. that there was *definitely* a difference in attitude;
2. that there was *probably* a difference in attitude;
3. that they were *unsure* if there was a difference in attitude;
4. that there was *probably* not a difference in attitude; or
5. that there was *definitely* not a difference in attitude.

The 124 synsets and their annotations are shown in Appendix A.

The divisions into *definitely* and *probably* were only to allow a more detailed analysis of performance on the Edmonds experiment subsequent to the annotation experiment. The performance of attitudinal and not-attitudinal sets of synonyms were then compared using the Edmonds methodology.

3.2.2.2 Results and discussion

Inter-annotator agreement for the annotation experiment is shown in Table 3.3 both individually for annotations that the annotators felt were *definitely* correct and those that

		Annotator		
Difference	Certainty	1	2	Overlap between annotators
Attitude	Definite	14	18	7
	Probable	26	18	9
	Definite *or* Probable	40	36	29
Not attitude	Definite	68	63	51
	Probable	15	18	5
	Definite *or* Probable	83	81	73
Unsure		1	7	0

Table 3.3: Break-down of categories assigned in the annotation experiment

Category division	κ score	Agreement
Attitudinal, not attitudinal and unable to decide	0·62	82%
Annotations where both annotators were sure of their annotation	0·85	97%

Table 3.4: Inter-annotator agreement and κ scores for the annotation experiment

they thought were *probably* correct and collectively, for all annotations regardless of the annotator's certainty.

Two divisions of the annotation results were used to compute a κ score and raw inter-annotator agreement: the agreement between annotators on the "attitudinal difference", "not attitudinal difference" and "unsure" categories regardless of whether they marked their choice as *definite* or *probable*; and the agreement between annotators on *only* the annotations they were definitely sure about, as per Wiebe and Mihalcea (2006).

In fact, we calculated two different κ scores for each of the above: κ_{Co} assuming different distributions of probabilities among the annotators (Cohen, 1960); and $\kappa_{S\&C}$ assuming identical distributions among the annotators (Siegel and Castellan, 1988) as recommended by Di Eugenio and Glass (2004).

In general, κ values are computed using equation (3.2), where $P(A)$ is the observed agreement between annotators and $P(E)$ the probability of annotators agreeing by chance:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{3.2}$$

κ_{Co} and $\kappa_{S\&C}$ differ in their method of estimating $P(E)$: κ_{Co} computes $P(E)$ as the sum over the probability of annotator agreement on each individual category, as estimated by the actual agreement observed on that category, and $\kappa_{S\&C}$ as the sum over the probability of annotator agreement on each individual category as estimated by the number of times that category was observed over the entire dataset independent of annotator.

However the κ_{Co} and $\kappa_{S\&C}$ values were the same to two significant figures and are thus reported as a single value κ in Table 3.4. Raw inter-annotator agreement is also shown.[7]

The results suggest we can be fairly confident in using this classification scheme, particularly if restricted to the definite classes.

3.3 Attitudinal and non-attitudinal near synonym sets: comparison of FITB performance

In this section we evaluate the performance of EDMONDS-COLLOCATE and WEB1T-PMI on test data that contains near synonym sets that have affective meaning and sets that don't, in order to test our hypothesis stated at the start of this chapter, that that a corpus statistics approach is sensitive to the attitudinality of the near synonyms.

3.3.1 Test synsets and test sentences

3.3.1.1 Test synsets

Test synsets were chosen from the annotation task described in Section 3.2. Synsets were chosen where both annotators are certain of their label, and where both annotators have the same label. This results in 58 synsets in total: 7 where the annotators agreed that there was definitely an attitude difference between words in the synset, and 51 where the annotators agreed that there were definitely not attitude differences between the words in the synset. These 58 synsets are shown in Appendix A

An example of a synset agreed to have attitudinal differences was:

(3.4) *bad*, *insecure*, *risky*, *high-risk*, *speculative*

An example of synsets agreed to not have attitudinal differences was:

(3.5) *sphere*, *domain*, *area*, *orbit*, *field*, *arena*

The synsets are not used in their entirety, due to the differences in the number of words in each synset (compare {*violence*, *force*} with two members to {*arduous*, *backbreaking*, *gruelling*, *gruelling*, *hard*, *heavy*, *laborious*, *punishing*, *toilsome*} with nine, for example). Instead, a certain number n of words are selected from each synset (where $n \in \{3, 4\}$) based on the frequency count in the 1989 Wall Street Journal corpus. For example *hard*, *heavy*,

[7] The *Annotations where both annotators were sure of their annotation* figure in Table 3.4 is computed by excluding any question which one or both annotators marked as only *probably* belonging to one category or the other, or for which one or both annotators declared themselves unable to decide at all

Test synsets	No. test sentences		Baseline correctness (%)	
	Attitudinal	Non-attitudinal	Attitudinal	Non-attitudinal
TOP3	45953	353155	59.52	69·71
TOP4	48515	357290	56.37	68·91

Table 3.5: Number of test sentences and performance of the baseline for each set of test synsets

gruelling and *punishing* are the four most frequent words in the {*arduous*, *backbreaking*, *gruelling*, *gruelling*, *hard*, *heavy*, *laborious*, *punishing*, *toilsome*} synset, so when $n = 4$ those four words would be selected. When the synset's length is less than or equal to n, for example when $n = 4$ but the synset is {*violence*, *force*}, the entire synset is used. This was done to make the synsets all comparable to the (Edmonds, 1997) ones, which were all of length 2, 3 or 4.

These test sets are referred to as TOP3 (synsets reduced to 3 or fewer members) and TOP4 (synsets reduced to 4 or fewer members).

3.3.1.2 Test contexts

We performed this experiment, as Edmonds and Inkpen did, using the 1987 Wall Street Journal as a source of test sentences.[8] Wherever one of the words in a test set is found, it is removed from the context in which it occurs to generate a gap for the algorithm to fill. So, for example, when sentence (3.6) is found in the test data, the word *error* is removed from it and the system is asked to predict which of *error*, *mistake* or *oversight* fills the gap:

(3.6) ... his adversary's characterization of that minor sideshow as somehow a colossal *error* on the order of a World War....

Table 3.5 shows the number of attitudinal and non-attitudinal test sentences for both top3 and top4.

3.3.2 Results

In this section we compare the improvement that each of EDMONDS-COLLOCATE and WEB1T-PMI is able to make over the FITB baseline described in Section 2.4.2.1, for attitudinal and non-attitudinal near synonyms. Table 3.5 shows the baseline performance, where we already observe substantially better performance on non-attitudinal synsets,

[8] All references to the Wall Street Journal data used in this chapter refer to Charniak et al. (2000).

Synsets	q	k	EDMONDS-COLLOCATE improvement %		WEB1T-PMI improvement %	
			Att.	Non-att.	Att.	Non-att.
top3	2	2			−0·01	0·24
	4	2			−2·59	0·24
	–	2	0·88	−0·27		
	–	5	0·74	−0·29		
top4	2	2			−4·12	−1·31
	4	2			−5·55	−1·32
	–	2	0·36	−4·05		
	–	5	−1·91	−4·09		

Table 3.6: Improvement over the baseline for all test sentences

and Table 3.6 shows the difference in performance between the baseline and EDMONDS-COLLOCATE and the baseline and WEB1T-PMI.[9]

Results are shown with varying parameters for EDMONDS-COLLOCATE and for WEB1T-PMI: for EDMONDS-COLLOCATE, results are shown using both $k = 2$ (2 words either side of the gap) and $k = 5$ (5 words either side of the gap) to compute significance scores of candidates; and for WEB1T-PMI, we use one setting for k, $k = 2$ (window size of 2 words either side of the gap) and vary q (the corpus query window) between $q = 2$ and $q = 4$.

3.4 Discussion of the comparison between Edmonds-Collocate and Web1T-PMI

3.4.1 Comparative strength of Edmonds-Collocate

The first surprising result is that EDMONDS-COLLOCATE performs better than WEB1T-PMI overall, with WEB1T-PMI always worse than the baseline on our data. This is contrary to what we would expect from our comparison of the two methods in Section 3.1.2.1 which showed that WEB1T-PMI was generally the superior of the two methods, comparable to INKPEN-PMI, and similarly contrary to the conclusions of Inkpen (2007b). We suggest two possible ways of accounting for this.

It is not our purpose here to determine the best approach to the FITB task, it is to determine whether or not attitudinal near synonyms behave differently under these methods. However, one possibility for the comparative strength of EDMONDS-COLLOCATE is that WEB1T-PMI (and perhaps INKPEN-PMI also) performs especially well on the test

[9]The numbers given in Table 3.6 for EDMONDS-COLLOCATE are different from those given in Gardiner and Dras (2007a). We have since corrected a bug in the program that produces them. The general thrust of the results is the same, however.

set given in Section 2.4.3. This is possible, as Edmonds (1997) seems to have selected those sets as exemplars of near synonyms where a choice between them would be important in the context. However, it is also the case that INKPEN-PMI performs well when compared to the anti-collocations method, as discussed in Section 2.4.3, and that evaluation was performed on a different dataset. The second is that WEB1T-PMI is hurt by its much larger number of predictive attempts than EDMONDS-COLLOCATE. EDMONDS-COLLOCATE and WEB1T-PMI both fall back to the baseline where they are unable to produce a score for any of the candidate words. In the case of EDMONDS-COLLOCATE, this fall-back is used considerably more often than in the case of WEB1T-PMI: between 68.4% and 89.5% for various runs of EDMONDS-COLLOCATE as opposed to between 0.03% and 0.25% of the time for WEB1T-PMI. This difference in predictiveness is explained by the vastly different sized training sets of EDMONDS-COLLOCATE and WEB1T-PMI: about ten million words for EDMONDS-COLLOCATE and one trillion for WEB1T-PMI, but WEB1T-PMI seems to have lost more in accuracy than it gained in predictions.

It is possible also that Web 1T is not an ideal corpus for this problem: some researchers have informally expressed reservations about the representativeness of n-grams within it (Daumé, 2010). Other approaches to the FITB problem as summarised in Section 2.4 have used the Web 1T corpus with success but again, concentrating on the particular test set given in Section 2.4.3. Sub-par performance using Web 1T features or with web-derived n-grams has also been reported named entity recognition (Ramana et al., 2010). Islam and Inkpen (2009) reported success for real-world spelling correction using the Web 1T 3-grams, but that the data sparsity of the 5-grams meant that useful recall could not be obtained.

However, Web 1T continues to be used widely and successfully. Klein and Nelson (2009) found that Web 1T term counts had a very high correlation with term counts found in the Web as Corpus (WaC) (Baroni et al., 2009), greatly increasing their confidence that Web 1T can be used to accurately estimate INVERSE DOCUMENT FREQUENCY statistic (IDF) counts for the web. Work on error correction, which is closely related to lexical choice insofar as it is about detecting incorrect lexical choices and sometimes offering more correct alternatives, continues to rely heavily on Web 1T n-grams. The findings of Bergsma et al. (2009) that lexical disambiguation tasks were amenable to approaches using Web 1T have been influential, finding in particular who found that on the tasks of preposition selection and context-sensitive spelling correction web n-gram models could achieve an improvement of up to 24% on the state of the art. It is therefore difficult to conclusively decide that Web 1T is a bad candidate for our model.

3.4.2 Performance of Edmonds-Collocate and Web1T-PMI on attitudinal words

In comparing the performance of either EDMONDS-COLLOCATE or WEB1T-PMI on attitudinal relative to non-attitudinal synsets; it is necessary to take into account the wildly different baselines (see Table 3.5). We therefore, when comparing either method to the baseline, compare only the times that the method in question made a different prediction from the baseline. We then calculate the proportion of times that the method is correct versus the baseline (see Tables 3.7[10] and 3.8).

When we compare the accuracy of each method on attitudinal synsets versus non-attitudinal synsets by calculating the z-statistic as is standard for comparing two proportions (Moore and McCabe, 2003), we find that the difference in performance is significant at the 1% level for each of the data sets with the exception of one case. (In WEB1T-PMI case where $q = 2$ and the test set is top3, the performance difference is not significant.) That is, EDMONDS-COLLOCATE improves significantly over the baseline for attitudinal synsets in three of the four test runs, whereas WEB1T-PMI never does.

3.4.3 Entropy analysis

In this section, we describe an analysis of the results in Section 3.3.2 in terms of whether the balance of frequencies among words in the synset contribute to the quality of our prediction result.

We note that the higher baseline for non-attitudinal near synonyms might be a reflection of characteristics of the specific synsets used that make the WEB1T-PMI method perform worse, rather than its 'non-attitudinality' as such. Specifically, it might be that these synsets are on average more highly skewed. That is, the most frequent word might be far more dominant and therefore more easily chosen. Consider for example, a hypothetical training corpus where there are two sets of near synonyms: *bad*, *awful* and *terrible*; and *good*, *great* and *terrific*. If *bad* was seen 98 times and *awful* and *terrible* only once, then it is comparatively easy for a system to perform very well by always predicting *bad*. If *good*, *great* and *terrific* are seen 40 times, 35 times and 25 times each, it is comparatively difficult to predict the correct one by selecting one or two very dominant terms. We wish to find out if attitudinal and non-attitudinal near-synonym sets differ from each other in this way.

[10]The numbers given in Table 3.7 for EDMONDS-COLLOCATE are different from those given in Gardiner and Dras (2007a) due to the same error that required Table 3.6 be updated.

		All words				Att. words				Non-att. words			
Synsets	q	Baseline		EDM		Baseline		EDM		Baseline		EDM	
top3	2	13828	51·0%	13282	49·0%	938	41·2%	1340	58·8%	12890	51·9%	11942	48·1%
top4	2	14565	51·2%	13899	48·8%	1552	44·8%	1909	55·2%	13013	52·0%	11990	48·0%
top3	5	39019	61·0%	24924	39·0%	2233	48·2%	2398	51·8%	36786	62·0%	22526	38·0%
top4	5	41868	61·4%	26360	38·6%	4468	55·8%	3544	44·2%	37400	62·1%	22816	37·9%

Table 3.7: Number of times each method is right when the baseline and EDMONDS-COLLOCATE predict a different word

		All words				Att. words				Non-att. words			
Synsets	q	Baseline		Web 1T		Baseline		Web 1T		Baseline		Web 1T	
top3	2	85859	49·8%	86695	50·2%	12794	50·0%	12793	50·0%	73065	49·7%	73902	50·3%
top4	2	92564	51·9%	85895	48·1%	14787	53·6%	12790	46·4%	77777	51·5%	73105	48·5%
top3	4	80210	50·1%	79875	49·9%	13152	52·4%	11966	47·6%	67058	49·7%	67909	50·3%
top4	4	86316	52·2%	78902	47·8%	14892	55·0%	12197	45·0%	71424	51·7%	66705	48·3%

Table 3.8: Number of times each method is right when the baseline and Web1T-PMI predict a different word

Test set	q	Category	Entropy
top3	2	$-0{\cdot}11$	$0{\cdot}41^{*}$
top3	4	$-0{\cdot}10$	$0{\cdot}36^{*}$
top4	2	$-0{\cdot}17^{*}$	$0{\cdot}38^{*}$
top4	4	$-0{\cdot}15^{*}$	$0{\cdot}34^{*}$

* Significant at the $p < 0.05$ level

Table 3.9: Regression co-efficients for WEB1T-PMI between independent variables synset category and synset entropy, and dependent variable prediction improvement over baseline

In order to measure a correlation between the balance of frequencies of words and the prediction result, we need a measure of 'balance'. In this case we have chosen information entropy (Shannon, 1948), the measure of bits of information required to convey a particular result. In general, the entropy $H(X)$ of a series $X : x_1, \ldots, x_n$ is given by equation (3.3) where $p(x_i)$ is the probability mass of x_i:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_2(p(x_i)) \tag{3.3}$$

In general $p(x_i)$ is estimated using ratio of the number of times x_i is observed, $c(x_i)$ over the number of observances in total:

$$p(x_i) = \frac{c(x_i)}{\sum_{j=1}^{n} c(x_j)} \tag{3.4}$$

Very even distributions require a lot of information to convey and thus have high entropy values, whereas very skewed distributions have low entropy. At the extreme, consider a coin that always comes up heads, where all one needs to communicate the outcome is the number of tosses—thus, entropy—as compared to a fair coin where one must use one bit to communicate the result of each toss. (Likewise the entropy of our very skewed *bad*, *awful* and *terrible* example above is 0.61, comparatively low compared with the entropy of the *good*, *great* and *terrific* example, which is 1.56.)

The entropy of a synset's frequencies here is measured using the proportion of total uses of the synset that each particular word represents. A synset in which frequencies are reasonably evenly distributed has high information entropy and a synset in which one or more words are very frequent as a proportion of use of that synset as a whole have low entropy.

We then carried out multiple regression analysis using the category of the synset (attitudinal or not attitudinal, coded as 1 and 0 for this analysis) and the entropy of the synset's members' frequencies as our two independent variables; this allows us to sepa-

Words in set	No. tests	Performance	Improvement
Non-attitudinal			
capable, *open*, *subject*	4449	70·0%	+26·8
rest, *remainder*, *balance*	4305	63·7%	+11·7
measure, *step*	3864	71·0%	+9·9
main, *independent*	4582	53·0%	+7·3
report, *composition*, *paper*	9098	69·1%	+5·4
Attitudinal			
hard, *heavy*, *arduous*	4224	56·6%	+9·6
bad, *risky*, *speculative*	2839	73·8%	+1·1
big, *large*, *great*	16641	49·3%	+0·9

Table 3.10: The best performing 5 non-attitudinal and 3 attitudinal sets using EDMONDS-COLLOCATE, compared to the baseline

rate out the two effects of synset skewness and attitudinality. Regression co-efficients for WEB1T-PMI are shown in Table 3.9. No statistically significant correlations were found for EDMONDS-COLLOCATE.

Table 3.9 shows that in general, for WEB1T-PMI, performance is negatively correlated with category but positively with entropy. The negative correlation with category implies that this statistical method works better for predicting the use of non-attitudinal near synonyms, even factoring out entropy. WEB1T-PMI may be disadvantaged in the comparison with EDMONDS-COLLOCATE partly due to its greatly increased tendency to make a prediction at all. While Inkpen found that a larger training corpus benefited INKPEN-PMI on her test set, it is possible that WEB1T-PMI is also disadvantaged by training on the Web1T corpus, which, while much larger, does not resemble the Wall Street Journal closely.

In Table 3.10 we see examples of the best performing synsets using EDMONDS-COLLOCATE(for the TOP3 setting with a window size of 4, since it had the best results), as measured by improvement over the baseline, and divided into attitudinal and not. Only three examples are shown for attitudinal as only three actually achieved noticeable positive difference from the baseline.

While the performance of the best two sets, *capable*, *open*, *subject*; and *rest*, *remainder*, *balance*, are clearly exceptional in this table, it demonstrates that either EDMONDS-COLLOCATE cannot achieve equivalent performance on attitudinal near synonyms or that our test data for this chapter is insufficient to demonstrate that it can[11].

Similarly, in Table 3.11 we see examples of the best performing synsets using WEB1T-PMI. We here see a greater range of potential improvement over the baseline of up to

[11]Recall that there are only 7 attitudinal synsets in this test data.

Words in set	No. tests	Performance	Improvement
Non-attitudinal			
individual, *separate*, *single*	4813	74·2%	+40·6
capable, *open*, *subject*	4453	80·5%	+37·3
independent, *main*	4595	79·7%	+34·1
run, *test*, *trial*	3968	68·8%	+33·0
deal, *hand*	3742	84·5%	+32·7
Attitudinal			
battle, *conflict*, *fight*	3156	58·0%	+11·0
force, *violence*	3210	91·1%	+8·0
big, *great*, *large*	16669	53·8%	+5·4
arduous, *hard*, *heavy*	4236	50·1%	+3·2
low, *modest*, *small*	9680	55·4%	−3·0

Table 3.11: The best performing 5 non-attitudinal and attitudinal sets using Web1T-PMI, compared to the baseline

Words in set	No. tests	Performance	Improvement
Edmonds-Collocate			
single, *separate*	10361	73·0%	−7·3
department, *section*	4221	79·8%	−6·0
Web1T-PMI			
conglutination, *union*	4261	28·8%	−71·2
alteration, *change*, *modification*	3614	44·4%	−53·5

Table 3.12: The 2 worst performing sets using Edmonds-Collocate and Web1T-PMI, compared to the baseline

41 percentage points, and five attitudinal sets can be shown because there are enough predictions to have interesting results even at the fifth item. It is, as with Edmonds-Collocate, unclear whether the smaller improvements for attitudinal sets relate to more difficulties predicting the correct choice, or absence of appropriate test data.

Conversely to the greater potential of Web1T-PMI for improvement over the baseline, Table 3.12 shows that the scope for failure by Web1T-PMI is much greater, especially in cases with a high baseline (although cases with a high baseline have further to fall, as it were, when percentage improvement is used as here). Again, our test set contains insufficient data to show whether or not this performance differs between attitudinal and non-attitudinal near synonyms.

3.5 Conclusion

In this chapter we have compared two different methods, Edmonds-Collocate and Web1T-PMI, for predicting near-synonym usage as regards their capacity to predict the

use of a particular word that differs from its near synonyms in attitude, rather than in some other way. We have shown that while both perform promisingly on the small test set used in Edmonds (1997), neither performs well on our larger test set. Both methods vary statistically significantly in their ability to predict the usage of attitude and non-attitudinal near synonyms, supporting our hypothesis that a corpus statistics approach is sensitive to the attitudinality of the near synonyms, although not exactly in the way that we expected! It is apparent that not just any corpus statistics method will do better for attitudinal near synonyms; as our EDMONDS-COLLOCATE does, but our WEB1T-PMI does not. Skewness of the synsets also plays a part.

This, together with the already different performance on the baseline method, suggests that new approaches to the FILL IN THE BLANKS (FITB) task may be called for, with emphasis on evaluating them in light of their performance on near synonyms with attitude differences.

It is also apparent at this point that a further test set of near synonyms needs to be developed, both one which contains more attitudinal near synonyms so that we can examine their performance in more detail, and one that does not tend towards minor senses or very polysemous words. Several problems with our present set of synsets remain:

1. while using the most frequent set of words gives us a lot of test sets, there is no reason to suppose that highly frequent synsets are particularly likely to contain or differ in sentiment, and indeed we found that they did not much differ in sentiment; and

2. there is some reason to suppose that some highly frequent synsets are highly frequent because some of the words in them are highly polysemous, and thus likely to be difficult to correctly predict in a way that is less interesting to us.

One possiblity to proceed from here might be to refine our existing test set more, for example to manually remove some of the words that resulted in annotator disagreement in order to be able to include the synsets with the disagreed words removed. However, in the next chapter we instead re-think our use of WordNet and return to a data source used in previous near-synonym work, a usage guide for second language writers of English, which is likely to present near-synonyms with particular semantic or collocational differences.

Chapter 4

Improving near-synonym choice with sentiment differences

In Chapter 3 we concluded with a suggestion that a new approach to choosing among near synonyms with attitudinal differences was needed. In this chapter we develop novel approaches to the Fill In the Blanks (FITB) task given in Section 2.4.1, and explore their performance, particularly on near-synonym sets with attitudinal differences.

In this chapter, we draw on work in sentiment analysis as reviewed in Section 2.5. First, we investigate a supervised approach; in addition to the recent popularity of supervised approaches to the FITB task including Wang and Hirst (2010); Yu et al. (2010) as discussed in Section 2.4.2, this is suggested by work in sentiment analysis that shows that supervised methods also work well for that task. However, we particularly evaluate our approach with respect to a wide range of near-synonym sets, including many that have affective meaning.

As in the first exploration of supervised methods in document sentiment classification, by Pang et al. (2002), we start with a simple unigram model. Second, we then look at broader aspects of the document to use as features. We hypothesise that affective differences between near synonyms, such as the difference in attitude between *slim* and *skinny*, may be more influenced by more general aspects of the document such as affect, than are near synonyms that differ in other aspects. We test this hypothesis by applying proven techniques from the domain of sentiment analysis to the lexical gap problem.

In Section 4.1 we describe our data set to be used for this task; and in Section 4.2 we discuss the selection of appropriate baselines. In Section 4.3 we describe our experimental setup, followed by the definition of our unigram models. We discuss the results of these,

which motivates the selection of some additional features. Section 4.4 then describes further experiments based on the document-level features arising from Section 4.3, while Section 4.5 describes those based on a notion of weighting in the feature space.

4.1 Affective Text: Near-Synonyms and Corpora

Our hypothesis in this chapter is that affective and non-affective synonyms behave differently in the context of the FITB task, and that this may be extended to new models as developed here. To test this hypothesis again require both a set of near synonyms, and we require a set of documents with known document sentiment. Our choice of these two datasets is described in this section.

4.1.1 Documents containing sentiment

As samples of our test words in context, we used the SCALE dataset v1.0 movie review data set (SCALE 1.0) (Pang and Lee, 2005). As discussed in Section 2.5.4.1, SCALE 1.0 consists of 5000 movie reviews authored by four reviewers on Internet sites, ranging from extremely negative to extremely positive reviews. The range of authors is narrow, this is one of the central corpora of sentiment analysis, and in addition, the narrow range of authors allows us to explore author identity as a feature later in the chapter.

4.1.2 Sentiment annotated near synonyms

Since we seek to examine the choice between near synonyms that differ in sentiment we require a source of such near synonyms.

In Chapter 3 we developed an annotated set of WordNet (Fellbaum, 1998) synsets (shown in full in Appendix A). However not only is WordNet not annotated for polarity, it encodes very fine-grained sense distinctions which usually precludes having near synonyms that differ in sentiment contained in a single synset. In addition, there were very few synsets—only 7 of the 58 used—that had any affective meaning, and as we are especially interested in these sets, we require an additional source of near synonyms.

There are some versions of WordNet annotated for sentiment, for example SentiWordNet (Esuli and Sebastiani, 2006), but examination of this data shows that it is not easy to use it to produce clear distinctions such as "this set of near synonyms differ in sentiment and these do not".

As an example of why SentiWordNet is a difficult source of data for this use-case,

POS	Words	Positive score	Negative score
Adjective	*rich*, *plentiful*, *plenteous*, *copious*, *ample*	0·125	0
Verb	*merit*, *deserve*	0·75	0·125
Verb	*swell*, *puff up*	0·125	0·625
Noun	*feat*, *exploit*, *effort*	0·375	0·125
Noun	*swan song*, *last hurrah*	0·125	0·125

Table 4.1: Example entries from SentiWordNet

consider the five entries from it shown in Table 4.1. While having real-numbered values for the positivity and negativity of synsets would be of use if we were seeking features for a sentiment classification learner, one of the uses to which SentiWordNet has been put, but it is not straightforwardly apparent how to identify "'synsets that have the same affect", "synsets that differ in affect" and "synsets that have no affective meaning" from the numerical values.

The near-synonym usage guide *Choose the Right Word* (Hayakawa, 1994) contains near synonym sets chosen by a human author specifically as a guide to the subtleties of near-synonym word choice for readers and writers of English. It is therefore a good source of near synonyms that differ in fundamental ways such as sentiment. It was used by Inkpen and Hirst (2006) as a source of near synonyms marked for differences such as denotational differences (in which near synonyms could differ in what they *suggest* or *imply*, for example) and attitude and style differences, including near synonyms that are more *pejorative*, *disapproving* or *favourable*.

Inkpen and Hirst (2006) derived data automatically from *Choose the Right Word* by a decision list algorithm, and this data is not available for reasons of copyright. In addition, the focus of this data was on all axes in which near synonyms can differ, rather than on the axis of positive or negative attitude to the subject of a description in particular. Thus we have annotated our own data.

We use sets of near synonyms drawn from an earlier edition of this work, *Use the Right Word* (Hayakawa, 1968). (An excerpt from an entry, not used in this thesis, in *Use the Right Word* is shown in Figure 4.1 as an example of its contents.) We have sampled the sets by annotating those that are listed under the letter A, of which there are 57 in total. Of these 57 total sets, we exclude 16 sets that do not include at least two words that are each used at least five times in our sentiment annotated documents described in Section 4.1.1. This leaves 37 sets, totalling 133 words.

In addition because this set only contained 5 sets with the same affect shared among

malign, *asperse*, *defame*, *libel*, *slander*, *vilify*
These words mean to say or write something, often misleading or false, that is damaging to a person or a group of people. *Malign* is perhaps the broadest word in the group in that the feelings which motivate a person who *maligns* another can range from... simple ill will... to bitter hatred...
Asperse and *vilify* imply false accusations made in order to ruin someone's reputation... *Apserse* however is extremely formal, and more commonly appears in the form of a plural noun...
Defame can specifically indicate an attempt to destroy someone's good name...
Libel and *slander*, in their most restricted sense, are legal terms pertaining to defamation... In popular usage, however, both words are applied to false accusations by any means. See *accuse*, *belittle*, *lie*.
ANTONYMS: *praise*

Figure 4.1: Excerpt from an entry in *Use the Right Word*

Set sentiment type	Word	Sentiment	Word	Sentiment	Word	Sentiment
Same	*ludicrous*	Negative	*senseless*	Negative	*foolish*	Negative
	preposterous	Negative	*ridiculous*	Negative	*farcical*	Negative
	absurd	Negative	*silly*	Negative	*irrational*	Negative
	unreasonable	Negative				
None	*attend*	Neutral	*accompany*	Neutral		
Differing	*precise*	Neutral	*accurate*	Neutral	*exact*	Neutral
	right	Positive	*nice*	Neutral	*correct*	Neutral
	true	Neutral				

Table 4.2: Examples of test sets annotated for overall set sentiment differences, and for the sentiment of individual words

all words, we developed an extra 10 sets with the same affect shared among all the words[1].

We thus have a total of 47 test sets. These near synonyms were annotated for affect by the supervisor of this author. The annotation scheme called on the annotator to rely on *Use the Right Word*'s interpretation of the sets, rather than personal linguistic intuition. For example, *Use the Right Word* suggests in the entry for *aloof*: "Both *reserved* and *detached* can be associated with attractive qualities, whereas *aloof* is seldom so considered". Annotations were of two kinds:

1. for a given set, whether *Use the Right Word* indicates that that set contains at least some words conveying sentiment ('affective', or 'not affective'); or

2. for every word within any set marked as 'affective', whether *Use the Right Word* indicates that that word has positive, negative, or neutral affect.

An example of three annotated test sets is shown in Table 4.2. Observe that the set

[1]This addition was made at the suggestion of reviewers of an unpublished version of this chapter.

	Development set		Test set	
Set type	No. sets	Mean no. words	No. sets	Mean no. words
All sets	12	3.6	35	3.6
no-affect words	8	3.1	12	2.8
same-affect words	2	4.5	13	3.5
differing-affect words	2	4.5	10	3.9

Table 4.3: Distribution of sentiment among *Use the Right Word* sets

containing *attend* and *accompany* has set type marked as NONE, meaning no affect. The marking of NEUTRAL against the individual words is thus implied. The set containing *ludicrous*, *senseless* and others is marked SAME, and all the words are indeed marked identically with NEGATIVE sentiment. The set containing *precise*, *accurate* and others is marked DIFFERING and the sentiment of the individual words does indeed differ, with words within it having varying POSITIVE and NEUTRAL seniment. The complete set of annotated test sets is given in Appendix B.

Based on the annotations, we divided the 47 sets into *no-affect* sets (e.g. *absorb*, *digest*, *incorporate*), *same-affect* sets (e.g. *absurd*, *ludicrous*, *silly*, ...) where the affect was the same for every word in the set, and different-affect sets (e.g. *aloof*, *detached*, *reserved*) where the set was marked affective but affect of individual words differed within the set. The usage of sentiment-marked words in this data is shown in Table 4.3.

It is important to note that some of the near synonyms described here as "differing" in affect differ in that they contain neutral and negative, or neutral and positive words, for example we so include *insight* and *perception* based on the former being positive and the latter neutral. Therefore the *same-affect* sets contain *only* actively positive or negative words, not neutral ones.

4.1.2.1 Development set

Feature comparison was initially performed with a development set, also as indicated in Table 4.3), of 12 sets among the 47 sets of words. These consist of 8 sets without affect, 2 sets in which all members shared the same affect, and 2 sets where some members differed in their affect. In the SCALE 1.0 corpus, the words from these sets appeared 8587, 310 and 3759 times respectively. The words from the equivalent test sets appeared 7303, 4441 and 4374 times respectively.

4.1.2.2 True synonymy

A key question we discuss in this chapter is whether FITB choice methods perform differently for different types of near-synonym sets, specifically whether they behave differently on the distinction between *no-affect*, *same-affect* and *differing-affect* kinds described above.

One possible factor that could distort our results is the presence of *true* synonyms in our data set. It is a common view, which we have followed in the thesis, that there is no such thing as perfectly true synonymy. For example, Cruse (1986, p. 270) writes "if [true] synonyms exist at all, they are extremely uncommon… one would expect either that one of the items would fall into obsolescence, or that a difference in semantic function would develop," and Clark (1992) presents some evidence from spoken language that speakers are careful to contrast any two near synonyms in their regular vocabulary.

For the practical purposes of this chapter, however, it might be the case that words in different categories are more closely, or very frequently, interchangeable: if, for example, the *same-affect* or *no-affect* sets contain extremely closely related near synonyms or even true synonyms relative to the other, we would expect them to behave differently. In particular, we would expect that true synonyms would be chosen by writers essentially randomly,[2] and so the baseline performance would be low and choice methods probably not improve greatly on it. (*differing-affect* sets by definition contain words that differ in affect, so we already have evidence that they consist of at best near synonyms rather than true synonyms.) We therefore examine this in a quantitative manner.

