Sentiment and near-synonymy: do they go together?

Mary Gardiner
Centre for Language Technology
Macquarie Univeristy

gardiner@ics.mq.edu.au

Abstract

Near-synonyms are words that mean approximately the same thing, and which tend to be assigned to the same leaf in ontologies such as WordNet. However choosing-between them is still a significant problem for natural language generation systems, as such words may differ in crucial respects, such as in having a positive, neutral or negative attitude to the subject of the generated language; or in having denotational differences that may be important to the discourse.

Previous work in identifying and using near-synonyms has treated these the same and has concluded, on the basis of initial investigations, that a corpus statistics approach is not useful for the problem. However, as they are different, then corpus statistics may still be applicable to a subtype. In particular if near-synonyms differing in attitude respond better to corpus statistics, this suggests that an approach based on the extensive work in sentiment analysis is worth pursuing. This paper presents initial results showing that this is in fact the case, and presents a research programme based on this.

1 Introduction

The problem of choosing an appropriate word or phrase from among candidate near-synonyms or paraphrases is important for language generation. Barzilay and Lee (2003) cite summarisation and rewriting as among the possible applications, and point out that a component of the system will need to choose among the candidates based on various criteria including length and sophistication. An application of near-synonym generation is the extension of the text generation system HALogen (Langkilde and Knight, 1998; Langkilde, 2000) to include near-synonyms (Inkpen and Hirst, 2006).

An aspect of the choice between synonyms or paraphrases that should not be neglected is any difference in meaning or attitude. Currently, synonyms and paraphrases are usually treated as completely interchangeable in computational systems. But ideally, for example, a system should be able to make a correct choice between *frugal* and *stingy* when trying to describe a person whom the system is intending to praise.

There are several alternative possible approaches to choosing between words like *frugal* and *stingy*:

1. choose between them using exactly the same methods as the system uses to choose between fairly semantically unrelated words;
2. choose between them using the same methods that the system uses to choose between any two closely semantically related words, even if they do not differ significantly in attitude (eg *battle* and *fight*); or
3. choose between them using a method especially designed for choosing between closely related words with attitude differences.

Some existing work explores using special methods to choose between closely related words

(usually near-synonyms) with success better than that of using general natural language generation techniques to choose between themb (Inkpen and Hirst, 2004; Inkpen and Hirst, 2006).

Inkpen et al. (2006) describe techniques for choosing words with either positive or negative sentiment with the aim of producing better texts, but do not specifically justify the need for a special technique to solve the attitude choice problem in particular. In general, especially given the relatively poor performance of the Edmonds (1997) method, the research seems to have tended away from using corpus statistics to solve this problem, at least until the work of Inkpen (2007).

Sentiment analysis work such as that of Pang et al. (2002) and Turney (2002) suggests that it is possible to acquire the sentiment or orientation of pieces of text ranging from words to documents using corpus statistics without needing to use lexicographic resources prepared by experts. This also suggests that the sentiment of a word may affect its collocational context quite broadly. For example, taking two cases from the classification scheme above, it seems intuitively plausible that differences between *placid* (positive) and *unimaginative* (negative) may be expressed throughout the document in which they are found, while for the denotational pair *invasion* and *incursion* there is no reason why the document more broadly should reflect the precise propositional differences that are the essence of the denotational subtype. Therefore, it is possible that the results of the Edmond's experiment vary depending on whether the near-synonyms differ in sentiment expressed towards their subject (attitudinal), or whether they differ in some other way.

In this paper we outline an inquiry into whether approach 2 or approach 3 is more promising: is distinguishing between closely related words that differ in affect different from distinguishing between closely related words that do not differ in affect? In particular, I are exploring whether or not context cues sentiment charged choices more than it cues choices between related words without sentiment differences. Given promising results indicating that context cues may be more important for choosing between sentimentally charged near-synonyms, we outline a possible approach to acquiring such differences automatically.

In Section 2 we outline the general method we are using to test whether sentiment differences between closely related words and other differences between closely related words can be predicted equally well by context or not. In Section 3 we describe an annotation
experiment dividing sets of near-synonyms into those differing in attitude and those which do no. In Section 4 we describe results from an early experiment using corpus statistics approaches to discriminate between near synonyms. In Section 5 we discuss planned future experiments extending the current method and a future research direction incorporating sentiment analysis techniques into acquisition of near-synonym properties.

2 Task description

Our test for choosing between closely related words is based on that of Edmonds (1997).

The problem that the system is asked to solve is this: given a set of closely related words, choose which word belongs in a lexical gap in a given sentence. For example, the system might be given the sentence below, with the blank indicating a lexical gap, and asked which of *error*, *mistake* or *oversight* best fits in that gap:

"However, such a move also of cutting deeply into U.S. economic growth, which is why some economists think it would be a big _____."