In order to attempt to quantify the relatedness of each near-synonym set in our data, we return to WordNet as a measure of word relatedness. As we have seen in Chapter 3, words in WordNet synsets are closely related to each other (although not necessarily or even usually perfectly synonymous). If we found, therefore, that words in each of our *same-affect* sets tended to be in the same synsets and words in our *no-affect* sets tended to be in different synsets, we would suspect that words in the *same-affect* sets are much more closely related.

We therefore, for each of our word sets, count how many WordNet 2.0 synsets contain *at least 2 words* from that word set. For example, consider our (differing-affect) set *feat*, *operation*, *act*, *exploit*, *action*, *performance*. There are two WordNet 2.0 synsets that each contain two or more words from this set:

operation*, (*functioning*), *performance "process or manner of functioning or operating"

[2] Or nearly so, if varying collocational restrictions are allowed for true synonyms.

Proportion of our sets having 2 or more words in	*no-affect*	*same-affect*	*differing-affect*
… at least 1 common synset	55%	75%	58%
… at least 2 common synsets	20%	26%	25%
… 3 common synsets	5%	13%	0%

Table 4.4: Similarity of words in each of *no-affect*, *same-affect*, *differing-affect* by shared WordNet synsets

feat*, (*effort*), *exploit "a notable achievement"

In Table 4.4 we see that words in *same-affect* are probably more closely related overall than words in *no-affect* and *differing-affect*, since for 75% of the sets in *same-affect*, two or more of the words appear in the same WordNet synset at least once.[3] However, while this may lead us to suspect a lower baseline for *same-affect* may occur, it does not strongly suggest large numbers of true synonyms in our data set. It may be reasonable to suspect a somewhat lower baseline performance.

A complete listing of words in our sets that share synsets can be found in Appendix B.

4.2 Comparison of baselines

There are a number of possible baselines, each with various merits. As candidates in this chapter, we consider four possible approaches.

Two of these are standard possible baselines:

***most frequent* category** which we used as a baseline in Chapter 3 and which has been used by other FITB researchers as discussed in Section 2.4.2.1; and

a language model encoding the likelihood of short strings of text, used for the FITB task by Islam and Inkpen (2010) as discussed in Section 2.4.2.6.

Two of these are based on our implementation of methods described in the literature, as discussed in Section 3.1:

Edmonds-Collocate (Edmonds, 1997, 1999) and as described in Section 2.4.2.2; and

Web1T-PMI (based on Inkpen (2007b)) as described in Section 3.1.2.

We outline these below, and discuss their relative performance, along with the resulting choice of baseline for the rest of the chapter.

[3]Note that, as in the example of *operation* and *performance* sharing a synset and *feat* and *exploit* sharing a second synset, the entries for the *2 common synsets* and *3 common synsets* rows in Table 4.4 may not actually concern the same words from each of our sets.

Type of set	No. tests			MF %			LM %		
	dev	test	both	dev	test	both	dev	test	both
No affect	8587	7303	15890	82·6	85·2	83·8	60·7	63·4	61·9
Same affect	310	4441	4751	44·8	71·4	69·6	45·8	30·8	33·8
Differing affect	3759	4374	8133	50·9	57·1	54·2	40·1	57·6	49·5
Total	12656	16118	28774	72·3	73·8	73·1	54·2	58·4	56·3

Table 4.5: Accuracy of *most frequent* (MF) and language model (LM) baselines for development and test sets on the SCALE 1.0 dataset

4.2.1 Candidate baselines

4.2.1.1 Most frequent

As shown in Section 2.4.3, all researchers approaching the FITB task compare with a *most frequent* baseline , that is, comparing with the method of always selecting the most frequent word in a set to fill the gap. This baseline can be quite high: Inkpen reports it as achieving 44.8% accuracy on the seven sets of test words used by herself and Edmonds (these seven sets contain between two and four words, with a mean of exactly three words per set). However, we find that on our dataset, even the *most frequent* baseline is considerably higher for most sets in our development set, as shown in Table 4.5.

4.2.1.2 Language model

As Inkpen (2007b) notes, and as seen in the 2007 lexical substitution task, statistical language models are the mainstream method of lexical choice. Inkpen and Hirst (2006) compared their system to a language model baseline that was implemented as part of the HALogen NLG system (Langkilde and Knight, 1998), trained on 250 million words of text from the news genre. HALogen's word choices when combined with the anti-collocation method presented by Inkpen and Hirst (2006) outperformed HALogen alone, and the method presented by Inkpen (2007b) outperforms anti-collocations, and thus Inkpen (2007b) concludes that language models would be outperformed by her newer method.

However, as discussed in Section 2.4.2.6, Islam and Inkpen (2010) implement a language model and find that it outperforms INKPEN-PMI on 5 of the 7 evaluation sets (see Table 2.7).

Here, like Islam and Inkpen, we implement a baseline language model choice system using the Web 1T data (Brants and Franz, 2006), which is described in more detail in Section 3.1.2. In short, Web 1T contains n-gram frequency counts, up to and including

5-grams, as they occur in a trillion words of World Wide Web text.

We make a word choice by estimating the most probable 3-gram, backing off to shorter n-grams where necessary. We fundamentally use the backoff method of Katz (1987).

The backoff method is required for smoothing the data: a common problem in statistical models of natural language. The problem is that linguistic data is sparse, meaning that many perfectly valid sequences of words will never be seen in training data. Estimating the probability of these sequences based on their count (ie, estimating it at 0) is not only falsely underestimating their probability, but ends up setting the probability of related strings to 0 as well (eg if *cat* is never seen in the corpus, and has an estimated probability of 0, the probability of *the cat in the hat* is also 0). Smoothing techniques reserve some of the probability space for unseen items, avoiding this problem. (Jurafsky and Martin, 2009)

The Katz language modelling technique uses smoothing below some threshold, with this threshold often empirically determined, and uses the more accurate raw high-frequency accounts above it. Given that Web 1T only provides counts frequencies of at least 40 for 2- to 5-grams and 200 for unigrams, we use these as our threshold.

To explain the method more fully, consider predicting a word choice as estimating the most probable 3-gram. Consider example (4.1), originally introduced by Edmonds (1997) and here in Chapter 1:

(4.1) However, such a move would also run the risk of cutting deeply into U.S. economic growth, which is why some economists think it would be a big _____.

In order to predict a near synonym to fill the gap in example (4.1), we would be estimating which of the following 3-grams are most likely:

(4.2) a big *error*

(4.3) a big *mistake*

(4.4) a big *oversight*

We would then choose from among *error*, *mistake* and *oversight* by choosing the word contained in the most probable 3-gram.

As is common in language models, we back off to 2-grams and 1-grams where necessary. We draw our discussion that follows from the original paper of Katz (1987), the more detailed explication of Gale and Sampson (1995), and the overview of Jurafsky and Martin (2009). In general, backoff models look like this:

- The count of ngram $w_1, \ldots, w_n$ in the training data is denoted by

$$C(w_1, \ldots, w_n) \tag{4.1}$$

- The smoothed count C^* of ngram $w_1, \ldots, w_n$ is given by the specific smoothing algorithm.

- The adjusted probability P^* of ngram $w_1, \ldots, w_n$ is given by

$$P^*(w_n|w_1, \ldots, w_{n-1}) = \frac{C^*(w_1, \ldots, w_n)}{C(w_1, \ldots, w_{n-1})} \tag{4.2}$$

- If a count for $w_1, \ldots, w_n$ is unavailable, $P(w_n|w_1, \ldots, w_{n-1})$ is estimated using the counts for ngram $w_2, \ldots, w_n$ using a back-off model, where α is a the proportion of the probability space reserved for unseen events:

$$P(w_n|w_1, \ldots, w_{n-1}) = \alpha(w_1, \ldots, w_{n-1})P(w_n|w_2, \ldots, w_{n-1}) \tag{4.3}$$

In the smoothing and backoff implementation of Katz (1987) the specifics of these functions are:

- The number of ngrams with count c in the training data is denoted by

$$r(c) \tag{4.4}$$

- The adjusted count C^* of an ngram $w_1, \ldots, w_n$ is smoothed by the Good-Turing estimation Good (1953) and is given by

$$C^*(w_1, \ldots, w_n) = \frac{(C(w_1, \ldots, w_n) + 1) \cdot r(C(w_1, \ldots, w_n) + 1)}{r(C(w_1, \ldots, w_n))} \tag{4.5}$$

- The proportion α of the total probability mass allocated to unseen words w_n following $w_1, \ldots, w_{n-1}$ is given by:

$$\alpha(w_1, \ldots, w_{n-1}) = \frac{\beta(w_1, \ldots, w_{n-1})}{\sum_{w_n : C(w_2, \ldots, w_n) > 0} P^*(w_n|w_2, \ldots, w_{n-1})} \tag{4.6}$$

  where the function β is given by:

$$\beta(w_1, \ldots, w_{n-1}) = 1 - \sum_{w_n : C(w_1, \ldots, w_n) > 0} P^*(w_n|w_1, \ldots, w_{n-1}) \tag{4.7}$$

In a typical language model implementation, at some sufficiently large value of r the probability of n-gram $w_1, \ldots, w_n$, $P(w_n|w_1, \ldots, w_{n-1})$ would be estimated using a maximum likelihood estimate instead, given in equation (4.8), as in for example Gale and Sampson (1995):

$$P(w_n|w_1, \ldots, w_{n-1}) = \frac{C(w_1, \ldots, w_n)}{C(w_1, \ldots, w_{n-1})} \tag{4.8}$$

Given that we are using Web 1T for our values of $C(w_1, \ldots, w_n)$, and Web 1T does not provide counts for $r < 40$ for 2- to 5-grams and for words when $r < 200$, we use equation (4.8) rather than equation (4.2). There are several reasons for this:

1. the standard methods for determining the value of this cut-off would usually be applied on the tail end of the data, precisely what's been removed from Web 1T; and

2. even for fairly low values of r available in Web1T, ie $r \approx 40$, $C^*(w_1, \ldots, w_n) > C(w_1, \ldots, w_n)$ when using equation (4.5) to compute C^*. This results in periodic cases where $\beta < 0$ in equation (4.7).

Given this, we also estimate α differently from equation (4.6). The Web 1T corpus does not include n-grams for $n \geq 2$ with counts of less than 40 (or unigrams with counts of less than 200). We therefore estimate the probability mass allocated to unseen n-grams by the proportion of the count of an n-gram $w_1, \ldots, w_n$ unaccounted for by known $n+1$-grams $w_1, \ldots, w_{n+1}$. To give an example of how this is done, let us assume that bigram *mistakes are* has a Web 1T count of 300, and the only trigrams beginning with *mistakes are*, *mistakes are bad* and *mistakes are good*, with counts of 75 each. We then have 150 unseen tokens following *mistakes are* and thus the value of α is 0.5.

This can be formally expressed as

$$\alpha(w_1, \ldots, w_{n-1}) = \frac{\sum_{w_n : C(w_1, \ldots, w_n) > 0} C(w_1, \ldots, w_n)}{C(w_1, \ldots, w_{n-1})} \tag{4.9}$$

4.2.1.3 Edmonds-Collocate and Web1T-PMI prediction method

We also consider the baselines EDMONDS-COLLOCATE and WEB1T-PMI as described in Sections 2.4.2.2 and 3.1.2 respectively.

4.2.2 Results and discussion

We discuss each individual potential baseline with respect to the most frequent baseline below. In summary, though, all performed worse, so we use the *most frequent* baseline in the rest of the chapter.

4.2.2.1 Language model results

Language model results are given in Table 4.5. The language model badly underperformed compared to the *most frequent* baseline, which was not expected from the results of Inkpen and Hirst (2006).

It is, however, difficult to compare this result directly with the one given in Inkpen and Hirst (2006). In that article, the HALogen system's language model, which predicts the correct near synonym between 58 and 83% of the time, is given as a baseline against which to compare the Xenon experimental system, but values for the *most frequent* baseline are not given for the same test sets.

One possible reason for the bad performance of the language model is that the *most frequent* baseline here, as noted earlier, is substantially higher than in other work, perhaps because of the nature of the near-synonym sets. A second possible reason that could interact with the first is suggested by Inkpen and Hirst (2006), who note that the collocations encoded in the language model will tend to be with function words, per the short n-gram distance.

However in light of the good performance of language models on the FITB task described in Islam and Inkpen (2010), also trained on Web 1T, this requires investigation in the future. There are some differences in our language model implementation to that of Islam and Inkpen, including a different smoothing technique and the use of 3-grams rather than 5-grams. However, without a high performing language model, we do not consider it as a baseline in this chapter.

4.2.2.2 Edmonds-Collocate results

Table 4.6 shows the results of the EDMONDS-COLLOCATE prediction method. This method performs poorly, unable to make a prediction in the vast majority of cases, and while it can be successful when we limit the cases to those where it does make a prediction at all, its success is not uniform, but is concentrated in the same affect group, and, on inspection, located almost entirely within the *bad*, *distasteful*, *objectionable*, *unpleasant* set of near synonyms, which dominates that group numerically. Otherwise EDMONDS-COLLOCATE

Query window size	4		10	
	Predictions	Correct	Predictions	Correct
Overall	5·2	65·5	11·9	67·1
No affect	4·8	51·0	11·5	62·5
Same affect	6·6	96·5	12·7	95·0
Differing affect	5·0	69·1	12·2	58·5

Table 4.6: Percentage of times EDMONDS-COLLOCATE made a prediction and percent of those predictions that were correct

Parameters		Overall		No affect		Same affect		Differing affect	
q	k	Value	+/-	Value	+/-	Value	+/-	Value	+/-
2	2	54·5	−18·6	61·6	−22·3	41·0	−28·6	48·6	−5·6
4	2	51·1	−22·1	57·3	−26·5	39·0	−33·2	45·9	−8·3
2	3	48·2	−24·9	56·5	−27·3	31·7	−40·6	41·8	−12·5
4	3	45·7	−27·4	50·9	−32·9	33·8	−38·8	42·5	−11·7
2	4	42·5	−30·6	51·1	−32·7	25·8	−46·5	35·6	−18·6
4	4	43·1	−30·0	49·6	−34·3	35·6	−31·3	38·6	−15·6

Table 4.7: WEB1T-PMI results compared to *most frequent* baseline

performs well below other methods even when its success is measured on cases where it can make a prediction. This is unsurprising given the finding that EDMONDS-COLLOCATE simply does not use a large enough training set to make useful predictions (Inkpen, 2007b).

4.2.2.3 Web1T-PMI results

WEB1T-PMI results for varying parameters are shown in Table 4.7. One noteworthy characteristic is that a wider window (parameter k) of context around a gap almost always diminishes performance, with the exception of the performance on same affect for $q = 4$ and $k = 4$.

In general these results also show a large decrease in performance over the baseline, contrary to results reported in Chapter 3, where results were approximately equal to the baseline. Several reasons may hold as to why WEB1T-PMI performs unexpectedly poorly:

- the SCALE 1.0 test data in this chapter is very different from that used in Chapter 3, which tested on the Wall Street Journal;
- the test near synonyms in this chapter are also different from those used in previous work, being near synonyms selected by a human editor rather than high frequency WordNet synsets; and
- the test near synonyms in Chapter 3 were trimmed to only three or four possible

choices, whereas this set from Appendix B contains up to 9 possible alternatives (in the case of the *same-affect* set containing *ludicrous*, *senseless*, *foolish*, *preposterous*, *ridiculous*, *farcical*, *absurd*, *silly* and *irrational*).

4.2.3 Conclusion

In this section we have tested EDMONDS-COLLOCATE approximating the method of Edmonds (1997, 1999); and WEB1T-PMI approximating the method of Inkpen (2007b), albeit on a different data set, on a new test set divided into NO AFFECT, SAME AFFECT and DIFFERING AFFECT classes. The EDMONDS-COLLOCATE method performs especially well on SAME AFFECT test sets, allowing us to suggest that these may be especially amenable to statistical techniques, but the overall poor performance of EDMONDS-COLLOCATE and WEB1T-PMI on our new test set requires us to develop better approaches to FITB and explore our hypothesis further.

4.3 Unigram models

In this section we examine the effectiveness of simply using unigrams as Pang et al. (2002) did for document sentiment classification, relative to our chosen baseline from Section 4.2.

The learner We use Support Vector Machines (SVMs) as a binary classifier. SVMs implement a kernel-based supervised learning method: they perform classification by modelling the training data in a high dimensionality feature space and perform efficient searches for hyperplanes dividing this space into categories (Cristianini and Shawe-Taylor, 2000).

We use the SVM-Light implementation of Support Vector Machines (Joachims, 1999), which implements a binary classifier. Therefore a separate machine m_{c_i} is trained for each binary decision: is the gap filled by candidate word c_i or not? We select the SVM m_c from among the set that returns the highest confidence score (as suggested by Rifkin and Klautau (2004); Liu and Zheng (2005)) and choose c to fill the gap. c is judged correct if it matches the word w the original author used.

Hypothesis testing A single method for comparing classification accuracy has not been universally accepted. What constitutes an invalid method is now more widely recognised: Salzberb (1997), for example, points out both the incorrectness of using a regular t-test on two accuracy scores, and the surprisingly widespread use of it in the machine learning community at that time. One issue is when the classifiers are evaluated on the same data

set, so a test that requires independent data is appropriate. He consequently defines an approach that uses a paired test—a paired t-test, McNemar test or similar—and k-fold cross-validation. The survey on cross-validation by Refaeilzadeh et al. (2009) notes that more complicated proposals, such as various $n \times k$-fold cross-validation approaches have not yet been widely accepted. We therefore use a simple 5-fold cross-validation.

With respect to the statistical test, our data, with their relatively high baselines as noted in Section 4.2 but with some quite low values as well, appears to be quite skewed. The Gaussian-based t-test is therefore not suitable, and we use the non-parametric McNemar test (Sprent and Smeeton, 2007) instead.[4]

4.3.1 The models

To construct our unigram features, we consider every other word in the context of the gap as a feature used to predict the correct word for the gap, giving us a feature space equal to the number of distinct tokens in the corpus. In producing the features for each set of words, we excluded all words in the set being predicted. We also, following Pang et al. (2002), excluded all tokens that did not occur at least 4 times in the training data. We did not use stemming or stop lists.

We tested four possible unigram models:

1. the *frequency* of each word in the document containing the gap (DocFreq)
2. the *presence* of each word in the document containing the gap (DocPres)
3. the *frequency* of each word in the sentence containing the gap (SentFreq)
4. the *presence* of each word in the sentence containing the gap (SentPres)

4.3.2 Results and discussion

The increase in prediction accuracy for each of these four baselines over the *most frequent* baseline is shown in Table 4.8.

First, as seen in Chapter 3, we again observe that the *most frequent* baseline itself differs for near-synonym sets with and without attitudinal meaning, and also when that meaning is the same or differs among the near synonyms. Observe in particular that the performance of the *most frequent* baseline is lower for attitudinal near-synonym sets, providing some support for suggesting that choosing between these may be a more difficult

[4] Foody (2008) gives a good overview of the use and applicability of these tests in the comparable field of machine learning for imaging.

	Overall			No affect			Same affect			Differing affect		
Unigram model	dev.	test	both	dev.	test	both	dev.	test	both	dev.	test	both
MF baseline	72·3	73·8	73·1	82·6	85·2	83·8	44·8	71·4	69·6	50·9	57·1	54·2
DocFreq	+3·2	+1·3	+2·1	+0·0	+1·5	+0·7	+10·0	+1·5	+2·1	+9·9	+0·7	+5·0
DocPres	+5·3‡	+2·0‡	+3·4‡	+0·2†	+2·1‡	+1·0‡	**+13·5‡**	**+2·4‡**	**+3·1‡**	+16·4‡	+1·4†	+8·4‡
SentFreq	+8·8	+4·6	+6·5	+2·5	+4·1	+3·3	+10·3	+1·5	+2·1	+23·1	+8·5	+15·3
SentPres	**+9·5$^{\ddagger\vee}$**	**+4·8$^{\ddagger\vee}$**	**+6·8$^{\ddagger\vee}$**	**+2·9$^{\ddagger\vee}$**	**+4·4$^{\ddagger\vee}$**	**+3·6$^{\ddagger\vee}$**	+11·6‡	+1·7‡	+2·3$^{\ddagger\times}$	**+24·2$^{\ddagger\vee}$**	**+8·7$^{\ddagger\vee}$**	**+15·9$^{\ddagger\vee}$**

Bold values are best performance for that column

* Difference from *most frequent* baseline significant at the $p < 0.05$ level (shown for DocPres and SentPres only)

† Difference from *most frequent* baseline significant at the $p < 0.01$ level (shown for DocPres and SentPres only)

‡ Difference from *most frequent* baseline significant at the $p < 0.001$ level (shown for DocPres and SentPres only)

$^{\times}$ Difference from DocPres significant at the $p < 0.05$ level (shown for SentPres only)

$^{\cup}$ Difference from DocPres significant at the $p < 0.01$ level (shown for SentPres only)

$^{\vee}$ Difference from DocPres significant at the $p < 0.001$ level (shown for SentPres only)

Table 4.8: Percentage increase of unigram models over *most frequent* baseline

task, perhaps partly influenced by the somewhat closer semantic relationships discussed in Section 4.1.2.2.

Second, like Pang et al. (2002) on this same data set, we have found that *presence* features equal or outperform *frequency* features, even though we are performing a fairly different task, lexical gap prediction rather than classifying documents by sentiment. The improvement of *presence* over *frequency* is less dramatic at the sentence level than the document level, presumably because most tokens will only occur at most once in a sentence in any case.

Third, the improvement of only using tokens in the sentence surrounding the gap over using tokens in the entire document in general echoes the result of Edmonds, that using features from a wide window (in his case, 50 words) around a gap diminished performance over using a smaller window. This is similar to other tasks such as word sense disambiguation (WSD): the only survey comprehensively discussing the issue of window size in WSD (Ide and Véronis, 1998) noted that the use of 'micro-context' provided most benefit, and the use of 'topical context' was variable and generally minimal.

Fourth, and most interestingly in that it is new and is relevant to our hypothesis, and not reflected in the literature to date, is that there is an exception to the second point, which is that *same-affect* words are more accurately chosen using *document frequency* rather than *sentence frequency*.

A possible explanation for this is that, whereas *no-affect* and different-affect have sufficient information at the sentence level, for *same-affect* sets there is a 'tone' that suffuses the document that is important in replicating word choice. As an example from the Movie Review Corpus of this kind of tone distributed throughout a document (our italics): "even though the film *suffers* from its *aloof* and *uninviting* approach … the *problem* with the picture seems on the surface to be its *plodding* pacing, but actually the *defect* has to do more with …".

Of the twelve sets of *same-affect* words in the development and test set, with 4751 instances in the corpus, only the smallest set (with 35 instances) was of positive affect; the other eleven sets and 4716 instances were negative. That is, there was vastly more negative language, and the somewhat low baseline for these (69.6%) possibly suggests some variety in insults used to criticise the movies. This would fit with the work of Wiebe et al. (2004) on subjectivity, where they found that subjective opinion pieces exhibited a greater amount of linguistic 'creativity' — "Apparently, people are creative when they are being opinionated" — as evidenced by aspects such as higher frequency of *hapax legomena*. This then suggested to us two possible ways that this 'tone' might be manifest in the

	aggregate	uniform
author #1	0·032	0·121
author #2	0·086	0·234
author #3	0·041	0·088
author #4	0·010	0·180

Table 4.9: Kullback–Leibler divergences for authors vs aggregate and uniform distributions

movie reviews: in the overall sentiment of the document, or through the writing style of a particular author.

That document sentiment might be useful is intuitive, and part of our reason for reviewing document sentiment analysis techniques in Section 2.5. The result that document-level classifiers did worse than sentence-level ones on different-affect sets of words (as opposed to *same-affect* ones) went against that intuition, and led to the authorial style idea inspired by the work of Wiebe et al.. To check this idea that a particular author's style might be distinguishable (and therefore perhaps useful in detecting this creative choice of *same-affect* words), we did a quick analysis of the development set. With most of the *same-affect* instances being negative, we looked at the distribution of negative words as broken down by the 10-point rating scale (from 0.1 to 1.0), for each author. We then compared each distribution against the aggregate distribution for all authors and against a uniform distribution as measured by their Kullback–Leibler divergence (Kullback and Leibler, 1951); results are in Table 4.9. Kullback–Leibler divergence (discussed in more detail in Chapter 5) gives a value for the difference between two distributions, or rather, the inadequacy of one distribution as a model for another.

Note that scores for individual authors vary by a factor of 4 (vs uniform) or 8 (vs aggregate). There is no generally agreed interpretation of absolute Kullback–Leibler divergence values, but the point to be drawn here is that some authors are much more different from the typical case in their use of negative words than are others. Also of note is that use of particular near synonyms and other similar linguistic phenomena where language allows a certain amount of choice at a local level is useful in the opposite task: given a document with certain features, identify its author (Koppel et al., 2006).

We discuss our models incorporating document sentiment and author information in the next section.

4.4 Sentiment-derived features with unigrams

4.4.1 Document sentiment

We constructed two types of models to incorporate document sentiment. The first was just to use the sentiment given in the movie review corpus. This is a gold-standard sentiment, but does not contribute many features. Our second type of model was to use the sentiment of individual words in the document, the sort of feature often used in document sentiment classification; for this, we used scores from MicroWNOP (Cerini et al., 2007) (see Section 2.5.4.3).

4.4.1.1 Gold-standard sentiment

The first set of features, DocSent, is as follows:

- the sentiment of the document in question, as assigned by the Subj measurement of the Scale 1.0 corpus; and
- the sentiment of the target word, as assigned in the annotation (Section 4.1.2).

4.4.1.2 Approximate sentiment

This set of features, MicroWNOP includes the features from Section 4.4.1.1 and adds features from MicroWNOP. We construct the following definition of a single MicroWNOP Positive score and MicroWNOP Negative score for a word. The MicroWNOP Positive score for a word is the highest single Positive score assigned to a synset containing that word, whether assigned in the Common, Group1 or Group2 categories. The MicroWNOP Negative score is the equivalent value for the negative scores.

We then define four features, intended to be a proxy for the document sentiment:

- the sentiment of the target word, as assigned by annotators in Section 4.1.2;
- the average of the MicroWNOP Positive scores of all of the words in the document, excepting the target word;
- the average of the MicroWNOP Negative scores of all the words in the document, excepting the target word; and
- the total MicroWNOP Positive scores of all the words in the document, excluding the target, minus the MicroWNOP Negative scores of all the words in the document, excluding the target.

4.4.2 Author identity

Author identity is also given in SCALE 1.0. We thus define AUTHORID, four binary features representing each of the four authors in the SCALE 1.0 corpus, positive when a particular author wrote the review in which the gap occurs.

4.4.3 Results and discussion

Results comparing various combinations of the above features with unigram features are shown in Table 4.10 as an accuracy rating and a percentage increase over the unigram baseline. Features were only tested in conjunction with DOCPRES and SENTPRES, as the better performing unigram baselines in Section 4.3.

Table 4.10 shows that for the most part these features have little to no impact on prediction accuracy. However, AUTHORID (whether by itself of in conjunction with DOCSENT) always produces the best results. In three of the cases it is only by a small margin, with the exception being the case of AUTHORID on the *sentence frequency* classifier. The addition of AUTHORID to the *sentence presence* unigrams, which outperform *document presence* unigrams on the other word set classes, comes closest to approximating the *document presence* results on the same affect word sets.

To look further into our intuition that knowledge of the author reflects the linguistic creativity discussed above, we examined the impact of a *document frequency* (DF) threshold (Yang and Pedersen, 1997). A major conclusion of Wiebe et al. (2004) was that rare events such as *hapax legomena* contain a lot of information for subjective texts, and so feature selection such as by DF thresholding would be harmful. Therefore, we might expect that DF thresholding would worsen results here if this linguistic creativity is what has led to the *same-affect* results. For this, we examined the development set (which has similar overall results on the various classifiers). Using *document frequency* thresholding values of 2, 4 and 8 caused small decreases (typically around +0.1% at a DF of 8) in the performance of all of the features above. The small magnitude of these suggested that the source of the *same-affect* results might be elsewhere. In the event that author identity had turned out to be a useful feature here, that would have indicated a need to approximate author identity, in order for the method to be effective on unseen text by unknown authors. However that is not indicated by these results.

Additional features	Overall dev.	Overall test	Overall both	No affect dev.	No affect test	No affect both	Same affect dev.	Same affect test	Same affect both	Differing affect dev.	Differing affect test	Differing affect both
None (unigram only)												
DocPres	77·6	75·8	75·6	82·8	87·3	84·9	58·4	**73·8**	**72·8**	67·3	58·6	62·6
SentPres	81·7	78·6	80·0	85·5	89·6	87·4	56·5	73·1	72·0	**75·1**	65·9	70·1
AuthorID												
DocPres	+0·0	+0·0	+0·0	−	+0·0	+0·0	+0·3	−0·0	0·0	+0·0	+0·0	+0·0
SentPres	**+0·0**	+0·4‡	**+0·3‡**	**+0·1**	**+0·1**	**+0·1**	**+1.3**	+0·8‡	+0·8‡	−0·2	+0·6⋆	**+0·3**
MicroWNOP												
DocPres	+0·0	+0·0	+0·0	−0·1	−0·1	−0·1⋆	−1·0⋆	−0·1	−0·1	+0·4⋆	+0·4⋆	+0·4†
SentPres	−1·4‡	−1·5‡	−1·4‡	−0·3†	−1·1‡	−0·7‡	−1·9	−0·1	−0·2	−3·8‡	−3·4‡	−3·6‡
DocSent												
DocPres	+0·0	−0·0	+0·0	−	−	−	−	+0·0	+0·0	+0·1	+0·0	+0·0
SentPres	−0·1⋆	+0·1	−0·0	−0·0	−0·0	−0·0	+0·3	+0·0	+0·0	−0·4⋆	+0·3⋆	−0·0
MicroWNOP and AuthorID												
DocPres	+0·0	+0·0	+0·0	−0·1	−0·1	−0·1	−1·0⋆	−0·0	−0·1	+0·4⋆	+0·3	+0·3⋆
SentPres	−1·4‡	−0·8‡	−1·1‡	−0·3†	−0·9‡	−0·6‡	+0·6	+0·8‡	+0·8‡	−3·9‡	−2·4‡	−3·1‡
DocSent and AuthorID												
DocPres	+0·0	+0·0	+0·0	−	+0·0	+0·0	**+0·0**	−0·0	−0·0	+0·1	+0·1	+0·1
SentPres	−0·1	**+0·5‡**	+0·3‡	+0·0	+0·1	+0·1	+1·9	+0·9‡	+1·0‡	−0·5†	**+0·8†**	+0·2

Bold values are best performance for that column

A − means model's choice of word was identical to the unigram *presence* model, +0.0 means an increase of < 0.05, −0.0 means a decrease with absolute value < 0.05

⋆ Difference from unigram model significant at the $p < 0.05$ level

† Difference from unigram model significant at the $p < 0.01$ level

‡ Difference from unigram model significant at the $p < 0.001$ level

Table 4.10: Performance of SVMs using unigrams with additional features, compared to unigram *presence* models

4.4.4 Conclusion

In this section, we have explored the idea that sentiment-derived features might improve upon the performance of unigram features in solving the FITB task. Unfortunately, we have found that their contribution is neglible at best, even on near synonyms that have sentiment or differ in sentiment, and in texts that convey opinions. Either our selection of features is not capturing sentiment in a useful way, or it cannot be straightforwardly used to contribute to FITB.

4.5 Unigram models accounting for distance

4.5.1 Distance measure

Our finding in Section 4.3 that *sentence presence* features usually outperform *document presence* features coheres with the finding of Edmonds (1997) that a very small window around the data provided better choice accuracy. (Subsequent authors have tended not to evaluate very wide windows in the first place.) Presumably, the noise in the more distant words overwhelms any useful information they convey. However, the finding for the *same-affect* set that *document presence* improves performance hints that document-level features can have an impact on the lexical gap task at least in some cases. This has parallels in work on the behaviour of context with respect to entropy, whose underlying principle as described by Qian and Jaeger (2010) is "that distant contextual cues tend to gradually lose their relevance for predicting upcoming linguistic signals". This idea was first presented in the context of an exploration of the behaviour of entropy by Genzel and Charniak (2002), where they propose the Constant Entropy Rate principle. By conditionally decomposing entropy with respect to local (i.e. sentence-level) and broader context, they empirically demonstrate support—through the consistently observed increase in entropy conditioned on local context throughout texts—for their principle, and conclude that broader context continues to influence later text. They develop this further in Genzel and Charniak (2003), where they find changes in entropy behaviours at paragraph boundaries, suggesting topic or other broader context characteristics are part of the effect. Their principle has subsequently been supported by work in psycholinguistics, such as that of Keller (2004), Levy and Jaeger (2007), and Gallo et al. (2008). Qian and Jaeger (2010) go on to investigate the precise type of relationship of broader context, and find that incorporating linear and sub-linear representations of broader context into entropy models improves the fit of these models of the development of entropy throughout a text.

As noted in Section 4.3, work in the structurally similar task of WSD has generally ignored broader context. Prompted by the results in the previous section and the work on discourse entropy, in this section we describe another model, in which unigram features are weighted by their distance from the gap.

Rather than using feature value 1 for *presence* or 0 for *absence*, as in the unigram models in Section 4.3, here we weight the presence of a token by its distance from the gap. For example, in the sentence fragment in example (4.5) the distance of the token *big* from the gap is 1, and the distance of *economists* from the gap is 7.

(4.5) ... some economists think it would be a big _____.

In order that features further away from the gap not be entirely eliminated by their distance from the gap, but that noise not overwhelm the information they bring, we experiment with weighting the distance using the following functions for the feature value $f(w)$ of a token w using distance $d(g, w)$ in number of tokens between w and gap g:

- Inverse linear weighting, InvLinear:

$$f(w) = \frac{1}{d(g, w)} \tag{4.10}$$

- Inverse square root weighting, InvSquareRoot:

$$f(w) = \frac{1}{\sqrt{d(g, w)}} \tag{4.11}$$

As in Section 4.3, every token in the document is considered as a feature, except those with total corpus frequency of less than 4 and the candidates to fill the gap themselves. If a token w is used more than once in a document, the largest value for $f(w)$, ie the smallest $d(g, w)$ for both InvLinear and InvSquareRoot for a given gap g is used.

We test the effectiveness of limiting the features to text surrounding the target word, using the following measures:

- every token in the document;
- every token in the sentence containing the gap and the two surrounding sentences; and
- every token in the sentence containing the gap, only.

We also test the InvSquareRoot unigram model with selected successful additional features from Section 4.4, giving us the following models:

1. InvSquareRoot
2. InvSquareRoot and AuthorID combined
3. InvSquareRoot and DocSent combined
4. InvSquareRoot, AuthorID and DocSent combined.

4.5.2 Results and discussion

Results are shown in Table 4.11 (see Appendix C for z-scores associated with the significance data shown). We see that weighting the unigrams for their distance from the gap is useful in all cases, giving very large relative improvements over the baseline (in error reduction, up to 48% for *differing-affect* sets with InvLinear). This general pattern accords with the behaviour of broader context in the entropy work discussed earlier.

While this result is to some extent expected—words closer to the gap have a higher weight, and thus the most predictive power over the word filling the gap—the most interesting result is that a wider context than sentence level remains useful. The result even extends beyond the 3-sentence level to the entire document. In at least some cases words quite far from the gap indeed are affecting the choice of near synonym. As with the unigram results from Section 4.3, this supports our hypothesis that features of the entire document influence the choice of near synonym, even though document-level features do not appear to have been well captured by our choice of features in Section 4.4.

Same-affect near synonyms also respond better to a different weighting function, InvSquareRoot rather than InvLinear. InvSquareRoot discounts the distance between a word and the gap less heavily than InvLinear does, especially, relatively speaking, at more extreme distances. This further demonstrates that more distant words are having an effect: some discounting is evidently required since InvSquareRoot outperforms no weighting, but there is an extent past which the discounting appears to under-weight features when the entire document is required as context.

As for the general utility of the weighting functions, having no weighting function results in the document context performing 3.4% worse than the single sentence context: this is consistent with the discussion in Section 4.3.2 on previous and related results, where the document context just adds noise. Weighting functions boost the results for both the single sentence context and the document context; for the *no-affect* and *differing-affect* sets and the InvLinear function, this is to the same (highest performing) level. The weighting function thus seems like a good way of ignoring noise, although it does not

Distance measure and span	Overall dev.	test	both	No affect dev.	test	both	Same affect dev.	test	both	Differing affect dev.	test	both
None												
Document	77·6	75·8	76·6	82·8	87·3	84·9	58·4	73·8	72·8	67·3	58·6	62·6
3 sentence	80·5	76·7	78·4	85·1	88·8	86·8	56·5	72·2	71·1	71·8	61·2	66·1
1 sentence	81·7	78·6	80·0	85·5	89·6	87·4	56·5	73·1	72·0	75·1	65·9	62·6
InvLinear												
Document	**+10·1**‡	**+7·8**‡	**+8·8**‡	**+7·1**‡	**+5·3**‡	**+6·3**‡	+2·6	+0·9*	+1·1*	**+17·7**‡	**+19·0**‡	**+18·4**‡
3 sentence	+7·1‡	+6·8‡	+6·9‡	+4·7‡	3·8‡	+4·3‡	+4·5*	**+2·3**‡	**+2·4**‡	+12·7‡	+16·4‡	+14·7‡
1 sentence	+5·9‡	+5·0‡	+5·4‡	+4·4‡	+3·0‡	+3·8‡	+4·2*	+1·7‡	+1·9‡	+9·6‡	+11·9‡	+10·8‡
InvSquareRoot												
Document	+7·6‡	+6·5‡	+7·0‡	+4·8‡	+4·3‡	+4·6‡	+3·5*	+1·2†	+1·4‡	+14·2‡	+15·7‡	+15·0‡
3 sentence	+5·2‡	+5·5‡	+5·4‡	+3·1‡	+2·9‡	+3·0‡	**+7·1**‡	+1·8‡	+2·2‡	+9·9‡	+13·5‡	+11·8‡
1 sentence	+4·6‡	+4·2‡	+4·4‡	+3·3‡	+2·5‡	+3·0‡	+6·8‡	+1·9‡	+2·2‡	+7·3‡	+9·5‡	+8·5

Bold values are best performance for that column
* Difference from non-distance weighted unigram model significant at the $p < 0.05$ level
† Difference from non-distance weighted unigram model significant at the $p < 0.01$ level
‡ Difference from non-distance weighted unigram model significant at the $p < 0.001$ level

Table 4.11: Performance of SVMs using distance measures, compared to unigram *presence* models

Additional features	Overall	No affect	Same affect	Differing affect
None	76·6	84·9	72·8	62·6
InvLinear	**+8·8**	**+6·3**	**+1·1**	**+18·4**
InvLinear and AuthorID	−1·3	−0·4	−1·2	−3·1
InvLinear and DocSent	−1·3	−0·4	−1·2	−3·1
InvLinear, AuthorID and DocSent	−1·3	−0·4	−1·2	−3·1

Bold values are the best performance for that column.

Table 4.12: Performance of SVMs using InvLinear with additional features, compared to the DocPres unigram model

entirely compensate for extending the context beyond what is necessary.

Results for the combination of the InvLinear weighting with other features is given in Table 4.12. In all of these case, adding the features harmful. Distance weighting thus appears to capture document tone better than the explicit features of author ID or document sentiment.

We have thus further confirmed our result that in many cases, a very wide context is useful, and we have gone some way to narrowing down exactly how to balance providing this context with weighting appropriately for noise. Further work is needed to determine how to distinguish contexts where words very distant from the gap should be included with appropriate weights, as in our *same-affect* set, and where they should be excluded entirely, or weighted even lower than InvLinear.

4.5.3 Examples of the best performing sets

The best performing five near synonym sets, relative to improvement over the baseline, are shown in Table 4.13 for both document InvLinear and sentence InvLinear. This shows that the best improvement our techniques deliver is in the order of a 25–35 percentage point improvement over the *most frequent* baseline. For both techniques, the largest improvements are in fact delivered over the full range of set types, including no affect: the only difference is the appearance of different same affect sets in the top five: *aghast* etc in the document list and *brashness* etc in the sentence list, each is seventh in the other list.

4.6 Conclusion

In this chapter, we have applied a novel supervised approach to the problem of choosing the right near synonym to fill a lexical gap. Our main conclusions from doing this are as follows. First, as per Pang et al. (2002) with document sentiment classification, unigrams alone

Words in set	Affect type	No. tests	Performance	Improvement
	Document InvLinear			
recommendation, *advice*	*no-affect*	146	91·8%	+37·7
feat, *operation*, *act*, *exploit*, *action*, *performance*	*differing-affect*	3596	85·6%	+35·5
charge, *attack*, *storm*, *assault*	*no-affect*	177	65·0%	+33·9
precise, *accurate*, *exact*, *right*, *nice*, *correct*, *true*	*differing-affect*	2181	76·9%	+30·7
aghast, *scared*, *frightened*, *afraid*	*same-affect*	187	67·4%	+26·2
	Sentence InvLinear			
recommendation, *advice*	*no-affect*	146	90·4%	+36·3
feat, *operation*, *act*, *exploit*, *action*, *performance*	*differing-affect*	3596	85·2%	+35·1
charge, *attack*, *storm*, *assault*	*no-affect*	177	62·7%	+31·6
precise, *accurate*, *exact*, *right*, *nice*, *correct*, *true*	*differing-affect*	2181	77·1%	+30·9
brashness, *brass*, *cheek*, *hide*, *nerve*	*same-affect*	70	60·0%	+25·7

Table 4.13: Best performing five sets for each of document and sentence InvLinear, relative to baseline performance

do well, with *presence* outperforming *frequency*, and with immediate-context (sentence) models generally outperforming wider-context (document) models. Second, that once appropriate weighting of distance features are incorporated, the technique performs near synonym choice notably better with document rather than sentence features; this may be a consequence of a particular 'tone' suffusing the document. Third, adding in one possible factor related to this tone, knowledge of the author of the text, gives slightly better results overall, in particular improving *sentence presence* results for *same-affect* near synonyms; this author effect was not expected at the start of the work. Fourth, the most significant improvement came from incorporating document-level information using a weighting scheme, which in fact improved over the earlier best sentence-level models, and which in its description of the effect of broader context mirrors work on the effect of the Constant Entropy Rate principle. This is true for all near-synonym sets, including the non-affect ones. It is still not yet clear how precisely this author and distance information are causing this improvement.

Chapter 5

Valence shifting text

In preceding chapters, we have seen that existing and new techniques addressing the Fill In the Blanks (FITB) task behave differently for near synonyms that have an affective aspect to their meaning, and in particular that the problem of distinguishing between near synonyms which have the *same* affect is difficult.

In particular, the baseline performance for distinguishing between near synonyms with affective meaning is low, so even large improvements in performance yield a comparatively low rate of absolute correctness. Yet, the same low most-frequent baseline suggests that it is particularly important to distinguish between such near synonyms, as there is no obvious "default" word to choose. *If* these words are all true synonyms as discussed in Section 4.1.2.2 this won't arise but the WordNet evidence presented there suggests that although some are closely related, they largely aren't true synonyms.

It also intuitively makes sense to hypothesize that choosing the right level of affect is important to authors in achieving their desired meaning: that choosing between, for example, *bad*, *awful* and *abominable* is an important choice in conveying a message. Consider the difference in meaning between these sentences, based on one from the Scale dataset v1.0 movie review data set (Scale 1.0) (Pang and Lee, 2005):

(5.1) If we have to have another parody of a post-apocalyptic America, does it have to be this *bad*?

(5.2) If we have to have another parody of a post-apocalyptic America, does it have to be this *awful*?

(5.3) If we have to have another parody of a post-apocalyptic America, does it have to be this *abominable*?

Original text	Stupid, infantile, redundant, sloppy, over-the-top, and amateurish. Yep, it's "Waking Up in Reno." Go back to sleep.
Lexical substitution	*Silly*, *juvenile*, *repetitive*, sloppy, *exaggerated*, and amateurish…
Word removal	*[…]*, *[…]*, redundant, *[…]*, *[…]*, and amateurish…
Significant rewrite	"Waking Up in Reno" is a slapstick comedy with low production values.

Table 5.1: Techniques for producing less negative paraphrases of example 5.4.

As we saw in Section 2.6, this problem is known as VALENCE SHIFTING: paraphrasing text in order to change its sentiment. Of course, as paraphrasing problem, valence shifting need not be limited to the lexical level. It could be addressed by any of the following paraphrasing techniques:

Lexical substitution Replacing individual words with a less negative closely related word, for example *terrible* to *bad*, or *mistake* to *incident*.

Word removal Removing a portion of more negative words, for example *terribly unfortunate mistake* to *unfortunate mistake*

Significant re-write Paraphrase at more than the lexical level, including syntactic changes or syntactic changes with lexical replacement, for example *a disaster of catastrophic proportions* might become *a difficult time* or even *an incident under investigation*.

Consider re-writing another sentence from SCALE 1.0. Valence-shifted paraphrases of example (5.4) using each of the techniques listed about are shown in Table 5.1.

(5.4) Stupid, infantile, redundant, sloppy, over-the-top, and amateurish. Yep, it's "Waking Up in Reno." Go back to sleep.

However, in this chapter[1] we limit ourselves to exploring the possibilities of valence shifting by lexical substitution rather than exploring paraphrasing techniques. Lexical choice remains an important problem: in addition to the use cases for valence shifting itself given in Section 5.1.1, ultimately Natural Language Generation itself needs high quality lexical choice techniques for the case of words with affective meanings, and NLG, at least using the present pipeline as described in Section 2.3, cannot rely solely on the kinds of statistical techniques used to address the FILL IN THE BLANKS (FITB) task in recent years, and in Chapters 3 and 4: typically the surface realisation of the output is not generated leaving only gaps for certain near-synonym clusters, thus an NLG system cannot rely on surrounding lexical items for the word choice.

[1] Work described in this chapter was also published in Gardiner and Dras (2012).

In addition, recall from Section 2.6.1 that authors including Inkpen et al. (2006) and Whitehead and Cavedon (2010) have had some trouble with inter-judge agreement when evaluating their work. It is therefore not even completely clear that the valence shifting task is a well-defined task, although it intuitively ought to be, given both that writing affective text has an enormous history, and the success of sentiment analysis at identifying sentiment, it ought to be the case that features can be identified that change both the judgement of automatic and of human sentiment judges. This encourages us to simplify the task in order to show that a fairly simple version of valence shifting achieves judge agreement on the direction of the change.

We therefore explore two questions:

1. Is it in fact true that altering a single lexical item in a sentence noticeably changes its polarity for readers?

2. Is there a quantitative measure of relative lexical valence within near-synonym sets that corresponds with human-detectable differences in valence?

We investigate these questions for negative words by means of a human experiment, presenting readers with sentences with a negative lexical item replaced by a different lexical item, having them evaluate the comparative negativity of the two sentences. we then investigate the correspondence of the human evaluations to certain metrics based on the similarity of the distribution of sentiment words to the distribution of sentiment in a corpus as a whole.

In Section 5.1 of this chapter, we discuss the definition of the valence shifting problem further, with an emphasis on defining it sufficiently for human evaluation of valence shifting; in Section 5.2 we discuss possible metrics for quantifying the relative valence of near synonyms in terms of their corpus distribution; in Section 5.3 we describe a human experiment in valence shifting, confirming that lexical valence shifting under certain parameters behaves as expected; in Section 5.4 we describe test data for the human experiment; in Section 5.5 we discuss the results of the human experiment; in Section 5.6 we investigate the effectiveness of the metrics described in in approximating them; and in Section 5.7 we discuss our success in confirming that lexical valence shifting behaves as expected and in discovering a possible avenue for measuring the valence-shifting capability of near synonyms.

5.1 Difficulties defining and solving the valence shifting problem

5.1.1 Use cases

The earlier example of rewriting *a disaster of catastrophic proportions* to *an incident under investigation* introduces a task definition problem: at what stage does a less negative text become an entirely different text? It is probably impossible to answer this question in the general case; it depends upon the use-case, and to what extent the negativity is key to the text.

For example, if one was to refer to a nuclear meltdown, and attempt to minimise panic (or liability concerns), rewriting *a disaster of catastrophic proportions* to *an incident under investigation* may be a desired change (although perhaps not an ethical one in many circumstances).

Any application which provides stylistic or semantic assistance with writing could use a module which moderates the negativity of text. Examples of possible use-cases for reducing the negativity of text include:

- writing texts in liability-sensitive areas such as journalism;
- writing texts in the form called "PR-friendly" (or at the extreme, "doublespeak");
- assisting writers working in a non-native language with achieving the correct level of negativity; and
- rewriting, paraphrasing or summarising text where the target text is in a different genre, particularly from subjective original genres such as movie reviews to objective genres like analysis or encyclopedia articles.

5.1.2 Faithfulness

To some extent, reducing the negativity of a text must almost inevitably alter the semantic content, and thus there is a question of at what point one is altering an existing text versus authoring an entirely new text on the same topic.

At the extreme, consider where the negativity could be considered so key to the text that rewriting it as less negative essentially involves writing an entirely new text on a similar subject. Many movie reviews are much like this, for example, consider this extract of a review of *Funny Games US*, a horror film:

> A cold thriller, exploring the exploitation of violence by heaping up a huge quantity of exploitive violence itself to make some points about what is the public's attitude and tolerance level for the random violence that is present in society… Paul calls Peter "fatty" several times to the mock displeasure of Peter, Tom, Beavis, or whoever this obese, snide monster really is… I stayed to see if this sadistic film works its point of view out , and I was disappointed that with all its novelty and games and pretenses to high art, it offered nothing much in the way of seeing things differently…

It is possible to rewrite this along the lines of "A cool thriller, exploring the portrayal of violence by portraying a moderate amount of violence itself to make some points..." but the major point of the text is to express strong disgust at and condemn the movie. Whether any use-case would be served by softening this text is unclear.

5.1.3 Subjectivity

A difficult problem in this task is to alter negative text that is judgemental. As discussed in Section 2.5.1.2, even truly polarised sentiment may be used in different ways. In subjectivity analysis the primary distinction is between SUBJECTIVE and OBJECTIVE, in which a speaker either self-reports an internal state (such as a negative opinion) or is describing the observable state of an object in the discourse. Depending on application, it may be desirable to shift the valence of only subjective or only objective statements. Consider altering the sentiment of the subjective example (5.5) as opposed to example (5.6), where all the words with negative sentiment are objective descriptions. For additional complexity, consider also the case of example (5.7), where the objective assessment *dour* is a factor influencing the subjective opinion *excessively earnest*.

(5.5) They should have chosen to make artistic films, instead of taking this more *exploitive* and *superficial* path.

(5.6) It is a film about urban white people: their *fears*, their *angers*, their *prejudices*, their *repressions*, and their eroticisms.

(5.7) The *excessively earnest* tale features a host of *dour* characters.

As an even more complex example consider the *Funny Games US* review introduced in Section 5.1.2. In this review *violence* in "a huge quantity of exploitative violence" and *displeasure* in "the mock displeasure of Peter" are actually descriptive terms rather than judgemental terms, they just happen to be describing something usually viewed as

negative. This text is especially complex because the author's negative opinion of the movie is also present in very close proximity to relatively neutral descriptions of its horror elements. Here is the quote again, where **bold text** indicates negative authorial opinion and *italic text* descriptions of the film's content or artistic purpose.

> A **cold** *thriller*, exploring the *exploitation* of *violence* by heaping up **a huge quantity of exploitive violence** itself to make some points about what is the public's attitude and tolerance level for *the random violence that is present in society*… Paul calls Peter *fatty* several times to *the mock displeasure* of Peter, **Tom, Beavis, or whoever this obese, snide monster really is**… I stayed to see if **this sadistic film** works its point of view out , and I was **disappointed** that with all its novelty and games and **pretenses to high art**, it offered **nothing much** in the way of seeing things differently…

5.1.4 Context-dependence

Intuitively, one of the chief difficulties of solving the valence-shifting problem is that of *context*: specifically that the polarity and strength of a statement may be highly dependent on domain, genre conventions, the speaker's or writer's personal style, and on local features. Consider for example negation in a loose paraphrase of example (5.1), where *not* transforms *be good* into a criticism:

(5.8) If we have to have another parody of a post-apocalyptic America, *does it have to be this bad*?

(5.9) If we have to have another parody of a post-apocalyptic America, *could it not be good*?

Likewise sarcasm presents difficulties as do counter-factals, with the sentence in example (5.10) using positive terms *good* and *great* in contrast to the reality of the film:

(5.10) *If* its story and its characters had been as consistent as its *good* intentions, it might have been a *great* film.

In example (5.11) *great* is not being used in its positive sense at all:

(5.11) He seems to have a *great* deal of fun mocking the image of the sullen detective.

These difficulties are very similar to unresolved difficulties in sentiment detection at the sentence level, with the review of Liu (2012, pp. 43–45) discussing several recent ap-

proaches to the more difficult problems. Narayanan et al. (2009) argue that no global approach to sentence sentiment classification (and therefore presumably identifying features for valence-shifting) is possible, and specifically focus on conditional sentences. Tsur et al. (2010) describe their work as the first approach to sarcasm classification, with their top performing system having precision of 91.2% on their test set: they note though that even human annotators do not strongly agree on sarcasm identification ($\kappa = 0.34$). González-Ibáñez et al. (2011) experimented with sarcasm detection of utterances on Twitter, again experiencing difficulty with inter-annotator agreement (with three judges agreeing on the classification of only about half their test set) and classifier accuracy as low as 43.3% on sentences the judges agreed on.

For a universal solution to valence shifting, all of these challenges with sentence sentiment detection would need to be met. Valence shifting is therefore clearly a difficult problem. In order to explore valence shifting in ideal conditions, throughout this chapter we deal with a dataset from which some of the more obviously difficult contexts have been removed, as described in Section 5.4.2.

5.2 Quantifying lexical valence

As discussed in the introduction to this chapter, a sensible preliminary hypothesis is that in the main replacing words with more negative words will render a text more negative. For example, we might transform the sentence in example (5.12) into example (5.13), which we argue is a more negative sentence:

(5.12) *Hideaway* is pretty *poor* entertainment, and what starts out as a superficial trip into the occult ends with a pointless, overblown fight to the death.

(5.13) *Hideaway* is pretty *excruciating* entertainment, and what starts out as a superficial trip into the occult ends with a pointless, overblown fight to the death.

If this holds, it should also be true that, statistically, a near synonym is more negative than another if it is associated with more negative contexts. In this section we discuss some possible measures of the negativity of a near synonym's contexts.

5.2.1 Related work

In Section 2.5.4 we described the many sentiment annotated word lists available. In addition Mohammad and Turney (2010, 2012) describe in detail the creation of EmoLex, a large polarity lexicon, using Mechanical Turk. Mohammad and Turney, rather than

Question	Possible answers
Which word is closest in meaning (most related) to startle?	{*automobile*, *shake*, *honesty*, *entertain*}
How positive (good, praising) is the word *startle*?	*startle* is {not, weakly, moderately, strongly} positive
How negative (bad, criticizing) is the word *startle*?	*startle* is {not, weakly, moderately, strongly} negative
How much is *startle* associated with the emotion {joy,sadness,...}?	*startle* is {not, weakly, moderately, strongly} associated with {joy,sadness,...}

Table 5.2: Sample annotation question posed to Mechanical Turk workers by Mohammad and Turney (2012).

asking annotators to evaluate words in context as we are proposing here, instead ask them directly for their analysis of the word, first using a synonym-finding task in order to give the worker the correct word sense to evaluate. Part of a sample annotation question given by Mohammad and Turney (2012) is given in Table 5.2. The word source used is the *Macquarie Thesaurus* (Bernard, 1986).

Mohammad and Turney include General Inquirer and WordNet Affect words in EmoLex to allow comparisons with the existing lexicons. Their analysis is primarily interested in the correctness of the annotator's negative vs positive distinctions, as they argue this is usually sufficient for their target applications in sentiment analysis. The evaluation of the polarity annotations show that while there is a substantial tendency towards "mixed" interpretations of words, for example, of the negative words from the General Inquirer, their annotation found 83 of them to be consistently found to be negative, but another 85 to be marked as both negative or positive by different annotators (and 1 word was found to be positive by all annotators). Likewise, 82 General Inquirer words were uniformly marked positive, but 84 either positive or negative (and 2 uniformly negative). Nevertheless, they find that their negative annotations have a κ score of 0.62 and positive 0.45, indicating strong and moderate agreement respectively.

Our work differs from that of Mohammad and Turney in that we rely on substitution evaluations, that is, having human judges rate specific contexts rather than supply their intuitions about the meaning of a word. As discussed in Section 2.2.2, Callison-Burch (2007, Section 4.1) argued for this evaluation of paraphrases. He writes:

> Because [in this thesis] we generate phrasal paraphrases we believe that the most natural way of assessing their correctness is through substitution, wherein we replace an occurrence of the original phrase with the paraphrase. In our evaluation we asked judges whether the paraphrase retains the same meaning

> as the phrase it replaced, and whether the resulting sentence remains grammatical. The reason that we ask about both meaning and grammaticality is the fact that what constitutes a "good" paraphrase is largely dictated by the intended application. For applications like information retrieval it might not matter if some paraphrases are syntactically incorrect, so long as most of them are semantically correct. Other applications, like natural language generation, might require that the paraphrases be both syntactically and semantically correct.

In our case, we are attempting to assess the effectiveness of valence-shifting, and we cannot pre-suppose that intuitions by the raters along the lines of feeling that the meaning of a word is more negative than that of another word translates into perceiving the desired effect when a word is used in context.

5.2.2 Measures of distribution

Our intuition is that words that make text more negative will tend to disproportionately be found in more negative documents, likewise words that make text less negative will tend to be found in less negative documents.

In order to quantify this, consider this as a problem of distribution. Among a set of affective documents such as SCALE 1.0, there is a certain, not necessarily even, distribution of words: for example, a corpus might be 15% negative, 40% neutral and 45% positive by total word count. However, our intuition leads us to hypothesise that the distribution of occurrences of the word *terrible*, say, might be shifted towards negative documents, with some larger percentage occurring in negative documents.

We then might further intuit that words could be compared by their relative difference from the standard distribution: a larger difference from the distribution implies a stronger skew towards some particular affective value, compared to word frequencies as a whole. (However, it should be noted that this skew could have any direction, including a word being found disproportionately among the neutral or mid-range sentiment documents.)

We thus consider two measures of differences of distribution: information gain and Kullback–Leibler divergence.

5.2.2.1 Information gain

Information gain is a measure that originated in information theory. The information gain $G(Y|X)$ associated with a distribution Y given the distribution X is the number of bits

saving in transmitting information from Y if X is known. A high information gain value thus suggests a strong predictive relationship between X and Y.

Information gain is typically used for feature selection in machine learning: a feature with a large information gain value is likely to be a useful feature on which to split a problem into pieces, since information gain in this case measures the expected reduction in entropy resulting from the partitioning (Mitchell, 1997). Yang and Pedersen (1997) evaluated information gain for feature selection for the problem of text classification, finding it one of the more effective methods.

Yang and Pedersen give an information gain formula for the gain G of term t in predicting categories $c_1 \dots c_m$ as follows, where $P_r(c_i)$ is the relative probability of category c_i, $P_r(c_i|t)$ the relative probability of c_i given term t and $P_r(c_i|\bar{t})$ the relative probability of c_i when term t is absent:

$$\begin{aligned} G(t) = &- \sum_{i=1}^{m} P_r(c_i) \log P_r(c_i) \\ &+ P_r(t) \sum_{i=1}^{m} P_r(c_i|t) \log P_r(c_i|t) \\ &+ P_r(\bar{t}) \sum_{i=1}^{m} P_r(c_i|\bar{t}) \log P_r(c_i|\bar{t}) \end{aligned} \tag{5.1}$$

5.2.2.2 Kullback–Leibler divergence

Cover and Thomas (1991) (as cited in Weeds (2003)) describe the Kullback–Leibler divergence (Kullback and Leibler, 1951) as a measure of "the inefficiency of assuming that the distribution is q when the true distribution is p". Weeds gives the formula for Kullback–Leibler divergence as:

$$D(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} \tag{5.2}$$

Weeds evaluated measures similar to Kullback–Leibler divergence for their usefulness in the distributional similarity task of finding words that share similar contexts. Our task is not an exact parallel: we seek the relative skewedness of words.

5.2.2.3 Worked example

Consider this worked example based on Table 5.3 in which we have three sentiment categories of documents: Negative, Neutral and Positive, and five texts (sentences in this case) within distributed among them.

Category	Word count	Example sentences
Negative	11	The production quality was barely *okay*
		Otherwise the movie was *bad*.
Neutral	16	The movie was *okay* I guess.
		The lead actress was *good* and the others were *okay*.
Positive	15	The acting was *good* and the direction quite *okay*
		Everything about this movie was *good*.
Total	42	

Table 5.3: Example sentences in different sentiment categories containing the words *bad*, *okay* and *good*.

Feature	Negative		Neutral		Positive	
Word count	11	26%	16	38%	15	36%
bad count	1	100%	0	0%	0	0%
okay count	1	25%	2	50%	1	25%
good count	0	0%	1	33%	2	67%

Table 5.4: Distribution of features among the Negative, Neutral and Positive categories in Table 5.3

Table 5.4 shows the distribution of features among the different categories. We could then consider the information gain and Kullback–Leibler divergence associated with each of the *bad*, *okay* and *good*.

As examples, we consider *bad* and *okay*. The information gain of *bad* is given by equation (5.3)[2] and that of *okay* by equation (5.4):

$$\begin{aligned} G(bad) = &-\left(\frac{11}{42}\log\frac{11}{42} + \frac{16}{42}\log\frac{16}{42} + \frac{15}{42}\log\frac{15}{42}\right) \\ &+ \frac{1}{42}\left(\frac{1}{1}\log\frac{1}{1} + \frac{0}{1}\log\frac{0}{1} + \frac{0}{1}\log\frac{0}{1}\right) \\ &+ \frac{41}{42}\left(\frac{10}{41}\log\frac{10}{41} + \frac{16}{41}\log\frac{16}{41} + \frac{15}{41}\log\frac{15}{41}\right) \\ \approx &3.27 \times 10^{-2} \end{aligned} \tag{5.3}$$

$$\begin{aligned} G(okay) = &-\left(\frac{11}{42}\log\frac{11}{42} + \frac{16}{42}\log\frac{16}{42} + \frac{15}{42}\log\frac{15}{42}\right) \\ &+ \frac{4}{42}\left(\frac{1}{4}\log\frac{1}{4} + \frac{2}{4}\log\frac{2}{4} + \frac{1}{4}\log\frac{1}{4}\right) \\ &+ \frac{38}{42}\left(\frac{10}{38}\log\frac{14}{38} + \frac{14}{38}\log\frac{14}{38} + \frac{14}{38}\log\frac{14}{38}\right) \\ \approx &3.70 \times 10^{-3} \end{aligned} \tag{5.4}$$

We thus see that *bad* has a higher information gain than *okay* relative to the example corpus in Table 5.3, which corresponds to *bad* being a more informative feature, as we'd

[2]The logarithm of 0 is undefined, however, per Weeds (2003), for the purposes of calculating information gain and Kullback–Leibler divergence $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$.

expect since it only occurs in one category of document. In our use case, we would hope that *bad*'s strong association with a particular class of document relative to *okay* implies that it is more strongly affective than *okay*, although as cautioned previously a strong association with a neutral document class would also result in a high information gain score.

The Kullback–Leibler divergence of the distribution of *bad* compared to all words is given by equation (5.5) and that of *okay* compared to all words is given by equation (5.6):

$$\begin{aligned} D(bad||\text{words}) =& \frac{1}{1}\log\frac{\frac{1}{1}}{\frac{11}{42}} + \frac{0}{1}\log\frac{\frac{0}{1}}{\frac{16}{42}} + \frac{0}{1}\log\frac{\frac{0}{1}}{\frac{15}{42}} \\ \approx& 1.34 \end{aligned} \tag{5.5}$$

$$\begin{aligned} D(okay||\text{words}) =& \frac{1}{4}\log\frac{\frac{1}{4}}{\frac{11}{42}} + \frac{2}{4}\log\frac{\frac{2}{4}}{\frac{16}{42}} + \frac{1}{4}\log\frac{\frac{1}{4}}{\frac{15}{42}} \\ \approx& 0.04 \end{aligned} \tag{5.6}$$

We again see a higher value for the value of $D(bad||\text{words})$ than we do for $D(okay||\text{words})$, reflecting its tendency to be associated with a particular class (or, in the Cover and Thomas (1991) formulation previously cited, the insufficiency of general words in the corpus to substitute for *bad*. Observe that the Kullback–Leibler divergence isn't a symmetric value, due to the asymmetry of the use of $p(x)$ and $q(x)$ in equation (5.2). This formulation measures the fitness of words to substitute for *bad*, $D(\text{words}||bad)$ would measure the ability of *bad* to substitute for other words.

5.2.3 Measure of corpus-centrality: inverse document frequency statistic (IDF)

In addition to the measures of distribution considered in Section 5.2.2, we consider a common measure of what we here call "corpus-centrality": the INVERSE DOCUMENT FREQUENCY statistic (IDF).

IDF was introduced to the information retrieval problem by Spärck Jones (1972), and is intended to capture the idea that a term used in nearly every document in a corpus is relatively unimportant for distinguishing that document's topic from others in that same corpus. Most often this captures closed-class STOP WORDS like *the* and *but*, but in a specialised corpus it might also capture the common topic of the corpus.

There are several different alternatives for computing IDF and related measures, here we follow Manning et al. (2008, pp 117–119). The DOCUMENT FREQUENCY df_t of a term

t is given in equation (5.7). The INVERSE DOCUMENT FREQUENCY idf_t of term t relative to a corpus of N documents is then as given in equation (5.8):

$$\mathrm{df}_t = \text{Number of documents in the corpus in which term } t \text{ occurs at least once} \tag{5.7}$$

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t} \tag{5.8}$$

Most often in information retrieval, IDF is combined with the TERM FREQUENCY $\mathrm{tf}_{t,d}$ of term t relative to document d as given in equation (5.9), in a statistic called TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY statistic (TF-IDF) as shown in equation (5.10):

$$\mathrm{tf}_{t,d} = \text{Number of occurences of term } t \text{ in document } d \tag{5.9}$$

$$\text{tf-idf}_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t \tag{5.10}$$

TF-IDF is often used as a feature in approaches to various topic-related problems, originally information retrieval—the retrieval of the most appropriate document from a collection for a given search query—and later text classification—the assignment of documents to one of several given categories or topics. Salton and Buckley (1988) showed that, in information retrieval, TF-IDF outperformed many considerably more complex models of the importance of a term to a document's topic. The text classification literature survey by Sebastiani (2002) describes similar findings in text categorisation.