The set of words that the system is asked to choose between might differ in sentiment from each other as, for example *error*, *mistake* and *oversight* do. However, they also might not. For example, the system might be asked to choose between the words *lawyer* and *attorney*, which do not differ in sentiment towards the referent.

The system always uses context cues to choose between the words it is presented with. We compare two sets of test data in terms of how well the system is able to predict the correct word:

1. sets of words where words in each set differ among themselves in affect; and
2. sets of words where words in each set do not differ among themselves in affect.

3 Evaluating near-synonym type

3.1 Method

We conducted an annotation experiment to provide a larger test set of near-synonyms to test our hypothesis against. The annotators were asked to decide whether crtain 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), where synset frequency is the sum of the frequencies of individual members of the synset. There was no normalisation to compare synsets with different numbers of member words.

Synsets were then manually excluded from this set by the author where they were deemed too similar to other more frequent synsets; were internally similar (for example, the synset consisting of *ad*, *advertisement*, *advertizement*); or contained purely dialectical variation (for example *lawyer* and *attorney*). This left 124 synsets of the original 300.

These 124 synsets were then independently annotated by two native English speakers, including the author of this paper, 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: they were either *definite* in their judgement that a synset did or did not differ in attitude; they considered that their judgement was *probably* correct; or they were completely *uncertain*.

3.2 Results

Inter-annotator agreement for the annotation experiment is shown in Table 1 both individually for certainty,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: agreement ``attitudinal difference'', ``not attitudinal difference'' and ``unsure'' regardless of certainty; and agreement between annotators on *only* the annotations they were definitely sure about, as per Wiebe and Mihalcea (2006). We calculated two κ scores: the Cohen (1960) κ and the Siegel and Castellan (1988) κ; however the two scores are identical to two significant figures and hence only one κ value is shown.

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

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

Table 1: 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 2: Inter-annotator agreement and scores for the annotation experiment

4 Edmonds' experiment

In the first experiment testing whether near-synonyms differing in attitude are more responsive to corpus statistics techniques than other near-synonyms, we used the methodology of Edmonds (1997). Edmonds' aim was slightly different from ours, in that his work was designed to explore whether contextual cues are sufficient for choosing between closely related words (near-synonyms) in general, rather than to explore whether some sets of closely related words behaved differently from others. However, we can use his method of prediction and then compare the performance of two groups of near-synonyms.

4.1 Method

Edmonds defined a measure designed to determine which of the set of words is cued most strongly by the sentence with the gap, as approximated by similarity score of that word with each of the set of words in the sentence. So the fittingness of, say, *error* for the gap in the example sentencein Section 2 is approximated by the tendency for each of *er-*

ror and *however*, *error* and *such* and so on to occur together in context.

In general the appropriateness score of any given candidate word *c* for sentence *S* is the sum of the significance scores $sig(c, w)$ for candidate *c* with every other word *w* in the sentence (barring stopwords):

$$appropriateness(c, S) = \sum_{w \in S} sig(c, w)$$

The significance score $sig(c, w)$ between two individual words is computed as follows (where $t(a, b)$ is the *t*-score for bigrams containing *a* and *b* in the training data):

1. if the *t*-score and mutual information score of *c* and *w* on the training data are greater than 2.0 and 3.0 respectively, then $sig(c, w)$ is given by: $sig(c, w) = t(c, w)$
2. if there is a word w_0 such that if the *t*-score and mutual information score of each of the pairs $c(w_0)$ and (w_0, w) are greater than 2.0 and 3.0 respectively, then $sig(c, w)$ is given by: $$sig(c, w) = \frac{1}{8}(t(c, w_0) + t\frac{(w_0, w)}{2})$$
3. otherwise $sig(c, w) = 0$

The candidate word $\underline{c}$ with the highest score $appropriateness(c, S)$ for sentence *S* is selected as the chosen word. If there is more than one candidate with that highest score, no candidate is chosen.

The *t*-scores and mutual information scores are calculated from bigram frequencies in the 1987 Wall Street Journal using 4 and 10 word windows for bigrams as calculated by the Ngram Statistics Package (Banerjee and Pedersen, 2003).

Edmonds' method is compared to a baseline, where the most frequent word in any test word set is chosen.

Our test word sets are drawn from the annotation experiment described in Section 3: they are the 58 synsets where the annotators agreed on the type of the synset and were both certain of their judgement. Thus we had 7 word sets agreed to differ internally in attitude and 51 agreed not to.