However, TF-IDF functions somewhat differently to the measures of distribution discussed in Section 5.2.2 because it does not return a single statistic representing the skewedness of the distribution of a word among all the document categories. The value of TF-IDF varies by choice of term and document, hence it is a *per document* measure of a word's importance.

While we cannot use TF-IDF in a way comparable to information gain or Kullback–Leibler divergence, IDF alone is a possible way to compare a word's comparitive importance to particular categories, where a lower IDF for a category indicates a higher importance to the category. Consider as an example two categories BAD MOVIES and GOOD MOVIES, in which the former has 10 documents of which 5 use the word *terrible*, and in which the latter has 20 documents of which 2 use the word *terrible*. In this case,

the value of $\text{idf}_{terrible}$ relative to BAD MOVIES is shown in equation (5.11) and the value of $\text{idf}_{terrible}$ relative to GOOD MOVIES is shown in equation (5.12):

$$\begin{aligned} \text{idf}_{terrible} &= \log \frac{10}{5} \\ &\approx 0.69 \end{aligned} \tag{5.11}$$

$$\begin{aligned} \text{idf}_{terrible} &= \log \frac{20}{2} \\ &\approx 2.30 \end{aligned} \tag{5.12}$$

We could thus consider IDF relative to corpora with different degrees of negativity as a potential feature in a system for learning valence.

5.3 Human evaluation of valence shifted sentences

As discussed above in Section 5.1, valence shifting, like many NLG tasks, is a difficult problem to fully evaluate, because it is difficult to determine what the best possible output text is, or, in some cases, *which* valence should be shifted. Hence, we confine ourselves to the narrower question of whether lexical substitution is an effective valence shifting strategy, and whether we can develop a way of predicting which words will shift valence effectively.

We ask human subjects to analyse sentences on two axes: ACCEPTABILITY and NEGATIVITY. This is loosely equivalent to the FLUENCY and FIDELITY axes that are used to evaluate machine translation (Jurafsky and Martin, 2009, Sections 25.3 and 25.9). As in the case of machine translation, a valence-shifted sentence needs to be fluent, that is to be a sentence that is acceptable or better in its grammar, semantics and so on, to listeners or readers. While some notion of fidelity or faithfulness to the original is also important in valence shifting, it is rather difficult to capture without knowing the intent of the valence shifting, since unlike in translation a part of the meaning is being deliberately altered. We therefore confine ourselves in this work to confirming that the valence shifting did in fact take place, by asking subjects to rate sentences.

In order to obtain a clear answer, we specifically evaluate valence shifting with sentences as close to ideal as possible, choosing words we strongly believe to have large valence differences, and manually selecting sentences where the subjects' assessment of the valence of these words is unlikely to be led astray by very poor substitutions such as replacing part of a proper name. (For example, consider the band name *Panic! at the Disco*: asking whether an otherwise identical sentence about a band named *Concern! at the Disco* is

less negative is unlikely to get a good evaluation of lexical valence shifting.) We then ask human subjects to evaluate these pairs of sentences for their relative fluency and negativity.

5.3.1 Using Mechanical Turk for linguistic experiments

Our subjects were recruited through Amazon Mechanical Turk[3]. Mechanical Turk is a web service providing cheap decentralised work units called Human Intelligence Tasks (HITs), which have been used by computational linguistics research for experimentation. Mechanical Turk launched in 2005 and was being evaluated as a tool to generated human reviews of datasets by 2007 (Su et al., 2007). Snow et al. (2008) cite a number of studies at that time which used Mechanical Turk as an annotation tool, including several which used Mechanical Turk rather than expert annotators to produce a gold standard annotation to evaluate their systems.

Snow et al. explore the reliability of using Mechanical Turk for annotation. Restricting themselves to short instructions and to tasks that only require a multiple-choice or numeric response, they compared the performance of workers on Mechanical Turk with expert annotators on several tasks: affective text analysis, word similarity, recognising textual entailment, event annotation and word sense disambiguation. Beginning with 10 annotators per unique question, they found that as few as 4 annotators per question, combined with bias correction techniques, were sufficient to approach expert-level performance in their problems.

While Mechanical Turk has been shown to be an effective annotation resource, Fort et al. (2011) question the ethics of using it: it is likely that at least a substantial number of workers are motivated to participate by money rather than entertainment; workers earn less than USD 2 an hour; and they do not have access to collective bargaining or workplace benefits. (Some studies such as Downs et al. (2010) elect to pay minimum wage.) In addition Fort et al. summarise several cases where it has been found to be less effective than desired by researchers: they report that some researchers find that the interface does not allow complex annotation questions to be asked (Tratz and Hovy, 2010; Gillick and Liu, 2010); that Mechanical Turk does not always approach the performance of trained annotators on all tasks (Bhardwaj et al., 2010); that the performance of Mechanical Turk workers is sometimes outperformed by standard machine learning techniques (Wais et al., 2010); and that having a large number of HITs completed can take a long time (Ipeirotis, 2010).

[3] http://www.mturk.com/

Callison-Burch and Dredze (2010) provide guidelines in appropriate design of tasks for Mechanical Turk:

1. provide clear and concise instructions suitable for non-experts;

2. insert appropriate controls from gold standard data where possible, allowing for removing the work of bad workers or weighting it differently;

3. report on your quality control measures; and

4. where possible provide the raw Mechanical Turk produced data for verification.

Wais et al. (2010) describe "filtering" techniques designed to remove the results of low-accuracy workers, noting that existing techniques such as that of Snow et al. (2008) involved assigning 10 or more workers to each unique question. For their task, which was the recognition of parts of business listings (such as choosing which of two strings was a telephone number, or extracting a telephone number from a full listing). They ended up developing a qualification test and comparing worker performance to expert annotations to select workers for their task. From a pool of over 4000 applicants, only 79 were selected to work on their annotation task. Molla and Santiago-Martinez (2011), who called upon workers to provide PubMed database IDs for references, also used a qualification task and even approved workers were double-checked: of their 10 assigned HITs, 2 had known answers which if answered caused the worker's answers to be rejected. In addition Molla and Santiago-Martinez rejected HITs if the workers provided invalid PubMed IDs, if the database entry with the provided ID did not have substantial overlap with the title shown to the worker, and if the worker disagreed with the majority (from 5 workers) often.

In our experiment, we ameliorate the risks of using Mechanical Turk using the following strategies:

1. We confined ourselves to asking workers for numerical ratings of sentences, rather than any more complex tasks, well within the type of tasks which Snow et al. reported success with.

2. We paid workers well compared to the typical Mechanical Turk HIT, which pays on the order of \$0.02 to \$0.05: accepted workers were paid \$0.96 for an accepted HIT, or \$0.02 for each question asked (including the elimination question described immediately below). We hope in this manner to have paid a more reasonable rate for the workers' time.

3. Similarly to the approach of Molla and Santiago-Martinez, all subjects were given two elimination questions in which the sentences within each pair were identical, that is, in which there was no lexical substitution. These, being identical, should receive identical scores—we also explicitly pointed this out in the instructions—and therefore we could easily eliminate workers who did not read the instructions from the pool.

Since we do not have a gold standard we did not implement the additional verification against gold standard data suggested by Callison-Burch and Dredze. Nor, for this task, can or should we agree that the majority is probably correct about our questions as Molla and Santiago-Martinez did.

5.3.2 Eliciting subjects' responses

In evaluating human responses to the question of whether a given sentence is more negative, we need to have a measurement of whether statistically significant differences in subjects' perceptions of the two sentences appear in the data.

In this section, we describe possible ways of eliciting subjects' perceptions of the two sentences.

5.3.2.1 Binary responses

A simple possible way of presenting the data to subjects would be an experiment which presented stimuli in this style:

> Select which of these two sentences reads like fluent English written by a native speaker, or select 'The two sentences are equally fluent.'
>
> Sentence 1: There are a few energetic scenes that save this production from being an utter DEBACLE.
>
> Sentence 2: There are a few energetic scenes that save this production from being an utter CATASTROPHE.

The subject then chooses among the following options:

- Sentence 1 reads more like fluent English written by a native speaker
- Sentence 2 reads more like fluent English written by a native speaker

- The two sentences are equally fluent.

After this judgement, they then might be presented with the negativity question:

Select which of the two sentences in the previous question is more negative, or select 'The two sentences are equally negative or I cannot choose which sentence is more negative.'

And choose among the following options:

- Sentence 1 is more negative
- Sentence 2 is more negative
- The two sentences are equally negative or I cannot choose which sentence is more negative.

This method is simple, but has several major disadvantages. The first is that subjects tend to choose the "equal"/"cannot tell" option frequently, and that there is no possibility of finding out the strength of any given metric: that is, it is impossible to determine if when the metric predicts a large difference in negativity individual subjects perceive such a difference.

5.3.2.2 Magnitude Estimation

Magnitude Estimation is a technique proposed by Bard et al. (1996) for adapting to grammaticality judgements. In this experimental modality, subjects are asked evaluate stimuli based not on a fixed rating scale, but on an arbitrary rating scale in comparison with an initial stimulus.

For example, subjects might be asked to judge the acceptability of this sentence initially:

(5.14) * The cat by chased the dog.

Assuming that the subject gives this an acceptability score of N, they will be asked to assign a multiplicative score to other sentences, that is, $2N$ to a sentence that is twice as acceptable and $\frac{N}{2}$ to one half as acceptable.

This same experimental modality was used by Lapata (2001) in which subjects evaluated the acceptability of paraphrases of adjectival phrases, for example, considering the acceptability of each of 5.16 and 5.17 as paraphrases of 5.15:

(5.15) a *difficult* customer

(5.16) a customer that is *difficult* to *satisfy*

(5.17) a customer that is *difficult* to *drive*

Magnitude Estimation was originally developed in the context of psychophysics, the study of perceptions of physical phenomena, such as how perception of warmth is related to both area and power output of a heat source. There are a number of concerns with its application to this experiment. In a standard design and analysis of a Magnitude Estimation experiment (Marks, 1974, chapter 2), all the stimuli given to the subjects have known relationships (for example, that the power level for one heat stimulus was half that of another stimulus), and the experimenter is careful to provide subjects with stimuli ranging over the known spectrum of strength under investigation.

In our case, we do not have a single spectrum of stimuli such as a heat source varying in power, or even the varying degrees of fluency given by Bard et al. (1996) or the hypothesised three levels of paraphrase acceptability (low, medium, high) that Lapata (2001) is testing that her subjects can detect. Instead, we have distinct sets of stimuli, each a pair of words, in which we hypothesise a reliable detectable difference within the pair of words, but not between a member of one pair and a member of any other pair. Thus, asking subjects to rate stimuli across the pairs of words on the same scale, as Magnitude Estimation requires, is not the correct experimental design for our task.

5.3.2.3 Rating scale

Given the difficulties with binary responses—either forcing subjects to choose or having a "can't tell" option that the subjects will rely on, and the inappropriateness of Magnitude Estimation for our task, we use an 11 point (0 to 10) rating scale. This allows subjects to rate two sentences as identical if they really perceive the sentences to be so, while allowing fairly subtle differences to be captured.

This is similar to the assessment of machine translation performance used by the National Institute of Standards and Technology. In their case they give very little context or explicit instructions, as Przybocki et al. (2008) observe. Their wording of fluency and adequacy questions are shown in Figures 5.1 and 5.2 and the background descriptions they provide in their guidelines (Linguistic Data Consortium, 2005) are:

How do you judge the fluency of this translation? It is:

5 Flawless English

4 Good English

3 Non-native English

2 Disfluent English

1 Incomprehensible

Figure 5.1: The machine translation fluency question posed in the LDC guidelines (Linguistic Data Consortium, 2005)

How much of the meaning expressed in the gold-standard translation is also expressed in the target translation?

5 All

4 Most

3 Much

2 Little

1 None

Figure 5.2: The machine translation adequacy question posed in the LDC guidelines (Linguistic Data Consortium, 2005)

Fluency Assessment

For each translation of each segment of each selected story, judges make the fluency judgement before the adequacy judgement. Fluency refers to the degree to which the target is well formed according to the rules of Standard Written English. A fluent segment is one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker of English.[...]

Adequacy Assessment

Having made the fluency judgement for a translation of a segment, the judge is presented with the "gold-standard" translation. Comparing the target translation against the gold-standard judges determine whether the translation is adequate. Adequacy refers to the degree to which information present in the original is also communicated in the translation. Thus for adequacy judgements, the gold-standard will serve as a proxy for the original source-language text.

We thus provide similar questions, although with more context in the actual instructions. The precise wording of one of our questions is shown in Figure 5.3 and full introductory instructions to participants are shown in Appendix E.

5.4 Test data

5.4.1 Selection of negative-less negative word pairs

In informal preliminary experiments[4], we initially pre-supposed that our metric could identify word pairs at either end of a negativity spectrum. We identified near synonym sets from *Use the Right Word* which appeared to us to contain negative words, and attempted to use the metrics to choose words from within the *Use the Right Word* near synonyms, with the two most distant scores forming the pair of words presumed to differ noticeably in negativity.

We initially demonstrate the source of our intuition by displaying results of these metrics on sample data: see Tables 5.5 and 5.6 for some sample information gain and

[4]Since these were informal, and involved linguistic experts and non-native speakers, the results are not detailed in this thesis.

Acceptability and negativity: concern/panic
Evaluate these two sentences for acceptability and negativity:

- Sentence 1: As they do throughout the film the acting of CONCERN and fear by Gibson and Russo is genuine and touching.
- Sentence 2: As they do throughout the film the acting of PANIC and fear by Gibson and Russo is genuine and touching.

Acceptability: first sentence of concern/panic pair
Give sentence 1 immediately above a score from 0 to 10 for its acceptability, where higher scores are more acceptable. The primary criterion for acceptability is reading like fluent English written by a native speaker.
Acceptability: second sentence of concern/panic pair
Give sentence 2 immediately above a score from 0 to 10 inclusive for its acceptability, where higher scores are more acceptable.
Negativity: first sentence of concern/panic pair
Give sentence 1 immediately above a score from 0 to 10 inclusive its negativity, where higher scores are more negative.
Negativity: second sentence of concern/panic pair
Give sentence 2 immediately above a score from 0 to 10 inclusive its negativity, where higher scores are more negative.

Figure 5.3: One of the acceptability and negativity questions posed to Mechanical Turk workers.

Near-synonym set	Lowest scoring word		Highest scoring word	
debacle, *disaster*, *catastrophe*	*catastrophe*	3×10^{-4}	*debacle*	1×10^{-2}
harshness, *bitterness*	*bitterness*	5×10^{-4}	*harshness*	5×10^{-4}
lumbering, *gawky*, *inept*, *awkward*, *clumsy*	*gawky*	2×10^{-4}	*inept*	5×10^{-3}

Table 5.5: Sample information gain scores for some negative words drawn from *Use the Right Word*

Kullback–Leibler divergence scores for words using the Scale 1.0 corpus in order to calculate the metrics' values. Results on our test data are shown later in Section 5.6.1.

As we see from Tables 5.5 and 5.6, larger word sets like *lumbering*, *gawky*, *inept*, *awkward*, *clumsy* might have two entirely different supposed extreme words selected by each of the two metrics. In addition, we cannot have enough faith in the metrics *a priori* to use them to select the best test words.

Near-synonym set	Lowest scoring word		Highest scoring word	
debacle, *disaster*, *catastrophe*	*debacle*	0·71	*catastrophe*	1·01
harshness, *bitterness*	*harshness*	0·80	*bitterness*	0·93
lumbering, *gawky*, *inept*, *awkward*, *clumsy*	*lumbering*	0·79	*clumsy*	1·00

Table 5.6: Sample Kullback–Leibler divergence scores for some negative words drawn from *Use the Right Word*

More Negative	Less Negative
dread	*anticipate*
conspiracy	*arrangement*
cowardly	*cautious*
despair	*concern*
worry	*concern*
frightening	*concerning*
war	*conflict*
assassination	*death*
toothless	*ineffective*
ignored	*overlooked*
stubborn	*persistent*
fad	*trend*
threat	*warning*
aggravating	*irritating*
heartbreaking	*upsetting*
tragedy	*incident*
scandal	*event*
panic	*concern*
idiotic	*misguided*
accusation	*claim*

Table 5.7: Negative and more neutral near synonyms chosen for Mechanical Turk workers

We therefore turn to hand-crafted data to test our hypotheses: words chosen so as to be noticeably negative, with a neutral or slightly negative near synonym. We chose 20 such word pairs, shown in Table 5.7. The more negative word of the pair is from the sentiment lists developed by Nielsen (2011)[5], typically rated about 3 for negativity on his scale (where 5 is reserved for obscenities) and the less negative chosen by us.

5.4.2 Selection of sentences containing negative words

We then selected two sentences for each word pair from the Scale 1.0 corpus. Sentences were initially selected by a random number generator: each sentence originally contained the more negative word. Since we are constructing an idealised system here, evaluating the possibility of valence shifting by changing a single word, we manually eliminated sentences where the part of speech didn't match the intended part of speech of the word pair, where the word was part of a proper name (usually a movie title) and where the fluency of the resulting sentence otherwise appeared terribly bad to us. Where a sentence was rejected another sentence was randomly chosen to take its place until each word pair had two accepted sentences for a total of 40 sentences. We then made changes to capitalisation where necessary for clarity (for example, capitalising movie titles, as the

[5] Available from http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

Its real achievement is creating a record of the kind of ridiculous *frightening/concerning* and grimly hysterical behavior everyone already knows goes on college campuses across the country; it's the realest Real World episode you'll ever see.

Cage's frozen wide-eyed expression tells it all about his absolute *panic/concern*.

Willis and Pfeiffer deftly turn their post-separation scenes into awkward shuffles between longing and *despair/concern* where most films treat divorce either as all-acrimony or all-indifference this one captures how much people want it all to work out.

Sitting through In the Company of Men is a cleaner more guilt-free experience but it's not entirely dissimilar, much as we *dread/anticipate* viewing what must happen we cannot tear ourselves away.

That film was smart enough to treat the story as grand melodrama a weepy transcendental *tragedy/incident* rather than an Oscar-season message movie.

Figure 5.4: 5 of the sentences accepted by us for presentation to test subjects

corpus is normalised to lower case.) Five sample accepted sentences are shown in Figure 5.4 and all considered sentences are shown in Appendix D.

5.4.3 Controlling for ordering effects

Since each subject is being presented with multiple sentences (40 in this experiment), rather than coming to the task untrained, it is possible that there are ordering effects between sentences, in which a subject's answers to previous questions influence their answers to following questions. Therefore we used a Latin square design to ensure that the order of presentation was not the same across subjects, but rather varied in a systematic way to eliminate the possibility of multiple subjects seeing questions in the same order. In addition, the square is balanced, so that there is no cyclical ordering effect (ie if one row of a Latin square is A-B-C and the next B-C-A, there is still an undesirable effect where C is tending to follow B). The Latin square is shown in Figure 5.5, each row was the ordering used for one subject.

In addition, although subjects are naturally not informed of the assignment of words to the More Negative and Less Negative categories, presentation of the words in a consistent order (for example, Sentence 1 always containing the More Negative word) may suggest answers to them. The presentation word order to subjects was therefore randomised at the time of generating each subject's questions.

0	1	39	2	38	3	37	4	36	5	35	6	34	7	33	8	32	9	31	10	30	11	29	12	28	13	27	14	26	15	25	16	24	17	23	18	22	19	21	20
1	2	0	3	39	4	38	5	37	6	36	7	35	8	34	9	33	10	32	11	31	12	30	13	29	14	28	15	27	16	26	17	25	18	24	19	23	20	22	21
2	3	1	4	0	5	39	6	38	7	37	8	36	9	35	10	34	11	33	12	32	13	31	14	30	15	29	16	28	17	27	18	26	19	25	20	24	21	23	22
3	4	2	5	1	6	0	7	39	8	38	9	37	10	36	11	35	12	34	13	33	14	32	15	31	16	30	17	29	18	28	19	27	20	26	21	25	22	24	23
4	5	3	6	2	7	1	8	0	9	39	10	38	11	37	12	36	13	35	14	34	15	33	16	32	17	31	18	30	19	29	20	28	21	27	22	26	23	25	24
5	6	4	7	3	8	2	9	1	10	0	11	39	12	38	13	37	14	36	15	35	16	34	17	33	18	32	19	31	20	30	21	29	22	28	23	27	24	26	25
6	7	5	8	4	9	3	10	2	11	1	12	0	13	39	14	38	15	37	16	36	17	35	18	34	19	33	20	32	21	31	22	30	23	29	24	28	25	27	26
7	8	6	9	5	10	4	11	3	12	2	13	1	14	0	15	39	16	38	17	37	18	36	19	35	20	34	21	33	22	32	23	31	24	30	25	29	26	28	27
8	9	7	10	6	11	5	12	4	13	3	14	2	15	1	16	0	17	39	18	38	19	37	20	36	21	35	22	34	23	33	24	32	25	31	26	30	27	29	28
9	10	8	11	7	12	6	13	5	14	4	15	3	16	2	17	1	18	0	19	39	20	38	21	37	22	36	23	35	24	34	25	33	26	32	27	31	28	30	29
10	11	9	12	8	13	7	14	6	15	5	16	4	17	3	18	2	19	1	20	0	21	39	22	38	23	37	24	36	25	35	26	34	27	33	28	32	29	31	30
11	12	10	13	9	14	8	15	7	16	6	17	5	18	4	19	3	20	2	21	1	22	0	23	39	24	38	25	37	26	36	27	35	28	34	29	33	30	32	31
12	13	11	14	10	15	9	16	8	17	7	18	6	19	5	20	4	21	3	22	2	23	1	24	0	25	39	26	38	27	37	28	36	29	35	30	34	31	33	32
13	14	12	15	11	16	10	17	9	18	8	19	7	20	6	21	5	22	4	23	3	24	2	25	1	26	0	27	39	28	38	29	37	30	36	31	35	32	34	33
14	15	13	16	12	17	11	18	10	19	9	20	8	21	7	22	6	23	5	24	4	25	3	26	2	27	1	28	0	29	39	30	38	31	37	32	36	33	35	34
15	16	14	17	13	18	12	19	11	20	10	21	9	22	8	23	7	24	6	25	5	26	4	27	3	28	2	29	1	30	0	31	39	32	38	33	37	34	36	35
16	17	15	18	14	19	13	20	12	21	11	22	10	23	9	24	8	25	7	26	6	27	5	28	4	29	3	30	2	31	1	32	0	33	39	34	38	35	37	36
17	18	16	19	15	20	14	21	13	22	12	23	11	24	10	25	9	26	8	27	7	28	6	29	5	30	4	31	3	32	2	33	1	34	0	35	39	36	38	37
18	19	17	20	16	21	15	22	14	23	13	24	12	25	11	26	10	27	9	28	8	29	7	30	6	31	5	32	4	33	3	34	2	35	1	36	0	37	39	38
19	20	18	21	17	22	16	23	15	24	14	25	13	26	12	27	11	28	10	29	9	30	8	31	7	32	6	33	5	34	4	35	3	36	2	37	1	38	0	39
20	21	19	22	18	23	17	24	16	25	15	26	14	27	13	28	12	29	11	30	10	31	9	32	8	33	7	34	6	35	5	36	4	37	3	38	2	39	1	0
21	22	20	23	19	24	18	25	17	26	16	27	15	28	14	29	13	30	12	31	11	32	10	33	9	34	8	35	7	36	6	37	5	38	4	39	3	0	2	1
22	23	21	24	20	25	19	26	18	27	17	28	16	29	15	30	14	31	13	32	12	33	11	34	10	35	9	36	8	37	7	38	6	39	5	0	4	1	3	2
23	24	22	25	21	26	20	27	19	28	18	29	17	30	16	31	15	32	14	33	13	34	12	35	11	36	10	37	9	38	8	39	7	0	6	1	5	2	4	3
24	25	23	26	22	27	21	28	20	29	19	30	18	31	17	32	16	33	15	34	14	35	13	36	12	37	11	38	10	39	9	0	8	1	7	2	6	3	5	4
25	26	24	27	23	28	22	29	21	30	20	31	19	32	18	33	17	34	16	35	15	36	14	37	13	38	12	39	11	0	10	1	9	2	8	3	7	4	6	5
26	27	25	28	24	29	23	30	22	31	21	32	20	33	19	34	18	35	17	36	16	37	15	38	14	39	13	0	12	1	11	2	10	3	9	4	8	5	7	6
27	28	26	29	25	30	24	31	23	32	22	33	21	34	20	35	19	36	18	37	17	38	16	39	15	0	14	1	13	2	12	3	11	4	10	5	9	6	8	7
28	29	27	30	26	31	25	32	24	33	23	34	22	35	21	36	20	37	19	38	18	39	17	0	16	1	15	2	14	3	13	4	12	5	11	6	10	7	9	8
29	30	28	31	27	32	26	33	25	34	24	35	23	36	22	37	21	38	20	39	19	0	18	1	17	2	16	3	15	4	14	5	13	6	12	7	11	8	10	9
30	31	29	32	28	33	27	34	26	35	25	36	24	37	23	38	22	39	21	0	20	1	19	2	18	3	17	4	16	5	15	6	14	7	13	8	12	9	11	10
31	32	30	33	29	34	28	35	27	36	26	37	25	38	24	39	23	0	22	1	21	2	20	3	19	4	18	5	17	6	16	7	15	8	14	9	13	10	12	11
32	33	31	34	30	35	29	36	28	37	27	38	26	39	25	0	24	1	23	2	22	3	21	4	20	5	19	6	18	7	17	8	16	9	15	10	14	11	13	12
33	34	32	35	31	36	30	37	29	38	28	39	27	0	26	1	25	2	24	3	23	4	22	5	21	6	20	7	19	8	18	9	17	10	16	11	15	12	14	13
34	35	33	36	32	37	31	38	30	39	29	0	28	1	27	2	26	3	25	4	24	5	23	6	22	7	21	8	20	9	19	10	18	11	17	12	16	13	15	14
35	36	34	37	33	38	32	39	31	0	30	1	29	2	28	3	27	4	26	5	25	6	24	7	23	8	22	9	21	10	20	11	19	12	18	13	17	14	16	15
36	37	35	38	34	39	33	0	32	1	31	2	30	3	29	4	28	5	27	6	26	7	25	8	24	9	23	10	22	11	21	12	20	13	19	14	18	15	17	16
37	38	36	39	35	0	34	1	33	2	32	3	31	4	30	5	29	6	28	7	27	8	26	9	25	10	24	11	23	12	22	13	21	14	20	15	19	16	18	17
38	39	37	0	36	1	35	2	34	3	33	4	32	5	31	6	30	7	29	8	28	9	27	10	26	11	25	12	24	13	23	14	22	15	21	16	20	17	19	18
39	0	38	1	37	2	36	3	35	4	34	5	33	6	32	7	31	8	30	9	29	10	28	11	27	12	26	13	25	14	24	15	23	16	22	17	21	18	20	19

Figure 5.5: The balanced Latin square used to control for ordering effects

5.5 Results

5.5.1 Number of participants

A total of 48 workers did the experiment. 8 were excluded from the analysis, for these reasons:

1. 6 workers failed to rate the identical sentence pairs in the elimination questions described in Section 5.3.1 identically, contrary to explicit instructions.
2. 1 worker confined themselves only to the numbers 5 and 10 in their ratings.
3. 1 worker awarded every sentence 10 for both acceptability and negativity.

Each of the 8 Latin square rows were re-submitted to Mechanical Turk for another worker to complete.

In addition, one worker returned a single score of 610 for the negativity of one of the Less Negative sentences: we assume this was a data entry error and the worker intended either 6 or 10 as the value. In our analysis we set this value to 10, since it is the worse (ie most conservative) assumption for our hypothesis that sentences containing Less Negative words will have a lower negativity score than those containing More Negative words.

5.5.2 Analysing scaled responses

In this chapter we consider two hypotheses:

1. that subjects will perceive a difference in *acceptability* between the original sentence and that containing a hypothesised less negative near synonym; and
2. that subjects will perceive a difference in *negativity* between the original sentence and that containing a hypothesised less negative near synonym, specifically, that subjects will perceive the latter sentence as less negative than the former.

We thus require hypothesis testing in order to determine if the means of the scores of the original sentences and those containing hypothesised less negative near synonyms differ significantly. In this situation, we can use a single-factor within-subject analysis of variance (ANOVA), also known as a single-factor repeated-measures ANOVA, which allows us to account for the fact that subjects are not being exposed to a single experimental condition each, but are exposed to all the experimental conditions. In this experiment

s_1		s_2		s_3	
More Neg	Less Neg	More Neg	Less Neg	More Neg	Less Neg
7	2	5	4	6	7
6	5	3	3	7	6
8	2	4	2	5	2

Table 5.8: Hypothetical data for an illustrative analysis using a single-factor within-subject ANOVA

we do not have any between-subjects factors—known differences between the subjects (such as gender, age, and so on)—which we wish to explore. A within-subjects ANOVA accounts for the lesser variance that can be expected by the subject remaining identical over repeated measurements, and thus has more sensitivity than an ANOVA without repeated measures (Keppel and Wickens, 2004, Part V).

Initially it may not appear that the data can be compared within word classes. We are hypothesising that, for example, sentences containing *stubborn* are read as more negative than otherwise identical sentences containing *persistent*, and that sentences containing *heartbreaking* are read as more negative than otherwise identical sentences containing *upsetting*, but we are not hypothesising any negativity relationship between *stubborn* and *upsetting* or between *heartbreaking* and *persistent*, or, for that matter, between *heartbreaking* and *stubborn*. However, since we are pairing the stimuli, it is reasonable to expect, should our hypothesis hold, that the *mean* scores of the more negative sample should be higher than the less negative sample, when the raters are asked about negativity. Therefore, we use an ANOVA, which tests whether the difference between the means is statistically significant.

We here provide a short worked example of a single-factor within-subjects ANOVA, with hand-crafted data, in order to illustrate the process; the analysis of our experimental data follows in Sections 5.5.3 and 5.5.4. For a full discussion of ANOVAs and within-subject designs see Keppel and Wickens (2004) or other experimental design and analysis texts. In this example we have three subjects s_1, s_2 ands_3 and they are presented with three sentences in the categories of More Neg and Less Neg, and asked to rank the negativity of the sentences from 0 to 10. Their responses are as shown in Table 5.8.

We analyse this data using a statistics package (a R program to do so is given in Appendix F) and it produces the summary statistics shown in Table 5.9. The p value is 0.03729, so $p < 0.05$ and we can say that Table 5.8, if it were real data, would show a statistically significant difference in observed negativity between the two categories More Neg and Less Neg.

	Deg. freedom	Sum Sq	Mean Sq	F value	p
Negativity	1	16·071	16·0714	5·2941	0·03729
Residuals	14	42·500	3·0357		

Table 5.9: The summary statistics produced by a within-subjects ANOVA on the data in Table 5.8

5.5.3 Acceptability results

The mean acceptability rating of sentences containing the More Negative words from Table 5.7 was 6.61. The mean acceptability rating of sentences containing the Less Negative words was 6.41. An ANOVA does not find this difference to be statistically significant. ($F(1, 39) = 1.5975, p = 0.2138$).

This is what we would expect: we manually selected sentences whose less negative versions were acceptable to us. Confirming that this appears to have been true for the Mechanical Turk workers also allows us to focus on the negativity results.

5.5.4 Negativity results

The mean negativity rating of sentences containing the More Negative words from Table 5.7 was 6.11. The mean negativity rating of sentences containing the Less Negative words was 4.82. An ANOVA finds this difference to be highly statistically significant. ($F(1, 39) = 29.324, p = 3.365 \times 10^{-6}$).

The means for each individual word are shown in Table 5.10.

In Table 5.10 we see that the effect is not only statistically significant overall, but very consistent: sentences in the Less Negative group *always* have a lower mean rating than their pair in the More Negative group.

5.5.5 Conclusion

In this section we have presented sentences that substitute a less negative word for a more negative word to subjects, and found that they judge the resulting sentences to be approximately equally acceptable, but less negative, confirming that valence shifting is an effect that can be reliably achieved by lexical substitution at least in idealised cases.

One possible concern with this experimental methodology is that the presentation of the sentences to subjects in pairs—review Figure 5.3 for an example—may have caused the subjects to rely too heavily on simply reviewing the highlighted differences between the sentences (the two words in each pair), and not the entire sentence, especially when it

More Negative word	Mean rating	Less Negative word	Mean rating	Difference
ignored	5·7	overlooked	5·0	0·7
cowardly	6·1	cautious	4·9	1·2
toothless	6·1	ineffective	5·1	1·0
stubborn	5·3	persistent	4·3	1·0
frightening	6·2	concerning	5·5	0·7
assassination	6·2	death	6·0	0·2
fad	5·5	trend	3·5	2·0
idiotic	6·3	misguided	5·6	0·7
war	6·5	conflict	5·4	1·1
accusation	6·3	claim	4·5	1·8
heartbreaking	5·8	upsetting	5·7	0·1
conspiracy	5·6	arrangement	4·1	1·5
dread	6·6	anticipate	3·9	2·7
threat	6·6	warning	5·1	1·6
despair	6·2	concern	4·5	1·7
aggravating	6·2	irritating	5·7	0·5
scandal	6·9	event	3·8	3·1
panic	6·5	concern	4·5	2·0
tragedy	5·9	incident	4·6	1·3
worry	5·3	concern	4·5	0·7

Table 5.10: Mean negativity ratings of word pairs from the More Negative and Less Negative pairs

came to negativity. Future work may wish to separate the paired. stimuli to control for this effect.

5.6 Automatically predicting the raters' scores using distribution statistics

Ultimately, the goal of this work is to be able to correctly predict the correct choice of near synonym so as to achieve the correct level of negativity in output. In the preceding section our data suggests that this can be accomplished with lexical substitution. However, this leaves the problem of determining the negativity of words automatically, rather than relying on hand-crafted data.

In this section, we attempt to determine whether the metric scores are useful in re-predicting the raters' scores.

More Negative	Inf. gain	Less Negative	Inf. gain	Difference
heartbreaking	1.43×10^{-3}	upsetting	4.57×10^{-4}	9.71×10^{-4}
despair	3.02×10^{-4}	concern	9.54×10^{-5}	2.07×10^{-4}
panic	6.51×10^{-4}	concern	9.54×10^{-5}	5.56×10^{-4}
aggravating	3.97×10^{-4}	irritating	2.40×10^{-3}	-2.00×10^{-3}
toothless	6.19×10^{-4}	ineffective	2.66×10^{-4}	3.53×10^{-4}
scandal	2.02×10^{-4}	event	4.87×10^{-4}	-2.85×10^{-4}
war	2.52×10^{-3}	conflict	4.91×10^{-4}	2.03×10^{-3}
frightening	3.49×10^{-4}	concerning	3.69×10^{-4}	-2.05×10^{-5}
cowardly	2.93×10^{-4}	cautious	2.35×10^{-4}	5.75×10^{-5}
assassination	1.34×10^{-4}	death	1.13×10^{-3}	-9.94×10^{-4}
idiotic	3.89×10^{-3}	misguided	1.65×10^{-3}	2.25×10^{-3}
fad	1.63×10^{-5}	trend	1.02×10^{-4}	-8.55×10^{-5}
threat	5.19×10^{-4}	warning	5.60×10^{-5}	4.63×10^{-4}
conspiracy	6.06×10^{-4}	arrangement	-3.96×10^{-4}	-2.10×10^{-4}
dread	1.95×10^{-3}	anticipate	3.64×10^{-4}	1.58×10^{-3}
accusation	3.55×10^{-4}	claim	4.99×10^{-6}	3.50×10^{-4}
stubborn	5.23×10^{-4}	persistent	3.34×10^{-4}	1.89×10^{-4}
ignored	1.23×10^{-4}	overlooked	1.60×10^{-3}	-1.48×10^{-3}
worry	4.06×10^{-4}	concern	9.54×10^{-5}	3.10×10^{-4}
tragedy	6.12×10^{-3}	incident	9.79×10^{-5}	6.02×10^{-3}

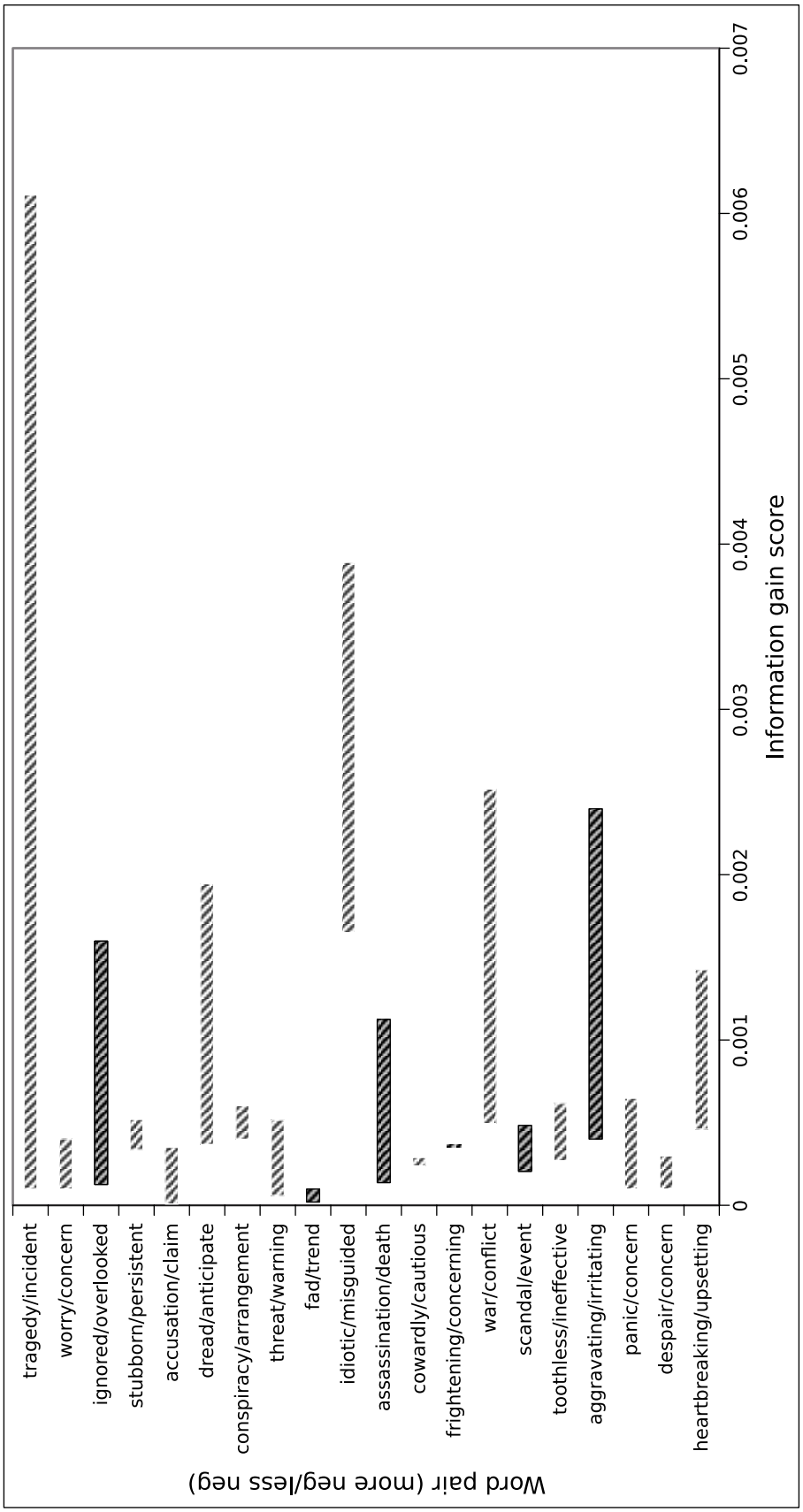
Table 5.11: The information gain values computed for the test data in Table 5.7

5.6.1 Information gain, Kullback–Leibler divergence and IDF values for the test data

The results of the information gain metric given in equation (5.1) on the test data are shown in Table 5.11. The difference between the Less Negative and More Negative information gain is shown in the final column, and also in Figure 5.6. No pattern in the data is immediately obvious, and in particular the ordering of More Negative and Less Negative is not maintained well by the metric.

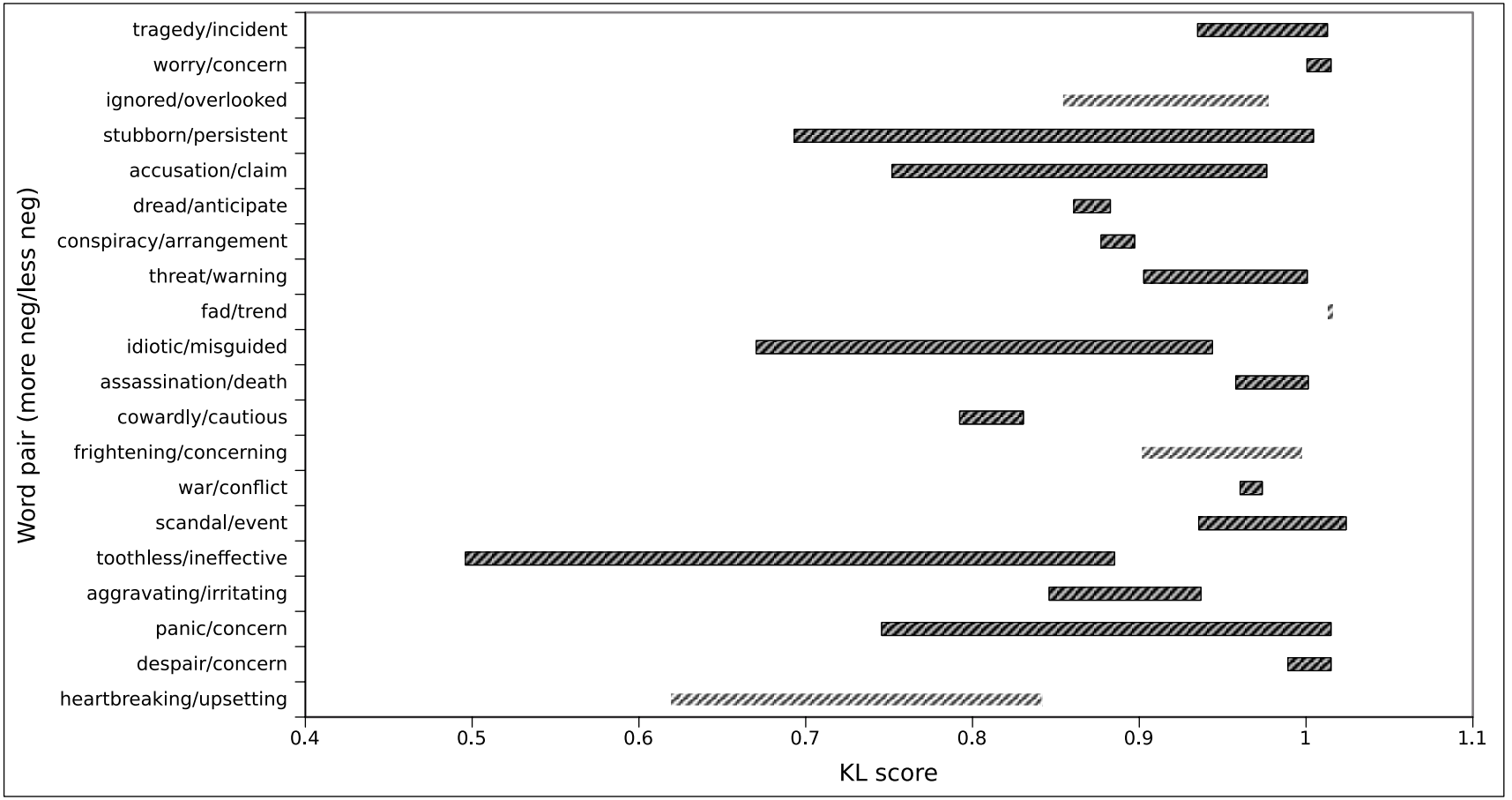
The results of the Kullback–Leibler divergence metric given in equation (5.2) on the test data are shown in Table 5.12. The difference between the Less Negative and More Negative Kullback–Leibler divergence is shown in the final column, and also in Figure 5.7. Here we see a much stronger pattern, that the word from More Negative tends to have a lesser Kullback–Leibler divergence value than the word from Less Negative (18 out of 20 word pairs).

The results of the IDF statistic given in equation (5.8) on the test data are shown in Table 5.13. The difference between the Less Negative and More Negative IDF is shown in the final column, and also in Figure 5.8. We do not see an especially strong pattern of one column's IDF values being larger than another's: eight pairs have a higher



Darker bars indicate that the MORE NEGATIVE word has a smaller information gain score than the LESS NEGATIVE word and its value appears on the left. Lighter bars indicate the reverse with the value for the MORE NEGATIVE word on the right.

Figure 5.6: The information gain values for each MORE NEGATIVE and LESS NEGATIVE pair.



Darker bars indicate that the More Negative word has a smaller Kullback–Leibler divergence value than the Less Negative word and its value appears on the left. Lighter bars indicate the reverse with the value for the More Negative word on the right.

Figure 5.7: The Kullback–Leibler divergence values for each More Negative and Less Negative pair.

More Negative	KL divergence	Less Negative	KL divergence	Difference
heartbreaking	0·84	upsetting	0·62	0·22
despair	0·99	concern	1·02	−0·03
panic	0·75	concern	1·02	−0·27
aggravating	0·85	irritating	0·94	−0·09
toothless	0·50	ineffective	0·89	−0·39
scandal	0·94	event	1·02	−0·09
war	0·96	conflict	0·97	−0·01
frightening	1·00	concerning	0·90	0·10
cowardly	0·79	cautious	0·83	−0·04
assassination	0·96	death	1·00	−0·04
idiotic	0·67	misguided	0·94	−0·27
fad	1·02	trend	1·01	0·00
threat	0·90	warning	1·00	−0·10
conspiracy	0·88	arrangement	0·90	−0·02
dread	0·86	anticipate	0·88	−0·02
accusation	0·75	claim	0·98	−0·23
stubborn	0·69	persistent	1·00	−0·31
ignored	0·98	overlooked	0·85	0·12
worry	1·00	concern	1·02	−0·02
tragedy	0·93	incident	1·01	−0·08

Table 5.12: The Kullback–Leibler divergence values computed for the test data in Table 5.7

value for the More Negative word and 12 for the Less Negative word.

Preliminary indications are thus that the Kullback–Leibler divergence may be a more useful metric for predicting the raters' scores most accurately, and thus perhaps for predicting negativity in usage more generally.

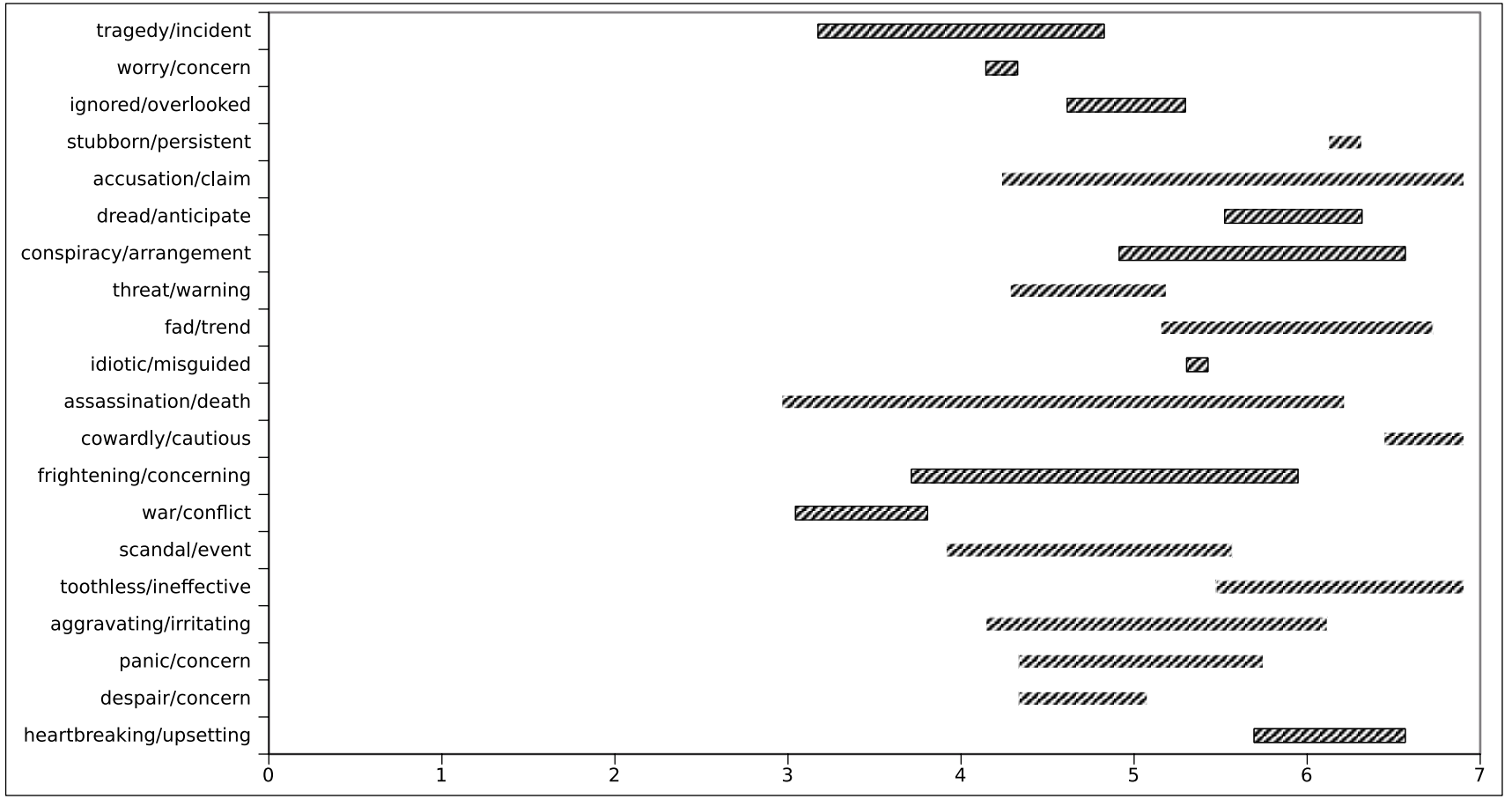
The next step is to transform the metrics into a model that predicts the raters' scores. There are two things that might be done here:

1. finding a function that maps each metric to the raters' scores; and/or
2. combining the two metrics in case there is any useful complementarity.

To achieve both of these simultaneously, we use Support Vector Regression (SVR). See Section 4.3 for a brief discussion of Support Vector Machines as background.

5.6.2 Predicting raters' scores from skewness statistics using Support Vector Machines

We test our hypothesis that the information gain and Kullback–Leibler divergence scores may be useful for predicting the perceived negativity of a word in its context by attempting to predict the raters' scores using Support Vector Machines with the information gain and/or Kullback–Leibler divergence values.



Darker bars indicate that the More Negative word has a smaller IDF score than the Less Negative word and its value appears on the left. Lighter bars indicate the reverse with the value for the More Negative word on the right.

Figure 5.8: The IDF values for each More Negative and Less Negative pair.

More Negative	IDF	Less Negative	IDF	Difference
heartbreaking	5·69	upsetting	6·57	0·89
despair	5·08	concern	4·33	−0·76
panic	5·75	concern	4·33	−1·42
aggravating	6·12	irritating	4·14	−1·98
toothless	6·91	ineffective	5·47	−1·44
scandal	5·57	event	3·91	−1·66
war	3·04	conflict	3·81	0·77
frightening	3·71	concerning	5·95	2·25
cowardly	6·91	cautious	6·44	−0·47
assassination	6·22	death	2·96	−3·26
idiotic	5·30	misguided	5·43	0·13
fad	6·73	trend	5·15	−1·58
threat	5·19	warning	4·28	−0·90
conspiracy	4·91	arrangement	6·57	1·67
dread	5·52	anticipate	6·32	0·80
accusation	6·91	claim	4·23	−2·68
stubborn	6·32	persistent	6·12	−0·20
ignored	4·61	overlooked	5·30	0·69
worry	4·14	concern	4·33	0·19
tragedy	3·17	incident	4·83	1·66

Table 5.13: The IDF values computed for the test data in Table 5.7

In order to do this, for each of the 38 words in Table 5.7 (*concern* appears three times, hence there are not 40 individual test words), we construct a leave-one-out SVM on trained using the values for the other 37 test words. Three separate SVMs are trained for each word:

SVM-IG using the information gain score of the 37 training words as the only feature;

SVM-KL using the Kullback–Leibler divergence score of the 37 training words as the only feature; and

SVM-IG-KL using two features for each training word, the information gain and the Kullback–Leibler divergence value.

SVM-IG and SVM-KL, having only one feature, are equivalent to a regular regression (ie not carried out by an SVM), in that it will transform the metric by rescaling and/or shifting the intercept to match the dependent variable as closely as possible, but having them allows us to evaluate whether the combination of features in SVM-IG-KL adds any information over the better performing of SVM-IG or SVM-KL.

All models are trained using SVM-Light (Joachims, 1999) on the default settings, other than training to perform regression rather than classification. We then ask the SVM to recompute the raters' mean score for that word.

		Predictions			
Category	Word	SVM-IG	SVM-KL	SVM-IG-KL	Mean rater score
More Negative	assassination	4·10	5·55	**5·56**	6·24
Less Negative	death	4·18	**5·51**	**5·51**	6·05
More Negative	panic	4·51	**5·76**	**5·76**	6·51
Less Negative	concern	**4·63**	5·58	5·58	4·53
More Negative	scandal	3·77	**5·58**	**5·58**	6·88
Less Negative	event	**4·63**	5·57	5·57	3·81

Prediction closest to the mean rater score shown in bold.

Table 5.14: Sample SVM predictions for the rater scores for each of the three SVM feature sets SVM-IG, SVM-KL and SVM-IG-KL

5.6.3 Results of prediction from skewness statistics

Sample predictions for individual words are shown in Table 5.14. In these examples we see that the SVM-IG-KL prediction is extremely close to the SVM-KL prediction, indicating that the information gain feature is not contributing to the predicted value when the Kullback–Leibler divergence feature is present. This is confirmed by inspection of the data: while the predicted values of SVM-KL and SVM-IG-KL do differ, the largest such difference is only 7.36×10^{-2} and the mean difference between the SVM-KL and SVM-IG-KL models across the 38 test words is 7.03×10^{-3}.

As in Section 5.6.3, we consider the overall success by examining both the mean average error (MAE) and the mean square error (MSE) of the predictions, with the former weighting all errors equally and the latter penalising outlying errors.

Evaluating the success of a regression without having an immediate application is tricky, because it is not clear how bad errors are: are extreme outliers very bad, or only bad proportionally? This is not yet clear for this task. We therefore give two common error measures in Table 5.15: mean average error and mean square error.

The mean average error (MAE) of a set of predictions $P : p_1, \ldots, p_n$ as compared to a set of known values $C : c_1, \ldots, c_n$ is simply the mean of the difference between each prediction p_i and the corresponding known value, c_i, and weights all errors, small and large, equally:

$$\mathrm{MAE}(P||C) = \frac{\sum_{i=1}^{n} |p_i - c_i|}{n} \tag{5.13}$$

The mean square error (MSE) of predictions P as compared to known values C is the mean of the squares of difference between each prediction p_i and the corresponding known value, c_i, and thus disproportionately penalises outlying errors:

	SVM-IG	SVM-KL	SVM-IG-KL
Mean average error	1·38	0·70	0·70
Mean squared error	2·48	0·75	0·75

Table 5.15: Mean average and mean squared error for each of the three SVM feature sets SVM-IG, SVM-KL and SVM-IG-KL

	SVM-IG	SVM-KL	SVM-IG-KL	Raters
Minimum predicted rating	3·67	5·51	5·51	3·52
Maximum predicted rating	7·37	5·93	5·93	6·88
Standard deviation	0·69	0·10	0·11	0·87

Table 5.16: Range of ratings predicted by SVM-IG, SVM-KL and SVM-IG-KL, compared to those of the raters

$$\mathrm{MSE}(P||C) = \frac{\sum_{i=1}^{n} |p_i - c_i|^2}{n} \tag{5.14}$$

We see in Table 5.15 that, as expected from the examples in Table 5.14, the overall performance of SVM-KL and SVM-IG-KL are nearly identical, and vastly exceed SVM-IG. In addition, the MSE value suggests that SVM-KL and SVM-IG-KL have less outliers than SVM-IG, since it is much closer, both relatively and absolutely, to the MAE value.

Returning to Table 5.14 however, it is clear that SVM-KL and SVM-IG-KL are not themselves a perfect solution to this problem: the range of predicted scores shown is very small. This is confirmed by an inspection of the complete range of predictions made by the models, a summary of which is shown in Table 5.16. The mean rating for a word by the raters themselves varies from 3.52 to 6.88, but the range of SVM-KL and SVM-IG-KL is only from 5.51 to 5.93. In addition, the standard deviation of the ratings is lower for *all* models, including SVM-IG, than the standard deviations of the mean scores assigned by the raters. This shows that the models do not adequately model the full spectrum of scores assigned by the raters.

In addition, we can consider whether or not SVM-IG-KL managed to preserve the *order* of ratings; that is, if say the sentences containing WAR were rated as especially negative, did SVM-IG-KL also return an especially high result relative to its other predictions. A simple proxy for this is whether both numbers tend to be on the same side of their respective means: if the worker rating for sentences containing *war* was above the mean, was the SVM-IG-KL prediction above the mean of SVM-IG-KL predictions? Inspection of the data shows that SVM-IG-KL is inconsistent at achieving this: the predicted value for 19 words had a matching relationship with the mean to the worker rating,

but for 19 words it had the opposed relationship, ie less than the mean where the worker rating was greater, or vice versa.

We therefore conclude that Kullback–Leibler divergence between the distribution of a word and the distribution of words overall in a sentiment-annotated corpus is a promising feature for this task, but certainly cannot be used in isolation.

A second issue to consider is why Kullback–Leibler divergence should perform better than information gain in this regard, when both provide a measure of a word's distributional difference from the underlying distribution. We hypothesise that this may be due precisely to the asymmetry of Kullback–Leibler divergence, as discussed in Section 5.2.2. In our valence-shifting pairs in Table 5.7, the pairs are generally going from the more specific (say, *panic*) to the more general (*concern*). Since Kullback–Leibler divergence in the direction we apply it amounts to, in the analysis of Weeds (2003), a measure of the ability of any word in the corpus to substitute for the word in question, it is likely the case that words in general substitute less poorly for the word *concern* than they do for the word *panic*. Therefore, use of this feature needs to evaluated further in light of whether less strongly attitudinal words are in fact also more general in meaning, or whether this is particular to our examples.

5.6.4 Predicting raters' scores from IDF using Support Vector Machines

In Section 5.6.2 we attempted to re-predict the raters' scores from the information gain and Kullback–Leibler divergence statistics. We found that Kullback–Leibler divergence was a potentially useful feature, and information gain appears not to be. In this section, we explore IDF as a feature, to discover whether it is useful, and potentially complementary to Kullback–Leibler divergence.

As discussed in Section 5.2.3, the IDF of two terms within a corpus could be considered as a measure of that term's importance to the corpus as a whole. While this will capture stop words (such as *the* and *and*) as important, we can use it as a feature for particular words, such as our 38 test words, and its value for stop words will not arise.

Therefore, for each of the 38 words in Table 5.7, we again construct a leave-one-out SVM on trained using the values for the other 37 test words. Two further SVMs are trained for each word:

SVM-IDF using the IDF score of the 37 training words as the only feature;

		Predictions			
Category	Word	SVM-KL	SVM-IDF	SVM-IDF-KL	Mean rater score
More Negative	assassination	5·55	**5·59**	**5·59**	6·24
Less Negative	death	5·51	**5·64**	5·63	6·05
More Negative	panic	**5·76**	5·61	5·62	6·51
Less Negative	concern	**5·58**	5·71	5·70	4·53
More Negative	scandal	5·58	**5·62**	**5·62**	6·88
Less Negative	event	**5·57**	5·72	5·72	3·81

Prediction closest to the mean rater score shown in bold.

Table 5.17: Sample SVM predictions for the rater scores for SVM feature sets SVM-KL, SVM-IDF and SVM-IDF-KL

	SVM-KL	SVM-IDF	SVM-IDF-KL
Mean average error	0·70	0·72	0·72
Mean squared error	0·75	0·81	0·81

Table 5.18: Mean average and mean squared error for the SVM feature sets SVM-KL, SVM-IDF and SVM-IDF-KL

SVM-IDF-KL using two features for each training word, the IDF and the Kullback–Leibler divergence value.

As in Section 5.6.2, models are trained using SVM-Light (Joachims, 1999) on the default settings, other than training to perform regression rather than classification and we then ask the SVM to recompute the raters' mean score for that word.

5.6.5 Results of prediction from IDF

Sample predictions for individual words are shown in Table 5.17. As with the SVM-IG-KL SVM, the SVM in these examples is apparently relying heavily on a single feature, this time IDF rather than Kullback–Leibler divergence. The largest difference between the predictions of SVM-IDF-KL and SVM-IDF is 1.84×10^{-2}, and the mean difference 1.21×10^{-3}.

As in Section 5.6.3, we consider the overall success by examining both the mean average error (MAE) and the mean square error (MSE) of the predictions, with the former weighting all errors equally and the latter penalising outlying errors.

In Table 5.18, we see that SVM-IDF and SVM-IDF-KL have virtually identical performance, but that this performance is *worse* than that of SVM-KL (and therefore SVM-IG-KL). Since IDF measures, in a sense, generality (as in, words that are more widespread in a corpus and hence more general have a lower IDF), this suggests that replacing words with more or less general words may not assist valence shifting.

Category	Sentence
self report	I read that most of the people in the film are loosely based on real stars but that is a distraction that I *ignored/overlooked* while watching the film.
reviewer opinion	The last half-hour especially drags as the interaction between Brackett and Peterson is curtailed in favor of shoot-outs chases and a lot of *idiotic/misguided* exposition.
film description	Director Kevin Hooks's Fled is a fast action movie that appears to be about two convicts on the lam but is actually a highly contrived and ridiculous conspiracy on top of *conspiracy/arrangement* plot.

Table 5.19: Example sentences assigned to each of the SELF REPORT, REVIEWER OPINION and FILM DESCRIPTION categories

5.6.6 Error analysis by example sentence objectivity

As the discussion in Section 5.1.3 demonstrates, sentiment analysis and valence shifting are both made more difficult by the varying ways sentiment may be used, including in objective or subjective contexts. In this section we consider whether the different contexts used affect the results given in Section 5.6.3 for the SVM-KL predictions, which is our best-performing feature.

We assign the sentences given to raters the following categories:

self report directed at the reviewer, that is, the reviewer is expressing a negative trait or opinion of themselves

reviewer opinion directed at the work, that is, the reviewer is expressing a negative sentiment about the film

film description neutrally describing a negative trait of a film character, setting, plot etc without expressing an opinion about it

One of sentences assigned to each of the three categories shown in Table 5.19 and the full assignment of categories to the 40 example sentences is shown in Appendix D. Table 5.20 shows the number of word-pairs (of 20, but one word-pair may have sentences in up to two categories) that are included in each category and the number of sentences (of 40) in each (since each MORE NEGATIVE-LESS NEGATIVE word-pair has two sentences in the test data).

The predictions of SVM-KL is on the level of individual words, and because it uses the mean rating as features, does not distinguish between the two (or six, in the case of *concern* occuring in three word pairs) sentences the raters were given, we here consider

Category	Number of sentences	Number of word pairs
self report	2	2
reviewer opinion	14	8
film description	24	12

Table 5.20: Number of test data sentences and word pairs per category

Category	MAE	MSE
self report	0·91	1·15
reviewer opinion	0·53	0·54
film description	0·83	0·90

Table 5.21: Mean average error and mean square error of SVM-KL at predicting the rating of words in each category

performance of the words by category when at least one of the sentences containing that word fell into that category.

The MAE and MSE per category are shown in Table 5.21. Both average errors for REVIEWER OPINION words are noticeably lower than the other two averages. Iff we compare the per-word rank of errors for the two large categories REVIEWER OPINION as opposed to FILM DESCRIPTION for words that have a sentence in one of the two categories but not both, we find that the distributions of the two samples differ significantly (Mann–Whitney $U = 218$, $n_1 = 13$, $n_2 = 22$, $p < 0.01$ two-tailed).

This means that SVM-KL is better at predicting the ratings of a word's negativity when at least one of that word's example sentences is in the REVIEWER OPINION category, where the reviewer is describing their subjective state. This indicates that Kullback–Leibler divergence–and possibly related measures—may be a more useful feature for valence-shifting in subjective contexts than otherwise.

5.7 Conclusion

In this chapter, we have shown that lexical substitution, as we hoped, can achieve valence shifting on its own, as judged by human raters with a substitution task. In addition, we have shown that at least one measure of the distribution of a word in a corpus, the Kullback–Leibler divergence, is a potentially promising feature for modelling the ability of a lexical substitution to achieve a valence shift, especially for sentiment in a subjective context.

We have not, in this chapter attempted to reproduce human judgements of the valence, or shifted valence, of pieces of text larger than the sentence level, and one weakness of

information gain, Kullback–Leibler divergence and IDF as features is that they are fairly limited to the lexical level. Extensions to n-grams would quickly run into a data sparsity problem. Even at the lexical level we also expect that they may not be useful as measures of the valence-shifting possibilities of closed class words, even though some closed class words like *very* contribute to sentiment (as discussed by for example Yessenalina and Cardie (2011)) and therefore presumably to valence shifting. Future work needs to be able to reproduce the comparitive sentiment of longer valence-shifted that differ in multiple places.

As we observed at the beginning of this chapter, ultimately the techniques used for the Fill In the Blanks (FITB) task may not be suitable for all tasks that require near-synonym choice, but only for those, such as the intelligent thesaurus or in other writing support tools, where surrounding context (largely) exists. Thus, the ability to use statistical information to inform lexical knowledge-base entries suitable for use in a general NLG system is also important, and our work here suggests some possible avenues for doing so. In particular, the investigation of other distributional measures as discussed by eg Weeds (2003) may prove to provide promising measures of the inherent sentiment, or other characteristics of near-synonym differences, when an appropriately annotated corpus is available.

Chapter 6

Conclusions

The goal of this thesis was to contribute to a solution to the problem of LEXICAL CHOICE, specifically by studying how to choose between words that have similar meanings but differ in sentiment, and by choosing between such words when there is a specific goal of VALENCE SHIFTING.

This thesis has shown that words that differ in sentiment or affect are a unique subset of the lexical choice and will need either specialised techniques or at least attention in performance and error analysis of lexical choice solutions. It has shown that valence-shifting is a meaningful problem, and one that can be addressed with lexical substitution as one possible approach, at least as a baseline.

There is considerable scope for further work in lexical choice and valence-shifting. In concluding this thesis, we first provide a summary of our major results in Section 6.1 and then outline future research directions in Section 6.2.

6.1 Summary of findings

6.1.1 Performance of FITB approaches on attitudinal near synonyms

We have examined the FILL IN THE BLANKS (FITB) task in which a system must suggest the most appropriate near synonym from a list in order to supply the missing word in a piece of text, such as in this example due to Edmonds (1997) and originally introduced here in Chapter 1:

(6.1) However, such a move also would run the risk of cutting deeply into U.S. economic

growth, which is why some economists think it would be a big {*error* | *mistake* | *oversight*}.

We have observed that the tendency with the FITB task is to evaluate performance on a fairly small set of near synonyms, which does not provide scope for performance to be analysed in terms of certain types of near synonyms, despite considerable discussion in the literature of the axes on which near synonyms can vary.

Based both on the large number of societal uses for writing attitude-infused text, and on the mature body of statistical approaches to sentiment analysis — primarily the detection of the polarity of texts — we concentrate on near synonyms that express, and sometimes differ, in attitude.

We reimplemented two existing approaches to the FITB problem, and found that they indeed vary in their ability to predict the usage of attitude and non-attitudinal near synonyms.

Limitations of this investigation included having imperfect re-implementations of methods described in the literature; not investigating the performance of more recent approaches to the FITB problem (see Section 2.4.2) on near synonyms that differ in sentiment; and not investigating related hypotheses on tasks like the lexical substitution task of McCarthy and Navigli (2007) (discussed in Section 2.4.4), which would bolster the claim that near synonyms that differ in sentiment behave differently in word choice or word prediction tasks. Overall, it is not clear how generalisable these results are to natural language generation applications, this requires considerable further work.

6.1.2 Development of FITB approaches that allow wide context to be incorporated

We have developed new supervised approaches to the FITB problem and investigated various feature sets. We found that incorporating distant lexical choices in a document can be a helpful feature if appropriate weighting is applied allowing them to be discounted for distance from the near synonym gap. We have investigated directly using author identity, document sentiment and proxies thereof as features, finding them to be only slightly useful at best, at this stage.

We have found that these techniques, like the existing ones we tested, behave differently for attitudinal near synonym sets, and also that these near-synonym sets tend to have a lower most-frequent baseline result also, continuing to suggest that statistical approaches may need to treat these near synonyms differently.

As a minor contribution in addition to this, we developed a set of annotated near synonyms for testing which include a sizeable proportion of attitudinal near synonyms.

It remains to be seen what, precisely, what the weighting of distance words is contributing to the FITB solution. Additional error analysis may contribute to our understanding. This approach to lexical choice may be of limited in natural language generation systems that cannot realise the remainder of the text before making a lexical choice between certain near-synonyms.

6.1.3 Demonstration of valence shifting by lexical substitution

We have investigated the problem of valence shifting, of automatically rewriting a text in order to change its sentiment — in our case, making it less negative — by lexical substitution. Previous work had had some difficulty with achieving valence shifts recognisable by human judges, we therefore constructed an idealised version of valence shifting which eliminated common confounding factors such as grammatical errors, showing that lexical valence shifting can be accomplished by solely by substituting a more negative word with a less negative word without significant loss of fluency. This result confirms that valence shifting is a meaningful problem for future investigation.

This approach is exploratory and its major limitation is therefore that it does not strongly suggest any particularly effective method for valence-shifting.

6.1.4 Investigation into measures of valence-shifting capability

We explored the idea that the valence-shifting capability of a word might correspond to that word's distribution in a sentiment-annotated corpus. We investigated two such measures of distribution, or rather of the extent to which a word's distribution varied from the overall distribution of words in the corpus: information gain and Kullback–Leibler divergence, together with IDF, a measure of a word's importance to a corpus.

We found that Kullback–Leibler divergence, which can be considered to be the substitutability of a lexical item for other lexical items in the document, had promise as a feature which could contribute to a model of how human judges perceive the results of valence shifting.

The major limitation of this approach is that we did not go on to investigate any of our measures as a feature in a valence-shifting system. They remain simply promising features to be investigated.

6.2 Future work

Each of the contributions of this thesis suggests further research avenues, both for improving lexical choice among near synonyms, and for improving valence shifting capabilities of systems that transform text.

6.2.1 Evaluation of other FITB approaches on attitudinal axes

Chapter 3 of this thesis describes a finding that several of the statistical approaches to the FITB task vary in their ability to predict the use of attitudinal near synonyms from the ability to predict non-attitudinal near synonyms. This result suggests that the several other statistical approaches have been proposed (see Section 2.4.2) should likewise be evaluated in terms of whether their success in predicting near-synonym use varies on the attitudinal axis.

6.2.2 Evaluation of FITB approaches on other near-synonym types

In this thesis we have considered one axis of difference between near synonyms: that of sentiment, as suggested by the large body of sentiment analysis research. However, in Section 2.1 we described many axes on which near synonyms can vary, and to date the performance of FITB approaches has not been broken down along these lines. Thus, another direction for future work would be evaluating FITB approaches, or at least the present best-performing ones, on other axes.

Several of these axes have one of the properties that encouraged us to examine sentiment differences, which is the potential for that axis to infuse the whole document. For example, formality is in some ways a property of an entire document, so as with attitudinal aspects of meaning, statistical approaches to near-synonym choice may vary on their performance on the axis of formality.

6.2.3 Further evaluation of the distance-weighted unigram model

The unigram distance model is yet to be compared specifically to the performance of more recent FITB approaches discussed in Section 2.4.2. While we argue in this thesis for a broader range of test data for the FITB problem on the grounds that the possible different performance of approaches on different types of near-synonyms, verifying this approach on the standard test words and Wall Street Journal corpus would allow us to determine which of the measures to use as a starting point for improving FITB performance when

it comes to near synonyms that have affective meaning.

6.2.4 Identification of the features used by the unigram distance model

The feature values in the unigram distance models for the FITB task are weighted only by their distance from the gap to be filled by the chosen near synonym.Despite having not discovered useful features in our investigations of author identity and sentiment scores, we continue to suspect that distance from the gap is not the only useful feature. Investigation of precisely which distant lexical features are having a measurable impact on the result would allow for more specific features to be developed, and perhaps, if patterns are observed in the lexical features which are assisting the model's performance, the use of non-lexical features that capture similar properties. It may be the case that these features are serving as a proxy for the sentiment of the document, the author of the document, the topic of the document or other global features of the document that could be used directly when known to be helpful.

One hypothesis that has been suggested is that of ELEGANT VARIATION, that is, a stylistic preference to not use an identical word repeatedly in a short sequence of text. Wider context may therefore provide clues to the choice of word that were stylistically prohibited in surrounding text. Another is that discourse level structures are being detected, perhaps contrast or alternative points of view. This could be investigated using existing discourse analysis tools or recent sentiment analysis work that focuses on sentiment-related discouse features.

6.2.5 Further investigation of the human perception of valence shifting

One approach to improving valence shifting is to tackle it from the other side: rather than valence shifting text and then having human judges evaluate it, find out what human judges find to unambiguously shift valence, and then attempt to have systems incorporate these strategies. This can be done in an explicit way ("what would you do, to valence shift this text?") but linguistic intuitions may be better captured if techniques are tested on human judges with ideal-condition implementations, without the judges needing to introspect about how they would valence shift. Successful techniques can then be incorporated into valence-shifting systems.

6.2.6 Experimentation with distributional similarity measures for valence shifting

Given the initial promising result using the Kullback–Leibler divergence to predict the rating of sentences by the human subjects investigation, other conceptually similar measures should be investigated. Given this finding, other measures of distributional similarity (or, perhaps, difference in this case) to the general distribution of words in a sentiment annotated corpus may be suitable: the many measures considered by Weeds (2003) are an obvious starting point.

6.2.7 Valence shifting between negative and positive text

In Chapter 5 we confined ourselves to valence shifting to less negative text. A range of questions about valence shifting between negative and positive text remain open: is this a valid task, to what extent is lexical substitution possible or effective, and are any of the same features explored in Chapter 5 useful when the words may have equal strengths but opposite polarities?

Appendix A

124 WordNet synsets annotated for sentiment differences

This appendix contains the annotations of the 124 high frequency synsets from WordNet described in Section 3.2.2.1. Section A.1 gives the instructions to annotators and a sample annotation question, and Section A.2 gives the full annotation results.

A.1 Instructions to annotators

The instructions to annotators are shown in Figure A.1 on page 163 and a sample annotation question in Figure A.2 on page 163.

A.2 Annotations for each of 124 synsets

Synsets that formed the final 58 selected synsets which were used as test data as described in Section 3.3.1.1 are in bold face.

		Synset has affective?	
POS	Words	Annotator 1	Annotator 2
Noun	*care*, *attention*, *aid*, *tending*	Definitely	Unsure
Noun	*deed*, *feat*, *effort*, *exploit*	Probably	Probably
Adjective	***legal*, *sound***	**Definitely not**	**Definitely not**
Noun	*effect*, *essence*, *burden*, *core*, *gist*	Probably not	Definitely
Noun	*care*, *charge*, *tutelage*, *guardianship*	Probably	Probably not
Adjective	***individual*, *separate*, *single***	**Definitely not**	**Definitely not**

Noun	***operation*, *functioning*, *performance***	**Definitely not**	**Definitely not**
Adjective	***former*, *late*, *previous***	**Definitely not**	**Definitely not**
Noun	***ascent*, *acclivity*, *rise*, *raise*, *climb*, *upgrade***	**Definitely not**	**Definitely not**
Noun	*fusion*, *merger*, *unification*	Definitely not	Probably not
Noun	*batch*, *deal*, *flock*, *good deal*, *great deal*, *hatful*, *heap*, *lot*, *mass*, *mess*, *mickle*, *mint*, *muckle*, *peck*, *pile*, *plenty*, *pot*, *quite a little*, *raft*, *sight*, *slew*, *spate*, *stack*, *tidy sum*, *wad*, *whole lot*, *whole slew*	Definitely	Probably
Adjective	***entire*, *full*, *total***	**Definitely not**	**Definitely not**
Noun	*matter*, *affair*, *thing*	Definitely not	Probably not
Noun	*consequence*, *effect*, *outcome*, *result*, *event*, *issue*, *upshot*	Definitely not	Unsure
Noun	*debris*, *dust*, *junk*, *rubble*, *detritus*	Probably	Probably
Noun	***sphere*, *domain*, *area*, *orbit*, *field*, *arena***	**Definitely not**	**Definitely not**
Adjective	*available*, *uncommitted*	Probably	Probably
Adjective	*clear*, *open*	Probably not	Unsure
Noun	***kind*, *sort*, *form*, *variety***	**Definitely not**	**Definitely not**
Noun	***measure*, *step***	**Definitely not**	**Definitely not**
Noun	*topic*, *subject*, *issue*, *matter*	Probably not	Probably not
Adjective	***respective*, *several*, *various***	**Definitely not**	**Definitely not**
Noun	***share*, *portion*, *part*, *percentage***	**Definitely not**	**Definitely not**
Adjective	***arduous*, *backbreaking*, *grueling*, *gruelling*, *hard*, *heavy*, *laborious*, *punishing*, *toilsome***	**Definitely**	**Definitely**
Adjective	***clean*, *clear*, *light*, *unclouded***	**Definitely not**	**Definitely not**
Noun	***path*, *track*, *course***	**Definitely not**	**Definitely not**
Adjective	***capable*, *open*, *subject***	**Definitely not**	**Definitely not**
Adjective	*significant*, *substantial*	Probably not	Definitely not
Noun	*health*, *wellness*	Probably not	Probably
Noun	*variety*, *change*	Definitely not	Probably not
Noun	***composition*, *paper*, *report*, *theme***	**Definitely not**	**Definitely not**
Noun	***commission*, *charge*, *direction***	**Definitely not**	**Definitely not**
Noun	***test*, *trial*, *run***	**Definitely not**	**Definitely not**

Noun	***sketch*, *study***	**Definitely not**	**Definitely not**
Adjective	***broad*, *full***	**Definitely not**	**Definitely not**
Noun	*speculation*, *venture*	Probably	Definitely
Noun	***place*, *property***	**Definitely not**	**Definitely not**
Noun	*undertaking*, *project*, *task*, *labor*	Probably not	Probably not
Noun	*person*, *individual*, *someone*, *somebody*, *mortal*, *human*, *soul*	Definitely not	Unsure
Adjective	*inadequate*, *poor*, *short*	Definitely	Probably
Noun	*bang*, *boot*, *charge*, *rush*, *flush*, *thrill*, *kick*	Probably	Definitely not
Noun	***topographic point*, *place*, *spot***	**Definitely not**	**Definitely not**
Noun	*drop*, *dip*, *fall*, *free fall*	Probably	Definitely not
Noun	***hand*, *deal***	**Definitely not**	**Definitely not**
Noun	*campaign*, *cause*, *crusade*, *drive*, *movement*, *effort*	Definitely	Probably
Noun	***seat*, *place***	**Definitely not**	**Definitely not**
Adjective	*able*, *capable*	Definitely not	Probably
Noun	***survey*, *study***	**Definitely not**	**Definitely not**
Noun	*support*, *keep*, *livelihood*, *living*, *bread and butter*, *sustenance*	Probably	Definitely
Noun	*coupling*, *mating*, *pairing*, *conjugation*, *union*, *sexual union*	Probably	Definitely
Noun	*mind*, *head*, *brain*, *psyche*, *nous*	Probably not	Probably not
Adjective	*chief*, *main*, *primary*, *principal*	Probably not	Definitely not
Noun	*argument*, *statement*	Definitely not	Probably
Noun	*cogitation*, *study*	Probably not	Definitely not
Adjective	***separate*, *single***	**Definitely not**	**Definitely not**
Adjective	***independent*, *main***	**Definitely not**	**Definitely not**
Adjective	*expected*, *likely*, *potential*	Probably	Probably
Noun	***view*, *survey*, *sight***	**Definitely not**	**Definitely not**
Noun	***union*, *conglutination***	**Definitely not**	**Definitely not**
Noun	*rise*, *boost*, *hike*, *cost increase*	Probably	Definitely
Noun	***position*, *post*, *berth*, *office*, *spot*, *billet*, *place*, *situation***	**Definitely not**	**Definitely not**
Noun	*opinion*, *sentiment*, *persuasion*, *view*, *thought*	Probably	Definitely

Noun	*human body*, *physical body*, *material body*, *soma*, *build*, *figure*, *physique*, *anatomy*, *shape*, *bod*, *chassis*, *frame*, *form*, *flesh*	Probably	Probably not
Adjective	***humble*, *low*, *lowly*, *modest*, *small***	**Definitely**	**Definitely**
Adjective	***broken*, *crushed*, *humbled*, *humiliated*, *low***	**Definitely**	**Definitely**
Noun	***construction*, *building***	**Definitely not**	**Definitely not**
Noun	***remainder*, *balance*, *residual*, *residue*, *residuum*, *rest***	**Definitely not**	**Definitely not**
Noun	*marriage*, *matrimony*, *union*, *spousal relationship*, *wedlock*	Probably not	Probably not
Noun	*approval*, *commendation*	Probably not	Definitely not
Noun	***violence*, *force***	**Definitely**	**Definitely**
Adjective	*bad*, *tough*	Unsure	Definitely
Noun	*job*, *task*, *chore*	Probably	Definitely
Noun	*view*, *aspect*, *prospect*, *scene*, *vista*, *panorama*	Definitely not	Probably not
Adjective	*full*, *replete*	Probably not	Probably not
Noun	*kernel*, *substance*, *core*, *center*, *essence*, *gist*, *heart*, *heart and soul*, *inwardness*, *marrow*, *meat*, *nub*, *pith*, *sum*, *nitty-gritty*	Probably	Probably
Adjective	*cardinal*, *central*, *fundamental*, *key*, *primal*	Definitely not	Probably not
Noun	***procedure*, *process***	**Definitely not**	**Definitely not**
Noun	*documentation*, *support*	Probably not	Definitely not
Adjective	*difficult*, *hard*	Probably	Unsure
Noun	*spending*, *disbursement*, *disbursal*, *outlay*	Definitely not	Definitely
Noun	*security*, *protection*	Probably	Definitely not
Noun	*course*, *trend*	Definitely not	Probably not
Noun	***union*, *sum*, *join***	**Definitely not**	**Definitely not**
Noun	***department*, *section***	**Definitely not**	**Definitely not**
Adjective	***avid*, *great*, *eager*, *zealous***	**Definitely**	**Definitely**
Noun	*study*, *work*	Definitely not	Probably not
Noun	***stead*, *position*, *place*, *lieu***	**Definitely not**	**Definitely not**
Adjective	*light*, *scant*, *short*	Probably	Definitely

Noun	***food*, *nutrient***	**Definitely not**	**Definitely not**
Noun	*imperativeness*, *insistence*, *insistency*, *press*, *pressure*	Probably	Probably
Noun	*effort*, *elbow grease*, *exertion*, *travail*, *sweat*	Probably	Probably not
Noun	*scheme*, *strategy*	Probably	Probably
Noun	*deficit*, *shortage*, *shortfall*	Probably not	Definitely not
Noun	***party*, *company***	**Definitely not**	**Definitely not**
Noun	*fairness*, *equity*	Probably	Probably
Noun	***advance*, *rise***	**Definitely not**	**Definitely not**
Adjective	***full*, *total***	**Definitely not**	**Definitely not**
Noun	*charge*, *complaint*	Definitely not	Probably
Adjective	*big*, *enceinte*, *expectant*, *gravid*, *great*, *large*, *heavy*, *with child*	Definitely	Definitely not
Noun	*caper*, *job*	Probably	Definitely
Adjective	***bad*, *insecure*, *risky*, *high-risk*, *speculative***	**Definitely**	**Definitely**
Noun	*return*, *issue*, *proceeds*, *take*, *takings*, *yield*, *payoff*	Definitely not	Probably not
Noun	***leverage*, *purchase***	**Definitely not**	**Definitely not**
Noun	***book*, *volume***	**Definitely not**	**Definitely not**
Noun	*bargain*, *deal*	Probably	Probably
Noun	***provision*, *supply*, *supplying***	**Definitely not**	**Definitely not**
Noun	***charge*, *billing***	**Definitely not**	**Definitely not**
Noun	*head*, *chief*, *top dog*	Probably	Definitely not
Noun	***committee*, *commission***	**Definitely not**	**Definitely not**
Noun	***option*, *alternative*, *choice***	**Definitely not**	**Definitely not**
Noun	***area*, *region***	**Definitely not**	**Definitely not**
Noun	*history*, *account*, *chronicle*, *story*	Definitely not	Probably not
Noun	***financing*, *funding***	**Definitely not**	**Definitely not**
Noun	***battle*, *conflict*, *fight*, *engagement***	**Definitely**	**Definitely**
Noun	***cause*, *reason*, *grounds***	**Definitely not**	**Definitely not**
Adjective	*especial*, *exceptional*, *particular*, *special*	Definitely not	Unsure
Noun	***choice*, *selection*, *option*, *pick***	**Definitely not**	**Definitely not**
Noun	***output*, *yield*, *production***	**Definitely not**	**Definitely not**
Adjective	*long*, *tenacious*	Definitely	Probably
Adjective	*hard*, *heavy*	Probably	Unsure
Noun	*probe*, *investigation*	Probably not	Definitely not

Noun	*basis*, *base*, *foundation*, *fundament*, *groundwork*, *cornerstone*	Definitely not	Probably not
Adjective	*blue*, *depressed*, *dispirited*, *down*, *downcast*, *downhearted*, *down in the mouth*, *low*, *low-spirited*	Definitely	Probably
Noun	***change*, *alteration*, *modification***	**Definitely not**	**Definitely not**

Near synonyms

In this experiment you will be presented with sets of near synonyms. Near synonyms are groups of words which mean nearly the same thing.
For each set of near synonyms, you will be asked whether you think the differences between all the words are mainly in attitude, or mainly in something else.

Examples

Difference in attitude

An example of two words that differ primarily in attitude to what they're describing are attitude are *stingy*, and *frugal*.

Other differences

An example of two words that differ primarily in meaning are *wood* and *forest*.
An example of two works that differ primarily in formality are *potato* and *spud*.

Figure A.1: Instructions to annotators annotating the 124 WordNet synsets.

Instructions

For each set of comma-separated words, choose whether you think they differ from each other mainly in attitude or mainly in some other way.

Words

position, *post*, *berth*, *office*, *spot*, *billet*, *place*, *situation*
(Meaning: a job in an organization; "he occupied a post in the treasury" (noun 556725))

- Definitely attitude difference
- Probably attitude difference
- Unsure
- Probably not attitude difference
- Definitely not attitude difference

Figure A.2: Sample annotation question for the annotators annotating the 124 WordNet synsets.

Appendix B

47 test word sets from *Use the Right Word* annotated for sentiment differences

These are the 47 test word sets drawn from *Use the Right Word* (Hayakawa, 1968) as described in Section 4.1.2.

Each word is also marked with words in the same set that it shares a WordNet 2.0 synset with, as discussed in Section 4.1.2.2. Words in the same set sharing a synset may be marked with $\times$, $\star$, or †. There is no meaning attached to either the same symbol used in different sets, or the choice of symbol within a set.

Set sentiment type	Word	Sentiment	Word	Sentiment	Word	Sentiment
None	*incorporate*	Neutral	*digest*	Neutral	*absorb*	Neutral
Same	*ludicrous*$^{\times\star}$	Negative	*senseless*	Negative	*foolish*	Negative
	preposterous*	Negative	*ridiculous*$^{\times\star\dagger}$	Negative	*farcical*$^{\times}$	Negative
	absurd*	Negative	*silly*†	Negative	*irrational*	Negative
	unreasonable	Negative				
None	*attend*	Neutral	*accompany*	Neutral		
None	*collect*$^{\times}$	Neutral	*gather*$^{\times}$	Neutral		
Differing	*precise**	Neutral	*accurate**	Neutral	*exact**	Neutral
	right$^{\times}$	Positive	*nice*	Neutral	*correct*$^{\times}$	Neutral
	true	Neutral				

Differing	*acknowledge*$^{\times}$	Neutral	*confess*	Negative	*admit*$^{\times}$	Neutral
Differing	*feat*$^{\times}$	Positive	*operation**	Neutral	*act*	Neutral
	exploit$^{\times}$	Positive	*action*	Neutral	*performance**	Neutral
None	*activity*	Neutral	*stir*	Neutral		
Differing	*insight*	Positive	*perception*	Neutral		
None	*fit*	Neutral	*conform*$^{\times}$	Neutral	*adapt*$^{\times}$	Neutral
None	*supplement*	Neutral	*addition*	Neutral		
None	*adequate*$^{\times}$	Neutral	*satisfactory*	Neutral	*enough*$^{\times}$	Neutral
	sufficient	Neutral				
None	*recommendation*	Neutral	*advice*	Neutral		
Same	*aghast*	Negative	*scared*$^{\times}$	Negative	*frightened*$^{\times}$	Negative
	afraid	Negative				
Differing	*drunk*	Negative	*alcoholic*	Neutral		
None	*fable*$^{\times}$	Neutral	*allegory*$^{\times}$	Neutral		
Differing	*aloof*	Negative	*reserved*	Positive	*detached*	Negative
None	*old*	Neutral	*ancient*	Neutral		
Same	*indignation*	Negative	*rage*$^{\times}$	Negative	*wrath**	Negative
	fury$^{\times}$	Negative	*anger**	Negative		
Differing	*creature*$^{\times}$	Neutral	*animal*$^{\times}$	Neutral	*beast*$^{\times}$	Negative
None	*reply*$^{\times\star\dagger}$	Neutral	*response*$^{\times\star}$	Neutral	*answer**†	Neutral
Same	*foreboding*	Negative	*anxiety*	Negative	*angst*	Negative
	worry	Negative	*dread*	Negative		
None	*aspect*$^{\times}$	Neutral	*look*$^{\times}$	Neutral	*appearance*	Neutral
None	*acclaim*	Neutral	*applause*	Neutral		
None	*around*$^{\times\star}$	Neutral	*about*$^{\times\star}$	Neutral	*approximately*$^{\times}$	Neutral
	roughly	Neutral				
Differing	*debate**	Neutral	*discuss*	Positive	*reason*$^{\times}$	Neutral
	argue$^{\times\star}$	Neutral				
None	*result**	Neutral	*issue*$^{\times\star}$	Neutral	*stem*	Neutral
	emerge$^{\times}$	Neutral	*arise*	Neutral		
None	*material*	Neutral	*weapons*	Neutral	*arms*	Neutral
Differing	*synthetic*	Neutral	*ersatz*	Negative	*false*$^{\times}$	Negative
	imitation$^{\times}$	Neutral				
Differing	*stylist*	Positive	*artist*	Neutral	*creator*	Neutral
	virtuoso	Positive	*painter*	Neutral		
Differing	*mannered*	Negative	*artificial*	Negative	*artistic*$^{\times}$	Neutral
	precious	Negative	*arty*	Negative	*stylized*	Positive

	aesthetic$^{\times}$	Positive				
None	*ally*	Neutral	*associate*$^{\times}$	Neutral	*fellow*$^{\times}$	Neutral
	partner	Neutral				
None	*promise*$^{\times}$	Neutral	*guarantee**	Neutral	*assure*$^{\times\star}$	Neutral
None	*charge*	Neutral	*attack*$^{\times}$	Neutral	*storm*	Neutral
	assault$^{\times}$	Neutral				
None	*sympathy*	Neutral	*attraction*	Neutral	*affinity*	Neutral
Differing	*credit*	Positive	*attribute*	Neutral		
Same	*bad*	Negative	*distasteful*	Negative	*objectionable*	Negative
	unpleasant	Negative				
Same	*banal*	Negative	*fatuous*$^{\times}$	Negative	*inane*$^{\times}$	Negative
	inspid*	Negative	*vapid**	Negative		
Same	*bait*$^{\times}$	Negative	*hector*	Negative	*hound*	Negative
	ride$^{\times}$	Negative				
Same	*bigotry*	Negative	*bias*$^{\times}$	Negative	*intolerance*	Negative
	prejudice$^{\times}$	Negative				
Same	*bitterness*	Negative	*harshness*	Negative		
Same	*bleak*$^{\times}$	Negative	*barren*$^{\times}$	Negative	*desolate*$^{\times}$	Negative
	gaunt	Negative				
Same	*brashness*	Negative	*brass*$^{\times}$	Negative	*cheek*$^{\times}$	Negative
	hide	Negative	*nerve*$^{\times}$	Negative		
Same	*abrupt*	Negative	*curt*$^{\times}$	Negative	*gruff*	Negative
	short$^{\times}$	Negative				
Same	*catastrophe*$^{\times}$	Negative	*debacle*	Negative	*disaster*$^{\times}$	Negative
Same	*clumsy*$^{\times\star\dagger}$	Negative	*awkward*$^{\times\dagger}$	Negative	*gawky**	Negative
	inept$^{\times}$	Negative	*lumbering*	Negative		

Appendix C

z scores corresponding with significance levels for Tables 4.8 and 4.11

z-scores resulting from the non-parametric McNemar test corresponding with the significance levels shown in Tables 4.8 and 4.11 are shown in Tables C.1 and C.2 respectively.

Comparison	Overall			No affect			Same affect			Differing affect		
	dev.	test	both	dev.	test	both	dev.	test	both	dev.	test	both
DocPres v. MF baseline	19·7	9·6	20·8	2·6	8·7	9·1	5·3	5·5	7·1	18·8	2·9	17·4
SentPres v. MF baseline	29·4	19·4	34·3	13·4	15·0	20·2	4·2	3·9	5·2	25·9	13·3	28·5
SentPres v. DocPres	14·4	11·3	18·0	12·8	9·5	15·8	0·9	1·5	1·7	9·6	10·9	14·4

Table C.1: z-scores associated with the significance levels shown in Table 4.8

Distance measure and span	Overall dev.	test	both	No affect dev.	test	both	Same affect dev.	test	both	Differing affect dev.	test	both
InvLinear												
Document	29·5	26·8	39·7	21·9	17·4	28·0	1·1	1·8	2·0	20·5	24·3	31·7
3 sentence	23·4	24·9	34·1	17·2	13·8	22·0	1·8	4·7	5·0	16·4	22·0	27·3
1 sentence	21·0	19·9	28·8	16·3	11·6	20·0	1·7	3·5	3·9	13·7	17·8	22·4
InvSquareRoot												
Document	26·0	25·4	36·3	18·7	15·7	24·4	2·2	2·8	3·3	18·4	22·3	28·9
3 sentence	21·1	23·1	31·2	14·1	11·9	18·5	3·3	4·5	5·3	15·5	20·6	25·7
1 sentence	19·0	19·5	27·2	14·2	11·0	17·9	3·1	4·8	5·6	12·5	16·4	20·6

Table C.2: z-scores associated with the significance shown in Table 4.11, all comparisons with a unigram model with no distance measure

Appendix D

Sentences considered for Mechanical Turk experiment

D.1 Accepted sentences

In this section we give all accepted sentences which were randomly chosen from the the Scale dataset v1.0 movie review data set (Scale 1.0) (Pang and Lee, 2005) corpus due to containing a more negative test word from Table 5.7, and which were manually chosen by us for presentation to Mechanical Turk workers, as described in Section 5.4. These sentences include manually corrected punctuation.

We also give the target of their sentiment, as discussed in Section 5.6.6, that is, whether the negative sentiment is:

self report directed at the reviewer, that is, the reviewer is expressing a negative trait or opinion of themselves

reviewer opinion directed at the work, that is, the reviewer is expressing a negative sentiment about the film

film description neutrally describing a negative trait of a film character, setting, plot etc without expressing an opinion about it

Latin square number	Sentence	Sentiment source
0	I read that most of the people in the film are loosely based on real stars but that is a distraction that I *ignored/overlooked* while watching the film.	self report

1	Ruben whose 1987 sleeper The Stepfather was one of the decade's best thrillers has in The Good Son *ignored/overlooked* everything that made that earlier film work: a tight script excellent atmosphere and solid characterization.	reviewer opinion
2	It's hard to argue with the rational message delivered, I just wished the film took on the politics of the scientists sleeping in the same bed with the government who are just as much a detriment to world peace as the xenophobic media and the *cowardly/cautious* bunch of world leaders.	film description
3	What we learn is that Goya was a flawed man, feeling himself to be *cowardly/cautious* and weak-spirited at times, whose great influences to overcome his shortcomings were Vel Zquez Rembrandt and his imagination tied to reason .	film description
4	EdTV is a shallow and jokey rendering of its subject a *toothless/ineffective* satire that fades before the last punchline.	reviewer opinion
5	It's curious how one of the most inventive made-for-cable tv series, HBO's Tales From the Crypt, could turn into one of the most *toothless/ineffective* movie series to haunt multiplexes.	reviewer opinion
6	Alvin's eyesight is so poor that he doesn't have a driver's license, he walks with a cane has a bad hip, has emphysema, and is as *stubborn/persistent* as a mule therefore even though it doesn't make too much sense to travel the way he does no one can talk him out of it.	film description
7	This *stubborn/persistent* driven hard-working man is not entirely pleasant to be around but he's fascinating and thoroughly human.	film description
8	It's not a horror movie although there are some *frightening/concerning* images.	film description
9	Its real achievement is creating a record of the kind of ridiculous *frightening/concerning* and grimly hysterical behavior everyone already knows goes on college campuses across the country it's the realest real world episode you'll ever see.	film description

10	*assassination/death* taking place in the crowded boxing venue is masterfully done and well worth the price of admission as de Palma goes crazy with his cameras offering numerous shots from all different angles of the crowd and the casino.	film description
11	Admittedly most of the humor is warped: how else could you describe a comedy with central themes of incest and an obsession with JFK's *assassination/death* but it would take an exceedingly bland viewer not to find at least a few amusing elements in the film.	film description
12	So what is the fascination, this *fad/trend* is much larger than previous ones like the Teenage Mutant Ninja Turtles and the power rangers.	reviewer opinion
13	The Star Wars *fad/trend* such as it was lasted into the early '80s and the original film received two additional theatrical runs.	reviewer opinion
14	Add to this an *idiotic/misguided* hollywood ending and First Knight turns out to be a mess of the first order.	reviewer opinion
15	The last half-hour especially drags as the interaction between Brackett and Peterson is curtailed in favor of shootouts chases and a lot of *idiotic/misguided* exposition.	reviewer opinion
16	Unlike those two films and numerous others that show the Bosnian *war/conflict* from an insider's viewpoint Welcome to Sarajevo makes its main character a British TV news reporter.	film description
17	The movie is not rated but would be an R for *war/conflict* carnage brief nudity and a little profanity.	film description
18	When the movie does deal with the children at all it is to have one of them scream out an *accusation/claim* and/or cry perhaps to be resolved later by a sensitive talk and a hug, perhaps not.	film description
19	Personally I don't buy the male bashing *accusation/claim* that has been leveled at Waiting to Exhale.	reviewer opinion
20	I felt very touched by the film and its ability to tell such a *heartbreaking/upsetting* story in such a sensitive and intelligent manner without it being an art house film.	reviewer opinion

21	The entire milieu of the Falls City youth has a *heartbreaking/upsetting* genuineness to it particularly noteworthy is how Lana's mother becomes Mom to everyone in their desperation for any family relationship and Peirce does a fine job of using Brandon to explore the effect of a nonconformist on a place where reality is assumed to be static.	reviewer opinion
22	Director Kevin Hooks's Fled is a fast action movie that appears to be about two convicts on the lam but is actually a highly contrived and ridiculous conspiracy on top of *conspiracy/arrangement* plot.	film description
23	Let me not dwell on this implausibility thing and rather than get into the intricate *conspiracy/arrangement* maze, let me give two small examples.	film description
24	Sitting through In the Company of Men is a cleaner more guilt-free experience but it's not entirely dissimilar much as we *dread/anticipate* viewing what must happen we cannot tear ourselves away.	self report
25	The film draws to an end with the 11 discussing the blurring of class distinctions in england and the effect of the series on their lives most claim to *dread/anticipate* it but tolerate it.	film description
26	The script heaps several unnecessary leaps of faith on the audience most surrounding the lack of seriousness american law enforcement officers place on the *threat/warning* of a major assassination.	film description
27	The curious mix of violence and charm with the implicit *threat/warning* of death lurking behind every toothy smile adds luster to the mystery.	film description
28	Willis and Pfeiffer deftly turn their post-separation scenes into awkward shuffles between longing and *despair/concern* where most films treat divorce either as all-acrimony or all-indifference this one captures how much people want it all to work out.	film description
29	Which signifies that there is something despairing about the foreign influence in Vietnam but what that *despair/concern* exactly means is never clearly stated in the context of the story.	film description

30	Agent whose facial tics grow a bit *aggravating/irritating* but is simply hilarious when he goes into full prowl mode in his attempt to think like a cat.	reviewer opinion
31	The incessant koan-like philosophizing becomes *aggravating/irritating* enough but it's even more frustrating that every one of the men from the law school-educated Staros to Kentucky-bred farm boy Witt thinks in the same college-sophomore-on-a-double-bong-hit-my-fingernail-could-be-a-whole-nother-universe terms.	reviewer opinion
32	This ignorance is bliss tale a tale with no given solution seems like it is a documentary but instead it mixes fact and fiction as it adds setup fictional pieces to its factual story and this national story in Iran which was widely reported in their media causing a national *scandal/event* plays here with a dazzling sense of wonderment and complexity.	film description
33	If you stay to see all of the credits you will learn that it was never proven that either had anything to do with the *scandal/event*.	film description
34	Cage's frozen wide-eyed expression tells it all about his absolute *panic/concern*.	film description
35	As they do throughout the film the acting of *panic/concern* and fear by Gibson and Russo is genuine and touching.	film description
36	That film was smart enough to treat the story as grand melodrama a weepy transcendental *tragedy/incident* rather than an Oscar-season message movie.	film description
37	But Liberty Heights isn't some heavy romantic *tragedy/incident* a la Romeo and Juliet.	film description
38	The director Peter Hyams has a spartan approach that sticks to action read killing without much *worry/concern* about plausibility.	film description
39	This is a special movie that some might detest or feel put upon on first viewing it but is for others an exhilarating work an outstanding example of a non-commercial film allowed to take on the personalities of the stars without *worry/concern* about it being a formula story or needing a glib ending.	film description

D.2 Rejected sentences

In this section we give all rejected sentences which were randomly chosen from the SCALE 1.0 corpus due to containing a more negative test word from Table 5.7, but which were manually rejected by us for various disfluencies, as described in Section 5.4. Since they were rejected, sentences have not been manually punctuated or recapitalised.

Sentence	Reason for rejection
he pops these little magnesium pills at the right time and instant courage just like the badge for bravery on the *cowardly/cautious* lion in the wizard of oz .	part of proper name
most films *fad/trend* in our minds like the setting sun .	wrong part-of-speech
after failing to provide the grandeur of the english patient in love and *war/conflict* then can't muster the kind of convincing characterizations which at least made the whole wide world watchable .	part of multi-word expression
our subject today is something i refer to as characterization via *accusation/claim* .	wrong sense
the first shadow *conspiracy/arrangement* was a terrible film that vanished from theaters almost as soon as it opened .	part of proper name
sad to say most wanted is yet another poorly-executed government conspiracy thriller joining the lackluster ranks of shadow conspiracy murder at 1600 and *conspiracy/arrangement* theory .	part of proper name
conspiracy/arrangement lovers will have fun sorting through the layers of cover-up and treachery here .	nonsensical
i've said before that i'm not a big believer in *conspiracy/arrangement* theories but that's not why i think most films about conspiracy theories are so mediocre at best .	part of multi-word expression
for while the material here is inherently better than that of shadow *conspiracy/arrangement* it takes strong performances and a sure steady hand at the helm to lift the production above the level of its uneven script .	part of proper name
perhaps you think this won't be just another plot-driven *conspiracy/arrangement* thriller .	part of multi-word expression

sad to say most wanted is yet another poorly-executed government conspiracy thriller joining the lackluster ranks of shadow *conspiracy/arrangement* murder at 1600 and conspiracy theory .	part of proper name
the storyline for *conspiracy/arrangement* theory is just as moronic as that of air force one which is saying a lot but at least wolfgang petersen's picture moved .	part of multi-word expression
for the record i don't tend to put much stock in *conspiracy/arrangement* theories .	part of multi-word expression
i put aside my critical notebook and tried to clear my mind of all preconceptions the thirty-two writers who worked on the script the mind-numbing marketing blitz the *dread/anticipate* at the appearance of yet another tv retread .	wrong sense
the combination of lyrical odd scenes of naturalism and the melodrama of a crime-fiction story as the film is full of disappearances and disguises sudden deaths and uncanny resurrections hidden trapdoors and secret tunnels bus chases and rooftop escapes- which gave the film its power its sense of *dread/anticipate* perfectly matching the public's mood at the time of world war1 .	wrong part-of-speech
interlaced with these close-ups are slow-motion shots that give the audience a constant feeling of *dread/anticipate* .	wrong part-of-speech
panic/concern plays like an episode for homicide does that has become more invigorated with psychological possibilities .	part of proper name
touching sad and sometimes funny *panic/concern* is the story of alex's rebellion at what has become his destiny .	part of proper name
panic/concern by writer/director henry bromell is a crisp tale rather like a short story .	part of proper name
the film ably depicts the city in a *panic/concern* and shows the victims foaming at the mouth reduced to madness .	wrong part-of-speech
don't *worry/concern* this isn't one of those manipulative disease-movie-of-the-week weepers .	wrong part-of-speech
i'll *worry/concern* tomorrow about whether or not the farrelly brothers portend the end of civilization as we know it today i'm still grateful someone is willing to push the comedic envelope into heretofore unknown zip codes .	wrong part-of-speech

the best that can be said of the movie is that it is so inoffensively bland that families need not *worry/concern* about the ages of their kids .	wrong part-of-speech
if the script had created characters i cared about which it didn't then i would have been more impressed by this scene because i would *worry/concern* that someone would get hurt .	wrong part-of-speech
the best line was smith's don't *worry/concern* i've been in worse scrapes than this but i don't remember any right now which sounds like it was lifted verbatim from one of the plethora of lethal weapon movies .	wrong part-of-speech
if your kids *worry/concern* about being abducted this could push them over the edge .	wrong part-of-speech
what were we going to do also if you are warned all of the time you soon learn to either ignore it or just *worry/concern* a lot .	wrong part-of-speech
don't *worry/concern* there aren't any subtitles and the adaptation is loose very loose .	wrong part-of-speech
but not to *worry/concern* guy has a rock solid alibi .	wrong part-of-speech
there is plenty of action so don't *worry/concern* you don't have to think of anything when you are watching this flick everything moves at break-neck speed .	wrong part-of-speech
don't *worry/concern* if you didn't .	wrong part-of-speech
if every non-disney animated film in production or on the drawing board is as good as anastasia the executives at the magic kingdom have good reason to *worry/concern* .	wrong part-of-speech
if the script had created characters i cared about which it didn't then i would have been more impressed by this scene because i would *worry/concern* that someone would get hurt .	wrong part-of-speech
don't *worry/concern* .	wrong part-of-speech
but that's not all he has to *worry/concern* about .	wrong part-of-speech
i'll *worry/concern* about what that means tomorrow .	wrong part-of-speech
if you did not cover pucker in your biology class don't *worry/concern* spalding quickly turns you into an expert .	wrong part-of-speech

like wings of desire faraway so close shows us the angels' world in black and white and opens with loosely connected scenes of the angels smiling knowingly over human shoulders as they work and *worry/concern* .	wrong part-of-speech
david needn't *worry/concern* since everyone is so relaxed that a massive sleeping sickness appears about to strike at anytime .	wrong part-of-speech
ford looks so out of breath that you begin to *worry/concern* that he may have a real-life heart attack .	wrong part-of-speech
an old timer consoles him by telling him don't *worry/concern* i've done it myself .	wrong part-of-speech
i'll *worry/concern* tomorrow about whether i've consigned myself to film critic hell today i'm still waiting for my cheeks and sides to recover .	wrong part-of-speech

Appendix E

Instructions to Mechanical Turk workers

In this appendix we give the full introductory instructions given to Mechanical Turk workers for the task described in Section 5.4.

Sentiment Polarity Word Paraphrasing

Please do HITs from this group only once. Only your first HIT in this group will be approved.

Information About this Research

This is a research study conducted by faculty members of Macquarie University, Australia, and funded by Macquarie University. We are inviting you to participate in this research project. Research studies are designed to obtain new knowledge. This new information may help people in the future.

In this research we are examining sentences that use similar words that differ in how negative they are, and want to examine which sentence you perceive to be more negative. In addition, we want to examine which sentence you perceive to be more fluent English.

What are the Benefits of this Research?

Research is designed to benefit society by gaining new knowledge. Though you may not receive any direct benefit from participating in this study, you will learn more about the kinds of research conducted by Macquarie University. There will be no personal benefit from participating in the study, other than contributing to science, and the MTurk.com payment.

What are the Risks of this Research?

There are no known risks associated with participating in this research.

What About Confidentiality?

All information collected in the study is confidential to the extent permitted by law. Please note that the data you provide will be grouped with data others provide for reporting and presentation, and that your personal information will not be used in any presentation of these results.

Voluntary Participation

Participation is completely voluntary. You can also withdraw from this study at any time and without penalty. However, you will be paid only if you complete the study. We will review your submission within 14 days of submission. If your submission is approved you will receive 0.96(0.02 per question for 48 questions total).

Warning: please do HITs from this group only once. Only your first HIT in this group will be approved.

Instructions

The purpose of this exercise is to get you to judge the acceptability and negativity of some English sentences. You will see a series of sentences in the Mechanical Turk interface.

Your task is to judge how good or bad each sentence is, and then how negative it is, by assigning a number to it.

Acceptability task

You will be presented with pairs of sentences that differ by one word. Some will seem perfectly okay to you, but others will not. What we're after is not what you think of the specific meaning of the sentence, but what you think of the way it's constructed.

You may decide that some sentences are perfectly acceptable, but that others are not, due to not being understandable, or being understandable but not sounding natural.

Your task is to judge how good or bad each sentence is by assigning a number from 0 to 10 to it, where the number assigned to each sentence reflects its acceptability relative to the other sentence.

Higher numbers mean more acceptable.

For example, in this pair of sentences you might judge the first sentence as acceptable, and the second sentence as understandable but not natural:

- the movie was *cool*.
- the movie was *snap-frozen*.

Thus, if you assigned the first sentence a 10 for acceptability, you might assign the second sentence 5, as half as acceptable.

Negativity task

You will be presented with pairs of sentences that differ by one word. Some will seem more negative to you than others. What we're after is what you think of the meaning of the sentence, independent of how it's constructed unless its construction is so bad it is meaningless to you.

Your task is to judge how negative (critical, depressed, angry, and expressing other negative emotions and opinions) each sentence is by assigning it a number from 0 to 10, where the number assigned to each sentence reflects its negativity relative to the other.

Higher numbers mean more negative.

For example,consider the sentence pair:

- the movie was *bad*.
- the movie was *abysmal*.

If you assigned the first sentence a 5 for negativity, you might give the second sentence a higher score, such as 7 or 9.

If you cannot meaningfully compare the two sentences because one is meaningless to you, as in this example, assign that sentence the special MEANINGLESS score:

- the movie was *snowing*.
- the movie was *cool*.

Scoring instructions

There are no 'correct' answers, so whatever seems right to you is a valid response.

If you are presented with the same word in both sentences, give them both the same score for acceptability and negativity. For example:

- the movie was *cool*. the movie was *cool*.

Since the sentences are exactly the same, you give them the same score for acceptability and negativity.

Warning: please do HITs from this group only once. Only your first HIT in this group will be approved.

Appendix F

Illustrative ANOVA implementation

This appendix gives the R implementation of the illustrative ANOVA performed in Section 5.5.2.

The data presentation R requires for the data in Table 5.8 is a Comma Separated Values (CSV) file as follows (tab-separated for readability:

Subject	Negativity	NegScore
1	moreneg	7
1	moreneg	6
1	moreneg	8
1	lessneg	2
1	lessneg	5
1	lessneg	2
2	moreneg	5
2	moreneg	4
2	moreneg	3
2	lessneg	4
2	lessneg	3
2	lessneg	2
3	moreneg	6
3	moreneg	7
3	moreneg	5
3	lessneg	7
3	lessneg	6
3	lessneg	2

The R code is:

```
sentenceData <- read.csv("example.csv")
aov.example <- aov(NegScore~Negativity+Error(Subject/Negativity),sentenceData)
summary(aov.example)
```

Appendix G

Ethics approval for Mechanical Turk research

The research presented in Section 5.3 of this thesis, using human subjects, was approved by the Macquarie University Ethics Review Committee, reference number 5201100823 (D), on November 7 2011. A copy of the final approval follows.

From:	**Faculty of Science Research Office <sci.ethics@mq.edu.au>**
Date:	Mon, Nov 7, 2011 at 4:00 PM
Subject:	**Final Approval for Ethics Application 5201100823_Dras(Gardiner)**
To:	A/Prof Mark Dras <mark.dras@mq.edu.au>
Cc:	Prof Richie Howitt <richie.howitt@mq.edu.au>,
	Ms Cathi Humphrey-Hood <cathi.humphrey-hood@mq.edu.au>,
	Faculty of Science Research Office <sci.ethics@mq.edu.au>

Dear Associate Professor Dras,

RE: Ethics project entitled: "Sentiment Polarity Word Paraphrasing"
Ref number: 5201100823 (D)

The Faculty of Science Human Research Ethics Sub-Committee has reviewed your application and granted final approval, effective 7th November 2011. You may now commence your research.

The following personnel are authorised to conduct this research:

Associate Professor Mark Dras
Ms Mary Gardiner

NB. STUDENTS: IT IS YOUR RESPONSIBILITY TO KEEP A COPY OF THIS APPROVAL EMAIL TO SUBMIT WITH YOUR THESIS.

In addition to the standard requirements of approval (listed below) - please note that your use of the Amazon Mechanical Turk for recruitment was considered carefully in the light of emerging discussion of wider ethical concerns with the operations of the AMT (see e.g. Fort K, Adda G and Cohen KB. (2011) Amazon Mechanical Turk: Gold Mine or Coal Mine? Computational Linguistics 37: 413-420). Following discussion with the University's ethics secretariat, it was concluded that your proposed use of the AMT in this study does not raise any specific ethical concerns, but we would appreciate your feedback, either to the Science Ethics Sub-Committee or the Ethics Secretariat, on the issues raised by Fort et al in relation to the sort of work you undertake.

Please note the following standard requirements of approval:
1. The approval of this project is conditional upon your continuing compliance with the National Statement on Ethical Conduct in Human Research (2007).

2. Approval will be for a period of five (5) years subject to the provision of annual reports. Your first progress report is due on 7 November 2012.

If you complete the work earlier than you had planned you must submit a Final Report as soon as the work is completed. If the project has been discontinued or not commenced for any reason, you are also required to submit a Final Report for the project.

Progress reports and Final Reports are available at the following website:

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/forms

3. You may not renew approval for a project lasting more than five (5) years. You will need to complete and submit a Final Report and submit a new application for the project. (The five year limit on renewal of approvals allows the University Human Research Ethics Committee to fully re-review research in an environment where legislation, guidelines and requirements are continually changing, for example, new child protection and privacy laws).

4. All amendments to the project must be reviewed and approved by the Faculty Sub-Committee before implementation. Please complete and submit a Request for Amendment Form available at the following website:

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/forms

5. Please notify the Faculty Sub-Committee immediately in the event of any adverse effects on participants or of any unforeseen events that affect the continued ethical acceptability of the project.

6. At all times you are responsible for the ethical conduct of your research in accordance with the guidelines established by the University. This information is available at the following websites:

http://www.mq.edu.au/policy/

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/policy

If you will be applying for or have applied for internal or external funding for the above project, you are responsible for providing the Macquarie University's Research Grants Management Assistant with a copy of this email as soon as possible. Internal and External funding agencies will not be informed that you have final approval for your project and funds will not be released until the Research Grants Management Assistant has received a copy of this email.

If you need to provide a hard copy letter of Final Approval to an external

organisation as evidence that you have Final Approval, please do not hesitate to contact the Faculty of Science Research team at sci.ethics@mq.edu.au.

Please retain a copy of this email as this is your official notification of final ethics approval.

Yours sincerely,
Richie Howitt, Chair
Faculty of Science Human Research Ethics Sub-Committee
Macquarie University
NSW 2109

Bibliography

Ahn, Kisuh; Johan Bos; James R. Curran; Dave Kor; Malvina Nissim; and Bonnie Webber (2005). Question answering with QED at TREC-2005. In *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings*. National Institute of Standards and Technology. URL http://trec.nist.gov/pubs/trec14/papers/uedinburgh-nissim.qa.pdf. 17

Anderson, Anne A.; Miles Bader; Ellen Gurman Bard; Elizabeth Boyle; Gwyneth Doherty; Simon Garrod; Stephen Isard; Jacqueline Kowtko; Jan McAllister; Jim Miller; Catherine Sotillo; Henry Thompson; and Regina Weinert (1991). The HCRC map task corpus. *Language and Speech*, 34:351–366. doi:10.1177/002383099103400404. 27

Andreevskaia, Alina and Sabine Bergler (2006). Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pages 209–216. Association for Computational Linguistics, Trento, Italy. URL http://www.aclweb.org/anthology/E06-1027. 43, 46

Androutsopoulos, Ion and Prodromos Malakasiotis (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187. doi:10.1613/jair.2985. 20

Aristotle (350 BCE). *Prior Analytics*. URL http://classics.mit.edu/Aristotle/prior.html. 20

Baccianella, Stefano; Andrea Esuli; and Fabrizio Sebastiani (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the ninth international conference on Language Resources and Evaluation (LREC 2010)*, pages 2200–2204. URL http://www.lrec-conf.org/proceedings/lrec2010/summaries/769.html. 49

Baker, Collin F.; Charles J. Fillmore; and John B. Lowe (1998). The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING 1998)*, volume 1, pages 86–90. URL http://www.aclweb.org/anthology/C98-1013. 18, 21

Banfield, Ann (1982). *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Routledge and Kegan Paul. 39

Bard, Ellen Gurman; Dan Robertson; and Antonella Sorace (1996). Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68. URL http://www.jstor.org/stable/416793. 126, 127

Baroni, Marco; Silvia Bernardini; Adriano Ferraresi; and Eros Zanchetta (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226. doi:10.1007/s10579-009-9081-4. 72

Barzilay, Regina and Kathleen R. McKeown (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57. Association for Computational Linguistics, Toulouse, France. doi:10.3115/1073012.1073020. URL http://www.aclweb.org/anthology/P01-1008. 19

Becker, Israela and Vered Aharonson (2010). Last but definitely not least: On the role of the last sentence in automatic polarity-classification. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 331–335. Association for Computational Linguistics, Uppsala, Sweden. URL http://www.aclweb.org/anthology/P10-2061. 41

Belz, Anja (2007). Putting development and evaluation of core technology first. In Dale and White (2007), pages 1–2. URL http://www.ling.ohio-state.edu/nlgeval07/papers/NLGEval07-Position-Papers.pdf. 26

Belz, Anja; Eric Kow; Jette Viethen; and Albert Gatt (2008). The GREC challenge 2008: Overview and evaluation results. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 183–193. URL http://www.aclweb.org/anthology/W08-1127. 26

Belz, Anja; Mike White; Dominic Espinosa; Eric Kow; Deirdre Hogan; and Amanda Stent (2011). The first surface realisation shared task: Overview and evaluation results. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 217–226. Association for Computational Linguistics, Nancy, France. URL http://www.aclweb.org/anthology/W11-2832. 26

Bentivogli, Luisa; Peter Clark; Ido Dagan; Hoa Dang; and Danilo Giampiccolo (2011). The seventh PASCAL Recognizing Textual Entailment challenge. In *Proceedings of the Fourth Text Analysis Conference (TAC 2011)*. National Institute of Standards and Technology, Gaithersburg, Maryland, USA. 20

Bergsma, Shane; Dekang Lin; and Randy Goebel (2009). Web-scale n-gram models for lexical disambiguation. In *Proceedings of the Twenty-First International Joint Conference on Artificial*

Intelligence (IJCAI-09). Pasadena, California. URL http://www.aaai.org/ocs/index.php/IJCAI/IJCAI-09/paper/view/398. 72

Bernard, John R. L., editor (1986). *The Macquarie Thesaurus*. Macquarie Library, Sydney, Australia. 116

Bhardwaj, Vikas; Rebecca Passonneau; Ansaf Salleb-Aouissi; and Nancy Ide (2010). Anveshan: A framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 47–55. Association for Computational Linguistics, Uppsala, Sweden. URL http://www.aclweb.org/anthology/W10-1806. 123

Bickmore, Timothy W.; Laura M. Pfeifer; and Brian W. Jack (2009). Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI '09)*, pages 1265–1274. doi:10.1145/1518701.1518891. 1

Binsted, Kim; Helen Pain; and Graeme Ritchie (1997). Children's evaluation of computer-generated punning riddles. *Pragmatics and Cognition*, 5:309–358. doi:10.1075/pc.5.2.06bin. 1, 26

Blackburn, Patrick and Johan Bos (2005). *Representation and Inference for Natural Language: A First Course in Computational Semantics*. Centre for the Study of Language and Information, Stanford, California, USA. 17

Bos, Johan (2011). A survey of computational semantics: Representation, inference and knowledge in wide-coverage text understanding. *Language and Linguistics Compass*, (5/6):336–366. doi:10.1111/j.1749-818x.2011.00284.x. 17, 18

Brants, Thorsten and Alex Franz (2006). Web 1T 5-gram Version 1. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13. 32, 33, 60, 88

Brennan, Michael and Rachel Greenstadt (2009). Practical attacks against authorship recognition techniques. In *Proceedings of the Twenty-First Innovative Applications of Artificial Intelligence Conference (2009)*, pages 60–65. Pasadena, California, USA. 54

Budanitsky, Alexander and Graeme Hirst (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47. doi:10.1162/coli.2006.32.1.13. 18

Callison-Burch, Chris (2007). *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh. 19, 116

Callison-Burch, Chris and Mark Dredze (2010). Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and*

Language Data with Amazon's Mechanical Turk, pages 1–12. Association for Computational Linguistics, Los Angeles. URL http://www.aclweb.org/anthology/W10-0701. 123, 125

Cerini, Sabrina; Valentina Compagnoni; Alice Demontis; Maicol Formentelli; and Caterina Gandi (2007). Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In *Language resources and linguistic theory: Typology, second language acquisition, English linguistics*. Franco Angeli, Milan, Italy. 49, 99

Charniak, Eugene (1979). With spoon in hand, this must be the eating frame. *American Journal of Computational Linguistics: Inference and Theory*, pages 187–193. URL http://www.aclweb.org/anthology/J79-1080. 20

Charniak, Eugene; Don Blaheta; Niyu Ge; Keith Hall; John Hale; and Mark Johnson (2000). BLLIP 1987-89 WSJ Corpus Release 1. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000T43. 70

Chen, Stanley F. and Joshua Goodman (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL 1996)*, pages 301–318. URL http://www.aclweb/anthology/P/P96/P96-1041.pdf. 33

Chklovski, Timothy and Patrick Pantel (2004). VerbOcean: Mining the web for fine-grained semantic verb relations. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 33–40. Association for Computational Linguistics, Barcelona, Spain. URL http://www.aclweb.org/anthology/W04-3205. 24

Church, Kenneth; William Gale; Patrick Hanks; and Donald Hindle (1991). Using statistics in lexical analysis. In Uri Zernick, editor, *Lexical Acquisition: Using On-line Resources to Build a Lexicon*. Lawrence Erlbaum Associates. 29, 30

Church, Kenneth and Patrick Hanks (1990). Word association norms and mutual information, lexicography. *Computational Linguistics*, 16(1):22–29. URL http://www.aclweb.org/anthology/J90-1003. 30

Church, Kenneth Ward and Patrick Hanks (1989). Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83. Association for Computational Linguistics, Vancouver, British Columbia, Canada. doi:10.3115/981623.981633. URL http://www.aclweb.org/anthology/P89-1010. 31

Clark, Eve V. (1992). Conventionality and contrast: pragmatic principles with lexical consequences. In Adrienne Lehrer and Eva Feder Kittay, editors, *Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization*, pages 171–188. Routledge, New York. 86

Clarke, Charles L. A. and Egidio L. Terra (2003). Passage retrieval vs. document retrieval for factoid question answering. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 427–428. Toronto, Canada. doi:10.1145/860435.860534. 32

Cohen, Jacob (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46. doi:10.1177/001316446002000104. 68

Cover, Thomas M. and Joy A. Thomas (1991). *Elements of Information Theory*. Wiley, New York, USA. 118, 120

Cristianini, Nello and John Shawe-Taylor (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK. 94

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press. 10, 13, 14, 86

Dagan, Ido; Oren Glickman; and Bernando Magnini (2006). The PASCAL Recognising Textual Entailment challenge. In Joaquin Quiñonero-Candela; Ido Dagan; Bernardo Magnini; and Florence d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11–13, 2005, Revised Selected Papers*, pages 177–190. Springer-Verlag. doi:10.1007/11736790_9. xiii, 20

Dale, Robert and Adam Kilgarriff (2011). Helping Our Own: The HOO 2011 pilot shared task. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 242–249. Association for Computational Linguistics, Nancy, France. URL http://www.aclweb.org/anthology/W11-2838. 26

Dale, Robert and Michael White, editors (2007). *Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation: Position Papers*. Arlington, Virginia, USA. URL http://www.ling.ohio-state.edu/nlgeval07/papers/NLGEval07-Position-Papers.pdf. 26, 194, 210, 212

Dasgupta, Sajib and Vincent Ng (2009). Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 701–709. Association for Computational Linguistics, Suntec, Singapore. URL http://www.aclweb.org/anthology/P09-1079. 41

Daumé, Hal, III (2010). Google 5gram corpus has unreasonable 5grams. Weblog entry. URL http://nlpers.blogspot.com/2010/02/google-5gram-corpus-has-unreasonable.html. 72

Di Eugenio, Barbara and Michael Glass (2004). The kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101. doi:10.1162/089120104773633402. 68

Dice, Lee R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26:297–302. URL http://www.jstor.org/stable/1932409. 27, 30

DiMarco, Chrysanne; Graeme Hirst; and Manfred Stede (1993). The semantic and stylistic differentiation of synonyms and near-synonyms. In *Proceedings of AAAI Spring Symposium on Building Lexicons for Machine Translation*, pages 114–121. Stanford, CA, USA. URL http://www.aaai.org/Papers/Symposia/Spring/1993/SS-93-02/SS93-02-025.pdf. 14, 15, 16, 57

Downs, Julie S.; Mandy B. Holbrook; Steve Sheng; and Lorrie Faith Cranor (2010). Are your participants gaming the system? screening Mechanical Turk workers. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI 2010)*, pages 2399–2402. Atlanta, GA, USA. 123

Dras, Mark (1999). *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Ph.D. thesis, Macquarie University. 18, 19

Dras, Mark; Debbie Richards; Meredith Taylor; and Mary Gardiner (2010a). Deceptive agents and language. In *Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2010)*. Toronto, Canada. 24

Dras, Mark; Debbie Richards; Meredith Taylor; and Mary Gardiner (2010b). Deceptive agents and language. In *Proceedings of the International Workshop on Interacting with ECAs as Virtual Characters*. Toronto, Canada. 24

Edmonds, Philip (1997). Choosing the word most typical in context using a lexical co-occurrence network. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 507–509. Association for Computational Linguistics, Madrid, Spain. doi:10.3115/976909.979684. URL http://www.aclweb.org/anthology/P97-1067. xiii, 3, 4, 7, 16, 27, 28, 31, 34, 35, 36, 58, 59, 60, 64, 65, 70, 72, 79, 87, 88, 89, 94, 102, 151

Edmonds, Philip (1999). *Semantic representations of near-synonyms for automatic lexical choice*. Ph.D. thesis, University of Toronto. xiii, 4, 13, 14, 15, 16, 18, 27, 28, 30, 34, 35, 59, 60, 65, 87, 94

Edmonds, Philip and Graeme Hirst (2002). Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144. doi:10.1162/089120102760173625. xvii, 13, 14, 15, 16, 17, 18, 24, 57

Edmonds, Philip and Adam Kilgarriff (2002). Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering*, 8(4):279–291. doi:10.1017/S1351324902002966. 65, 66

Esuli, Andrea and Fabrizio Sebastiani (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 417–422. Genova, Italy. 47, 53, 82

Evans, Vyvyan and Melanie Green (2006). *Cognitive Linguistics: An Introduction*. Lawrence Erlbaum Associates and Edinburgh University Press Ltd., Mahwah, New Jersey, USA. 12

Fellbaum, Christiane, editor (1998). *WordNet: An Electronic Lexical Database*. The MIT Press. 12, 18, 44, 65, 66, 82, 202

Fleischman, Michael and Eduard Hovy (2002). Towards emotional variation in speech-based natural language generation. In *Proceedings of the International Natural Language Generation Conference*. Arden House, New York. 23

Foody, Giles M. (2008). Sample Size Determination for Image Classification Accuracy Assessment and Comparison. In *Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, pages 154–162. Shanghai. 95

Fort, Karën; Gilles Adda; and K. Bretonnel Cohen (2011). Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420. doi:10.1162/COLI_a_00057. 123

Frege, Gottlob (1892). Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50. 10, 11, 14, 58

Fujita, Atsushi; Kentaro Furihata; Kentaro Inui; Yuji Matsumoto; ; and Koichi Takeuchi (2004). Paraphrasing of Japanese light-verb constructions based on lexical conceptual structure. In *Proceedings of the ACL Workshop on Multiword Expressions: Integrating Processing*, pages 9–16. Barcelona, Spain. URL http://www.aclweb.org/anthology/W04-0402. 19

Furbach, Ulrich; Ingo Glöckner; and Björn Pelzer (2010). An application of automated reasoning in natural language question answering. *AI Communications*, 23(2–3):241–265. doi:10.3233/AIC-2010-0461. 17

Gale, William A. and Geoffrey Sampson (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2:217–232. doi:10.1080/09296179508590051. 89, 91

Gallo, Carlos Gómez; T. Florian Jaeger; and Ron Smyth (2008). Incremental Syntactic Planning across Clauses. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci08)*, pages 845–850. 102

Gamon, Michael and Anthony Aue (2005). Automatic identification of sentiment vocabulary: Exploiting low association with known sentiment terms. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 57–64. Association for Computational Linguistics, Ann Arbor, Michigan. URL http://www.aclweb.org/anthology/W05-0408. 44

Gardent, Claire and Bonnie Webber (2001). Towards the use of automated reasoning in discourse disambiguation. *Journal of Logic, Language, and Information*, 10(4):487–509. URL http://www.jstor.org/stable/40180248. 17

Gardiner, Mary and Mark Dras (2007a). Corpus statistics approaches to discriminating among near-synonyms. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007)*, pages 31–39. Melbourne, Australia. 58, 71, 73

Gardiner, Mary and Mark Dras (2007b). Exploring approaches to discriminating among near-synonyms. In *Proceedings of the Australasian Language Technology Workshop 2007*, pages 31–39. Melbourne, Australia. URL http://www.aclweb.org/anthology/U07-1007. 34, 58

Gardiner, Mary and Mark Dras (2012). Valence shifting: Is it a valid task? In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 42–51. Dunedin, New Zealand. URL http://www.aclweb.org/anthology/U/U12/U12-1007. 110

Gatt, Albert and Anja Belz (2008). Attribute selection for referring expression generation: New algorithms and evaluation methods. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 50–58. URL http://www.aclweb.org/anthology/W08-1108. 26

Gatt, Albert; Anja Belz; and Eric Kow (2008). The tuna challenge 2008: Overview and evaluation results. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 198–208. URL http://www.aclweb.org/anthology/W08-1131. 26

Gatt, Albert and Ehud Reiter (2009). SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93. Association for Computational Linguistics, Athens, Greece. URL http://www.aclweb.org/anthology/W09-0613. 25

Genzel, Dmitriy and Eugene Charniak (2002). Entropy Rate Constancy in Text. In *Proceedings of the 40th Annual Meetings of the Association for Computational Linguistics (ACL'02)*, pages 199–206. Philadelphia, US. doi:10.3115/1073083.1073117. URL http://www.aclweb.org/anthology/P02-1026. 102

Genzel, Dmitriy and Eugene Charniak (2003). Variation of Entropy and Parse Trees of Sentences as a Function of the Sentence Number. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 65–72. Sapporo, Japan. URL http://www.aclweb.org/anthology/W03-1009. 102

Gillick, Dan and Yang Liu (2010). Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151. Association for Computational Linguistics, Los Angeles. URL http://www.aclweb.org/anthology/W10-0722. 123

Goldberg, Eli; Norbert Driedger; and Richard Kitteredge (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53. doi:10.1109/64.294135. 1

González-Ibáñez, Roberto; Smaranda Muresan; and Nina Wacholder (2011). Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586. Association for Computational Linguistics, Portland, Oregon, USA. URL http://www.aclweb.org/anthology/P11-2102. 115

Good, Irving John (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3–4):237–264. doi:10.1093/biomet/40.3-4.237. 90

Gorniak, Peter and Deb Roy (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470. doi:10.1613/jair.1327. 27

Grefenstette, Gregory (1999). The World Wide Web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and Computers*. London, UK. 31

Guerini, Marco; Carlo Strapparava; and Oliviero Stock (2008). Valentino: A tool for valence shifting of natural language texts. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakesh, Morocco. 51, 52

Guerini, Marco; Carlo Strapparava; and Oliviero Stock (2011). Slanting existing text with Valentino. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI '11)*, pages 439–440. Palo Alto, CA, USA. doi:10.1145/1943403.1943488. xvii, 52

Hassan, Samer; Andras Csomai; Carmen Banea; Ravi Sinha; and Rada Mihalcea (2007). UNT: SubFinder: Combining knowledge sources for automatic lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 410–413. Association for Computational Linguistics, Prague, Czech Republic. URL http://www.aclweb.org/anthology/S07-1091. 37

Hatzivassiloglou, Vasileios and Kathleen R. McKeown (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 174–181. Association for Computational Linguistics, Madrid, Spain. doi:10.3115/976909.979640. URL http://www.aclweb.org/anthology/P97-1023. 44, 46

Hawker, Tobias; Mary Gardiner; and Andrew Bennetts (2007). Practical queries of a massive n-gram database. In *Proceedings of the Australasian Language Technology Workshop 2007*, pages 40–48. Melbourne, Australia. URL http://www.aclweb.org/anthology/U07-1008. 61

Hayakawa, Samuel I., editor (1968). *Use The Right Word: Modern Guide to Synonyms and Related Words*. The Reader's Digest Association Pty. Ltd, 1st edition. 83, 165

Hayakawa, Samuel I., editor (1994). *Choose the Right Word*. Harper Collins Publishers, 2nd edition. Revised by Eugene Ehrlich. 51, 66, 83

Hearst, Marti A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics (COLING '92)*, volume 2, pages 539–545. Stroudsburg, PA, USA. URL http://www.aclweb.org/anthology/C92-2082. 12

Hearst, Marti A. (1998). Automated discovery of WordNet relations. In Fellbaum (1998), pages 131–153. 12

Hirst, Graeme (1995). Near-synonymy and the structure of lexical knowledge. In *Working notes, AAAI Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity*, pages 51–56. Stanford University. 15, 16

Hirst, Graeme (2003). Paraphrasing paraphrased. URL http://ftp.cs.toronto.edu/pub/gh/Hirst-IWP-talk.pdf. Invited talk, *ACL International Workshop on Paraphrasing*. 18

Holmes, David I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117. doi:10.1093/llc/13.3.111. 53

Ide, Nancy and Jean Véronis (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40. URL http://www.aclweb.org/anthology/J98-1001. 97

Inkpen, Diana (2004). *Building a Lexical Knowledge-Base of Near-Synonym Differences*. Ph.D. thesis, University of Toronto. 13, 18

Inkpen, Diana (2007a). Near-synonym choice in an intelligent thesaurus. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 356–363. Association for Computational Linguistics, Rochester, New York. URL http://www.aclweb.org/anthology/N07-1045. 24

Inkpen, Diana (2007b). A statistical model for near-synonym choice. *ACM Transactions of Speech and Language Processing*, 4(1):1–17. doi:10.1145/1187415.1187417. 4, 5, 7, 13, 24, 27, 28, 30, 31, 32, 33, 35, 36, 59, 60, 62, 71, 87, 88, 93, 94

Inkpen, Diana and Graeme Hirst (2006). Building and using a lexical knowledge-base of near-synonym differences. *Computational Linguistics*, 32(2):223–262. doi:10.1162/coli.2006.32.2.223. 5, 13, 15, 16, 18, 24, 57, 66, 83, 88, 92

Inkpen, Diana Zaiu; Ol'ga Feiguina; and Graeme Hirst (2006). Generating more-positive or more-negative text. In James G. Shanahan; Yan Qu; and Janyce Wiebe, editors, *Computing Attitude and Affect in Text (Selected papers from the Proceedings of the Workshop on Attitude*

and Affect in Text, AAAI 2004 Spring Symposium), pages 187–196. Springer, Dordrecht, The Netherlands. doi:10.1007/1-4020-4102-0_15. 51, 111

Inkpen, Diana Zaiu and Graeme Hirst (2002). Acquiring collocations for lexical choice between near-synonyms. In *Unsupervised Lexical Acquisition: Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 67–76. Association for Computational Linguistics, Philadelphia, USA. URL http://www.aclweb.org/anthology/W02-0909. xiii, 30, 31

Ipeirotis, Panos (2010). Analyzing the Amazon Mechanical Turk marketplace. Technical Report CeDER-10-04. URL http://hdl.handle.net/2451/29801. 123

Islam, Aminul and Diana Inkpen (2009). Real-word spelling correction using Google Web 1T n-gram dataset. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*. Hong Kong. doi:10.1145/1645953.1646205. 72

Islam, Aminul and Diana Inkpen (2010). Near-synonym choice using a 5-gram language model. *Research in Computing Science: Special issue on Natural Language Processing and its Applications*, 46:41–52. 13, 27, 28, 32, 33, 87, 88, 92

Islam, Md. Aminul (2011). *An Unsupervised Approach to Detecting and Correcting Errors in Text*. Ph.D. thesis, University of Ottawa. URL http://www.ruor.uottawa.ca/en/handle/10393/20049. 24, 27, 28, 34

Jaccard, Paul (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579. 27

Janarthanam, Srini and Oliver Lemon (2011). The gruve challenge: Generating routes under uncertainty in virtual environments. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 208–211. Association for Computational Linguistics, Nancy, France. URL http://www.aclweb.org/anthology/W11-2830. 26

Joachims, Thorsten (1999). Making large-scale SVM learning practical. In Bernhard Schölkopf; Christopher J.C. Burges; and Alexander J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. The MIT Press, Cambridge, USA. 94, 143, 147

Joshi, Aravind; Bonnie Webber; and Ralph M. Weischedel (1984). Preventing false inferences. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 134–138. Association for Computational Linguistics, Stanford, California, USA. doi:10.3115/980491.980520. URL http://www.aclweb.org/anthology/P84-1029. 20

Joshi, Mahesh; Dipanjan Das; Kevin Gimpel; and Noah A. Smith (2010). Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296. Association for Computational Linguistics, Los Angeles, California. URL http://www.aclweb.org/anthology/N10-1038. 39

Jurafsky, Daniel and James H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall, 2nd edition. 89, 122

Kacmarcik, Gary and Michael Gamon (2006). Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 444–451. Association for Computational Linguistics, Sydney, Australia. URL http://www.aclweb.org/anthology/P06-2058. 54

Kamps, Jaap and Maarten Marx (2002). Words with attitude. In *Proceedings of the 1st International Conference on Global WordNet*, pages 332–341. Mysore, India. 44

Kamps, Jaap; Robert J. Mokken; Maarten Marx; and Maarten de Rijke (2004). Using WordNet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, volume 4, pages 1115–1118. European Language Resources Association, Paris. 44

Kang, Sin-Jae and Jong-Hyeok Lee (2001). Semi-automatic practical ontology construction by using a thesaurus, computational dictionaries, and large corpora. In *Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management*. URL http://www.aclweb.org/anthology/W01-1006. 12

Katagiri, Shigeru; Biing-Hwang Juang; and Chin-Hui Lee (1998). Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method. *Proceedings of the IEEE*, 86(11):2345–2373. doi:10.1109/5.726793. 34

Katz, Slava M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35:400–401. doi:10.1109/TASSP.1987.1165125. 89, 90

Keller, Frank (2004). The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 317–324. Association for Computational Linguistics, Barcelona, Spain. URL http://www.aclweb.org/anthology/W04-3241. 102

Kennedy, Alistair and Diana Inkpen (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22:110–125. doi:10.1111/j.1467-8640.2006.00277.x. 41

Keppel, Geoffrey and Thomas D. Wickens (2004). *Design and Analysis: A Researcher's Handbook*. Pearson Prentice Hall, Upper Saddle River, New Jersey, USA, fourth edition. 135

Keshtkar, Fazel (2011). *A Computational Approach to the Analysis and Generation of Emotion in Text*. Ph.D. thesis, University of Ottawa. xiii, 19, 24, 25

Kietz, Jörg-Uwe; Alexander Maedche; and Raphael Volz (2000). A method for semi-automatic ontology acquisition from a corporate intranet. In *Proceedings of the ECAW-2000 Workshop on Ontologies and Text*. 12

Kilgarriff, Adam (2001). English lexical sample task description. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 17–20. Association for Computational Linguistics, Toulouse, France. URL http://www.aclweb.org/anthology/S01-1004. 65

Kipper, Karin; Anna Korhonen; Neville Ryant; and Martha Palmer (2008). A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40. doi:10.1007/s10579-007-9048-2. 18

Klein, Martin and Michael L. Nelson (2009). Correlation of term count and document frequency for Google n-grams. In Mohand Boughanem; Catherine Berrut; Josiane Mothe; and Chantal Soulé-Dupuy, editors, *Advances in Information Retrieval - 31th European Conference on IR Research - ECIR 2009*, pages 620–627. Toulouse, France. doi:10.1007/978-3-642-00958-7_58. 72

Koppel, Moshe; Navot Akiva; and Ido Dagan (2003). A corpus-independent feature set for style based text categorization. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*. Acapulco, Mexico. 39

Koppel, Moshe; Navot Akiva; and Ido Dagan (2006). Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology*, 57(11):1519–1525. doi:10.1002/asi.20428. 98

Koppel, Moshe; Shlomo Argamon; and Anat Rachel Shimoni (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412. doi:10.1093/llc/17.4.401. 39

Koppel, Moshe and Jonathan Schler (2004). Authorship verification as a one-class classification problem. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, pages 62–68. ACM, New York, NY, USA. ISBN 1-58113-828-5. doi:10.1145/1015330.1015448. 54

Krahmer, Emiel and Kees van Deemter (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218. doi:10.1162/COLI_a_00088. 26

Krahmer, Emiel and Mariët Theune, editors (2010). *Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation*. Springer, Berlin / Heidelberg. doi:10.1007/978-3-642-15573-4. 22

Kullback, Solomon and Richard A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86. URL http://www.jstor.org/stable/2236703. 98, 118

Landauer, Thomas and Susan Dumais (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240. doi:10.1037/0033-295X.104.2.211. 33

Langkilde, Irene and Kevin Knight (1998). The practical value of N-grams in generation. In *Proceedings of the 9th International Natural Language Generation Workshop*, pages 248–255. Niagra-on-the-Lake, Canada. URL http://www.aclweb.org/anthology/W98-1426. 88

Lapata, Maria (2001). A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 63–70. Pittsburgh, PA. URL http://aclweb.org/anthology/N01-1009. 127

Lee, Yoong Keok and Hwee Tou Ng (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 41–48. Association for Computational Linguistics. doi:10.3115/1118693.1118699. URL http://www.aclweb.org/anthology/W02-1006. 21

Levy, Roger and T. Florian Jaeger (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf; J. Platt; and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA. 102

Lin, Dekang (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 768–774. Association for Computational Linguistics, Montreal, Quebec, Canada. doi:10.3115/980691.980696. URL http://www.aclweb.org/anthology/P98-2005. 18, 19

Lin, Dekang and Lin Pantel (2001). DIRT - Discovery of Inference Rules from Text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 323–328. San Francisco, USA. doi:10.1145/502512.502559. 19

Linguistic Data Consortium (2005). Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report. URL http://wayback.archive.org/

web/*/http://projects.ldc.upenn.edu/TIDES/Translation/TransAssess04.pdf. xvii, 19, 127, 128

Liu, Bing (2012). *Sentiment Analysis and Opinion Mining: A Comprehensive Introduction and Survey*. Morgan & Claypool. doi:10.2200/S00416ED1V01Y201204HLT016. 38, 39, 41, 114

Liu, Chien-Liang; Chia-Hoang Lee; Ssu-Han Yu; and Chih-Wei Chen (2011). Computer assisted writing system. *Expert Systems with Applications*, 38(1):804–811. doi:10.1016/j.eswa.2010.07.038. 1

Liu, Jingjing; Yunbo Cao; Chin-Yew Lin; Yalou Huang; and Ming Zhou (2007). Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 334–342. Association for Computational Linguistics, Prague, Czech Republic. URL http://www.aclweb.org/anthology/D07-1035. 39

Liu, Yi and Yuan F. Zheng (2005). One-against-all multi-class SVM classification using reliability measures. In *Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, 2005. (IJCNN '05.)*, volume 2, pages 849–854. doi:10.1109/IJCNN.2005.1555963. 94

Lyons, John (1977). *Semantics*. Cambridge University Press, Cambridge, UK. 10, 11

Maas, Andrew L.; Raymond E. Daly; Peter T. Pham; Dan Huang; Andrew Y. Ng; and Christopher Potts (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics, Portland, Oregon, USA. URL http://www.aclweb.org/anthology/P11-1015. xiii, 46

Madnani, Nitin and Bonnie J. Dorr (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3). doi:10.1162/coli_a_00002. 18, 19

Mairesse, François and Marilyn A. Walker (2010). Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20:227–278. doi:10.1007/s11257-010-9076-2. 23

Manning, Christopher D.; Prabhakar Raghavan; and Hinrich Schütze (2008). *An Introduction to Information Retrieval*. Cambridge University Press. 120

Marks, Lawrence E. (1974). *Sensory Processes: The New Psychophysics*. Academic Press, New York. 127

Matveeva, Irina; Gina-Anne Levow; Ayman Farahat; and Christiaan Royer (2005). Term representation with generalized latent semantic analysis. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-05)*. Borovets, Bulgaria. 37

McCarthy, Diana; Rob Koeling; Julie Weeds; and John Carroll (2004). Finding predominant word senses in untagged text. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 279–286. Barcelona, Spain. doi:10.3115/1218955.1218991. URL http://www.aclweb.org/anthology/P04-1036. 18

McCarthy, Diana and Roberto Navigli (2007). SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53. Association for Computational Linguistics, Prague, Czech Republic. URL http://www.aclweb.org/anthology/W07-2009. 37, 152

McKeown, Kathleen R. (1979). Paraphrasing using given and new information in a question-answer system. In *Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics*, pages 67–72. Association for Computational Linguistics, La Jolla, California, USA. doi:10.3115/982163.982182. URL http://www.aclweb.org/anthology/P79-1016. 19

Mejova, Yelena and Padmini Srinivasan (2011). Exploring feature definition and selection for sentiment classifiers. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 546–549. Association for the Advancement of Artificial Intelligence. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2808. 41

Mill, John Stuart (1843). *A System of Logic*. Longmans, London, England. 11

Mitchell, Tom M. (1997). *Machine Learning*. The McGraw-Hill Companies, Inc, New York, USA. 118

Mititelu, Verginica Barbu (2006). Automatic extraction of patterns displaying hyponym-hypernym co-occurrence from corpora. In *Proceedings of the First Central European Student Conference in Linguistics (CESCL)*. URL http://www.nytud.hu/cescl/proceedings/Verginica_Mititelu_CESCL.pdf. 12

Mohammad, Saif and Peter Turney (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34. Association for Computational Linguistics, Los Angeles, CA. URL http://www.aclweb.org/anthology/W10-0204. 115

Mohammad, Saif M. and Peter D. Turney (2012). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*. doi:10.1111/j.1467-8640.2012.00460.x. xiv, 115, 116

Molla, Diego and Maria Elena Santiago-Martinez (2011). Development of a corpus for evidence based medicine summarisation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 86–94. Canberra, Australia. URL http://www.aclweb.org/anthology/U11-1012. 124, 125

Mooney, Raymond J. (1996). Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 82–91. URL http://www.aclweb.org/anthology/W96-0208. 21

Moore, David S. and George P. McCabe (2003). *Introduction to the Practice of Statistics*. W. H. Freeman and Company, 4th edition. 73

Narayanan, Ramanathan; Bing Liu; and Alok Choudhary (2009). Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 180–189. Association for Computational Linguistics, Singapore. URL http://www.aclweb.org/anthology/D09-1019. 115

Navigli, Roberto (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10:1–10:69. ISSN 0360-0300. doi:10.1145/1459352.1459355. 21

Navigli, Roberto; Kenneth C. Litkowski; and Orin Hargraves (2007). Semeval-2007 task 07: Coarse-grained English all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35. Association for Computational Linguistics, Prague, Czech Republic. URL http://www.aclweb.org/anthology/W07-2006. 65

Nielsen, Finn Årup (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In Matthew Rowe; Milan Stankovic; Aba-Sah Dadzie; and Mariann Hardey, editors, *Proceedings, 1st Workshop on Making Sense of Microposts (#MSM2011): Big things come in small packages*, pages 93–98. Heraklion, Crete, Greece. URL http://ceur-ws.org/Vol-718/paper_16.pdf. 131

Palmer, Martha; Daniel Gildea; and Paul Kingsbury (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106. doi:10.1162/0891201053630264. 18

Pang, Bo and Lillian Lee (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 271–278. Barcelona, Spain. doi:10.3115/1218955.1218990. URL http://www.aclweb.org/anthology/P04-1035. 40, 43

Pang, Bo and Lillian Lee (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124. Association for Computational Linguistics, Ann Arbor, Michigan. doi:10.3115/1219840.1219855. URL http://www.aclweb.org/anthology/P05-1015. xvii, 38, 43, 47, 48, 82, 109, 173

Pang, Bo and Lillian Lee (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1—135. doi:10.1561/1500000011. 38, 41

Pang, Bo; Lillian Lee; and Shivakumar Vaithyanathan (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86. Association for Computational Linguistics. doi:10.3115/1118693.1118704. URL http://www.aclweb.org/anthology/W02-1011. 39, 40, 41, 81, 94, 95, 97, 106

Paris, Cécile; Nathalie Colineau; and Ross Wilkinson (2007). NLG systems evaluation: a framework to measure impact on and cost for all stakeholders. In Dale and White (2007), pages 16–17. URL http://www.ling.ohio-state.edu/nlgeval07/papers/NLGEval07-Position-Papers.pdf. 26

Paşca, Marius and Péter Dienes (2005). Aligning needles in a haystack: Paraphrase acquisition across the Web. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP 2005)*, pages 119–130. URL http://www.aclweb.org/anthology/I05-1011. 19

Pedersen, Ted (1996). Fishing for exactness. In *In Proceedings of the South-Central SAS Users Group Conference*, pages 188–200. 30

Przybocki, Mark; Kay Peterson; and Sébastien Bronsart (2008). Translation adequacy and preference evaluation tool (TAP-ET). In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odjik; Stelios Piperidis; and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), Marrakech, Morocco. ISBN 2-9517408-4-0. URL http://www.lrec-conf.org/proceedings/lrec2008/. 127

Qian, Ting and T. Florian Jaeger (2010). Close = relevant? the role of context in efficient language production. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, pages 45–53. Association for Computational Linguistics, Uppsala, Sweden. URL http://www.aclweb.org/anthology/W10-2006. 102

Quirk, Randolph; Sidney Greenbaum; Geoffrey Leech; and Jan Svartvik (1985). *A comprehensive grammar of the English language*. Longman. 39

Raaijmakers, Stephan and Wessel Kraaij (2008). A shallow approach to subjectivity classification. In *Proceedings of the Second International Conference on Weblogs and Social Media, ICWSM 2008*, pages 216–217. Seattle, Washington, USA. URL http://www.aaai.org/Library/ICWSM/2008/icwsm08-051.php. 43

Raaijmakers, Stephan; Khiet Truong; and Theresa Wilson (2008). Multimodal subjectivity analysis of multiparty conversation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 466–474. Association for Computational Linguistics, Honolulu, Hawaii. URL http://www.aclweb.org/anthology/D08-1049. 43

Ramana, Aditya; Srinath Srinivasa; Sudarshan Murthy; Avinash Reddy Palleti; and Ramya Krishna Y. (2010). Comparing web n-grams and other means of identifying named entities in corporate blogs. In *Proceedings of Web N-gram Workshop, Workshop of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–46. ACM, Geneva, Switzerland. 72

Rao, Josyula R. and Pankaj Rohatgi (2000). Can pseudonymity really guarentee privacy? In *Proceedings of the 9th USENIX Security Symposium*. Denver, Colorado, USA. 54

Rapp, Reinhard (2003). Word sense discovery based on sense descriptor dissimilarity. In *MT Summit IX: proceedings of the Ninth Machine Translation Summit*. New Orleans, USA. 37

Rapp, Reinhard (2008). The automatic generation of thesauri of related words for English, French, German, and Russian. *International Journal of Speech Technology*, 11(3–4):147–156. doi:10.1007/s10772-009-9043-7. 33

Refaeilzadeh, Payam; Lei Tang; and Huan Liu (2009). Cross Validation. In M. Tamer and Ling Liu, editors, *Encyclopedia of Database Systems*. Springer. 95

Reiter, Ehud (2011). Task-based evaluation of nlg systems: Control vs real-world context. In *Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop*, pages 28–32. Association for Computational Linguistics, Edinburgh, Scotland. URL http://www.aclweb.org/anthology/W11-2704. 26

Reiter, Ehud and Anja Belz (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558. doi:10.1162/coli.2009.35.4.35405. 25

Reiter, Ehud and Robert Dale (2000). *Building Natural Language Generation Systems*. Cambridge University Press. xiii, 22, 26

Rifkin, Ryan and Aldebaro Klautau (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5(2):101–141. URL http://jmlr.csail.mit.edu/papers/volume5/rifkin04a/rifkin04a.pdf. 94

Riloff, Ellen; Siddharth Patwardhan; and Janyce Wiebe (2006). Feature subsumption for opinion analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 440–448. Association for Computational Linguistics, Sydney, Australia. URL http://www.aclweb.org/anthology/W06-1652. 41

Riloff, Ellen and Janyce Wiebe (2003). Learning extraction patterns for subjective expressions. In Michael Collins and Mark Steedman, editors, *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105–112. URL http://www.aclweb.org/anthology/W03-1014. 40, 42

Salton, Gerard and Christopher Buckley (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523. 121

Salzberb, Steven L. (1997). On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. *Data Mining and Knowledge Discovery*, 1:317–327. 94

Scott, Donia and Johanna Moore (2007). An NLG evaluation competition? eight reasons to be cautious. In Dale and White (2007), pages 22–23. URL http://www.ling.ohio-state.edu/nlgeval07/papers/NLGEval07-Position-Papers.pdf. 26

Sebastiani, Fabrizio (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47. doi:10.1145/505282.505283. 121

Shannon, Claude E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656. 76

Siegel, Sidney and N. John Castellan, Jr. (1988). *Nonparametric statistics for the behavioural sciences*. McGraw Hill, Boston. 68

Snow, Rion; Daniel Jurafsky; and Andrew Y. Ng (2006). Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 801–808. Association for Computational Linguistics, Sydney, Australia. doi:10.3115/1220175.1220276. URL http://www.aclweb.org/anthology/P06-1101. 18

Snow, Rion; Brendan O'Connor; Daniel Jurafsky; and Andrew Ng (2008). Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics, Honolulu, Hawaii. URL http://www.aclweb.org/anthology/D08-1027. 123, 124

Spertus, Ellen (1997). Smokey: Automatic recognition of hostile messages. In *Innovative Applications of Artificial Intelligence (IAAI) '97*, pages 1058–1065. AAAI Press, Menlo Park, California. 42

Sprent, Peter and Nigel C. Smeeton (2007). *Applied Nonparametric Statistical Methods*. Texts in Statistical Science. Chapman and Hall/CRC, 4th edition. 95

Spärck Jones, Karen (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 60(1):11–21. 120

Stone, Philip J.; Dexter C. Dunphy; Marshall S. Smith; and Daniel M. Ogilvie (1966). *General Inquirer: A Computer Approach to Content Analysis*. The MIT Press. 44, 50, 51

Strapparava, Carlo and Alessandro Valitutti (2004). WordNet-Affect: an affective extension of WordNet. In *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC 2004)*, pages 1083–1086. 50

Striegnitz, Kristina; Alexandre Denis; Andrew Gargett; Konstantina Garoufi; Alexander Koller; and Mariët Theune (2011). Report on the second challenge on Generating Instructions in Virtual Environments (GIVE-2.5). In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 270–279. Association for Computational Linguistics, Nancy, France. URL http://www.aclweb.org/anthology/W11-2845. 26

Stubbs, Michael (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell Publishing, Oxford, UK. 13

Su, Fangzhong and Katja Markert (2008a). Eliciting subjectivity and polarity judgements on word senses. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 42–50. Coling 2008 Organizing Committee, Manchester, UK. URL http://www.aclweb.org/anthology/W08-1207. 50

Su, Fangzhong and Katja Markert (2008b). From words to senses: A case study of subjectivity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 825–832. Coling 2008 Organizing Committee, Manchester, UK. URL http://www.aclweb.org/anthology/C08-1104. 50

Su, Qi; Dmitry Pavlov; Jyh-Herng Chow; and Wendell C. Baker (2007). Internet-scale collection of human-reviewed data. In *Proceedings of the Sixteenth International World Wide Web Conference (WWW2007)*, pages 231–240. Banff, Alberta, Canada. 123

Taboada, Maite; Julian Brooke; Milan Tofiloski; Kimberly Voll; and Manfred Stede (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307. doi:10.1162/COLI_a_00049. 41

Takamura, Hiroya; Takashi Inui; and Manabu Okumura (2006). Latent variable models for semantic orientations of phrases. In *Proceedings of the 11 th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006*, pages 201–208. URL http://www.aclweb.org/anthology/E06-1026. 45

Tarski, Alfred (1936). Der Wahrheitsbegriff in den formalisierten Sprachen (the concept of truth in formalized languages). *Studia Philosophica*, 1:261–405. English translation by J. H. Woodger in Tarski, Alfred (1956) *Logic, Semantics, Metamathematics: Papers From 1923 to 1938*. 10

Tratz, Stephen and Eduard Hovy (2010). A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687. Association for Computational Linguistics, Uppsala, Sweden. URL http://www.aclweb.org/anthology/P10-1070. 123

Tsur, Oren; Dmitry Davidov; and Ari Rappoport (2010). A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 162–169. Association for the Advancement of Artificial Intelligence. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1495. 115

Turney, Peter (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML 2001)*, pages 491–502. Freiburg, Germany. 31

Turney, Peter (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA. doi:10.3115/1073083.1073153. URL http://www.aclweb.org/anthology/P02-1053. 38, 41, 43, 44, 45

Turney, Peter D. and Michael Littman (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346. doi:10.1145/944012.944013. 38, 44

Turney, Peter D.; Michael L. Littman; Jeffrey Bigham; and Victor Shnayder (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference RANLP-2003 (Recent Advances in Natural Language Processing)*, pages 482–489. Borovets, Bulgaria. 37

Viethen, Jette and Robert Dale (2006). Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 63–70. Association for Computational Linguistics, Sydney, Australia. URL http://www.aclweb.org/anthology/W06-1410. 27

Wais, Paul; Shivaram Lingamneni; Duncan Cook; Jason Fennell; Benjamin Goldenberg; Daniel Lubarov; David Marin; and Hari Simons (2010). Towards building a high-quality workforce with Mechanical Turk. In *Proceedings of Computational Social Science and the Wisdom of Crowds (NIPS)*, pages pages 1–5. Whister. 123, 124

Wang, Tong and Graeme Hirst (2010). Near-synonym lexical choice in latent semantic space. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1182–1190. Coling 2010 Organizing Committee, Beijing, China. URL http://www.aclweb.org/anthology/C10-1133. 13, 27, 28, 33, 81

Weaver, Warren (1955). Translation. In William N. Locke and A. Donald Booth, editors, *Machine Translation of Languages: Fourteen Essays*, pages 15–23. Technology Press of MIT, Cambridge, MA, USA. 21

Weeds, Julie Elizabeth (2003). *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, University of Sussex. 118, 119, 146, 150, 156

Whitehead, Simon and Lawrence Cavedon (2010). Generating shifting sentiment for a conversational agent. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 89–97. Association for Computational Linguistics, Los Angeles, CA. URL http://www.aclweb.org/anthology/W10-0211. 49, 52, 53, 111

Wiebe, Janyce and Rada Mihalcea (2006). Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1065–1072. Association for Computational Linguistics, Sydney, Australia. doi:10.3115/1220175.1220309. URL http://www.aclweb.org/anthology/P06-1134. 45, 68

Wiebe, Janyce; Theresa Wilson; Rebecca Bruce; Matthew Bell; and Melanie Martin (2004). Learning subjective language. *Computational Linguistics*, 30(3):277–308. doi:10.1162/0891201041850885. 42, 97, 98, 100

Wiebe, Janyce; Theresa Wilson; and Claire Cardie (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3):165–210. doi:10.1007/s10579-005-7880-9. 43, 50

Wiebe, Janyce M. (1990). Identifying subjective characters in narrative. In *Papers presented to the 13th International Conference on Computational Linguistics (COLING 1990)*, volume 2, pages 401–406. URL http://www.aclweb.org/anthology/C90-2069. 39

Wiebe, Janyce M. (1994). Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287. URL http://www.aclweb.org/anthology/J94-2004. 39

Wilson, Theresa; Janyce Wiebe; and Paul Hoffmann (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics, Vancouver, British Columbia, Canada. URL http://www.aclweb.org/anthology/H05-1044. 45, 50

Wilson, Theresa; Janyce Wiebe; and Rebecca Hwa (2006). Recognizing strong and weak opinion clauses. *Computational Intelligence*, 22(2):73–99. doi:10.1111/j.1467-8640.2006.00275.x. 42

Yang, Yiming and Jan O. Pedersen (1997). A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420. Morgan Kaufmann Publishers Inc., San Francisco, USA. 39, 100, 118

Yano, Tae and Noah A. Smith (2010). What's worthy of comment? content and comment volume in political blogs. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 359–362. Association for the Advancement of Artificial Intelligence. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1503. 39

Yarowsky, David (1992). Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, pages 454–460. Nantes, France. URL http://www.aclweb.org/anthology/C92-2070. 21

Yessenalina, Ainur and Claire Cardie (2011). Compositional matrix-space models for sentiment analysis. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 172–182. Association for Computational Linguistics, Edinburgh, Scotland, UK. URL http://www.aclweb.org/anthology/D11-1016. 46, 150

Yessenalina, Ainur; Yisong Yue; and Claire Cardie (2010). Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1046–1056. Association for Computational Linguistics, Cambridge, MA. URL http://www.aclweb.org/anthology/D10-1102. 41

Yu, Liang-Chih; Hsiu-Min Shih; Yu-Ling Lai; Jui-Feng Yeh; and Chung-Hsien Wu (2010). Discriminative training for near-synonym substitution. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1254–1262. Coling 2010 Organizing Committee, Beijing, China. URL http://www.aclweb.org/anthology/C10-1141. 27, 28, 33, 34, 35, 81

Yuret, Deniz (2007). KU: Word sense disambiguation by substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 207–214. Association for Computational Linguistics, Prague, Czech Republic. URL http://www.aclweb.org/anthology/S07-1044. 33, 37

Zagibalov, Taras and John Carroll (2008a). Automatic seed word selection for unsupervised sentiment classification of chinese text. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1073–1080. Coling 2008 Organizing Committee, Manchester, UK. URL http://www.aclweb.org/anthology/C08-1135. 44

Zagibalov, Taras and John Carroll (2008b). Unsupervised classification of sentiment and objectivity in chinese text. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 304–311. Hyderabad, India. URL http://www.aclweb.org/anthology/I08-1040. 44

Zhao, Shiqi; Xiang Lan; Ting Liu; and Sheng Li (2009). Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 834–842. Association for Computational Linguistics, Suntec, Singapore. URL http://www.aclweb.org/anthology/P09-1094. 19