Our test data consists of test sentences drawn from either the 1987 Wall Street Journal, or the 1988 Wall Street Journal. There were two sets of training data: the 1989 Wall Street Journal using bigrams drawn from 4 word windows around the target word (the experiments called *4win-* were trained this way) or from 10 word windows around the target word (the experiments called *10win-* were trained this way).

Abbreviation	Training window size	Synset size (min 2)	Wall Street Journal year
4win-top3-1987	4	max 3	1987
4win-top4-1987	4	max 4	1987
4win-top4-1988	4	max 4	1988
10win-top3-1987	10	max 3	1987
10win-top4-1987	10	max 4	1987

Table 3: Test runs for the Edmonds experiment

4.2 Results

Since the Edmonds method cannot always make a prediction, we directly compare the baseline and the Edmonds predictions only on sentences where the Edmond method can make a prediction. The number of times that the Edmonds method can make a prediction at all is shown in Table 4, which also shows the baseline correctness on the sentences described, and the Edmonds method correctness where it can make a prediction. A sentence that contains *n* words from test synsets is counted as *n* separate test sentences in this table.

4.3 Discussion

There are several results of interest here. First, the baselines perform noticeably differently for attitudinal versus non-attitudinal success ratios for each of the five data sets. Calculating the *z*-statistic for comparing two proportions, we find that this difference is significant at the 1% level for each of the data sets, with the attitudinal baseline always higher. Similarly, the difference between attitudinal and non-attitudinal success ratios for Edmonds are also significant at the 1% level.

Because of this first result regarding baselines, the second result, which does show that gross success rates for attitudinal near-synonyms is significantly higher under the Edmonds corpus statistics approach, is less interesting: these higher success ratios could be due to the naturally higher baseline alone.

	Sentences containing test word		Baseline correctness		Edmonds prediction		Edmonds precision	
Test set	**Attitudinal**	**Non-attitudinal**	**Attitudinal**	**Non-attitudinal**	**Attitudinal**	**Non-attitudinal**	**Attitudinal**	**Non-attitudinal**
4win-top3-1987	7588	340246	86.6%	66.2%	5.7%	14.2%	94.7%	67.5%
4win-top4-1987	29453	350038	84.0%	65.9%	8.1%	15.5%	72.3%	62.9%
4win-top4-1988	27023	295437	85.4%	64.0%	7.7%	15.4%	69.2%	62.0%
10win-top3-1987	7588	340246	86.0%	67.8%	14.7%	28.9%	90.3%	58.7%
10win-top4-1987	29453	350038	82.7%	67.4%	15.2%	31.6%	65.6%	54.1%

Table 4: Performance of the baseline and Edmonds method on all test sentences

Test data	All words		Attitudinal words		Non-attitudinal words	
	Baseline	**Edmonds**	**Baseline**	**Edmonds**	**Baseline**	**Edmonds**
4win-top3-1987	9715	10824	5	40	9710	10784
4win-top4-1987	13924	12456	604	326	13320	12130
4win-top4-1988	11752	10861	594	256	11158	10605
10win-top3-1987	28900	20825	14	54	28886	20771
10win-top4-1987	37850	23245	1214	449	36636	22796

Table 5: Number of times each method is right when the baseline and the Edmonds method predict a different word

We inspected some of the data, and noted that for attitudinal synsets, the distribution was much more skewed than for non-attitudinal synsets: one element dominated, and the others were infrequent. In some cases this dominant element appeared to the neutral one, perhaps reflecting the nature of the corpus, but in other cases there was no discernible pattern.

To take into account the varying baselines, we extracted cases where only one method predicted correctly, disregarding those cases where both were right or both wrong. The counts of these are presented in Table 5. We then considered as a `success' any correct prediction by Edmonds, and calculated the proportion of successes for attitudinal and non-attitudinal for each of the five data sets. Then, for each of the data sets, we compared the success ratios for attitudinal and non-attitudinal, again using the *z*-statistic for comparing two proportions. The differences are again significant at the 1% level. In this analysis, the attitudinal synsets perform better only for 4win-top3-1987 and 10win-top3-1987; that is, for the cases where there are at most three elements in the synset. For the cases with four elements in the synset, the non-attitudinal synsets perform better with respect to the baseline. We speculate that this is due to the nature of the synsets discussed above: the attitudinal synsets are distributionally very skewed, and adding a very low probability element (to move from three to four elements in the synset) does not

make the task of the baseline noticeably harder, but does add extra noise for Edmonds.

5 Future research plan

The initial experiment provides some support for the hypothesis that closely related words that differ in sentiment amongst themselves can be predicted by context more easily than closely related words that do not differ in sentiment amongst themselves. Section 5.1 describes several additional experiments planned that will further test this hypothesis and Section 5.2 outlines a longer term research project based on sentiment analysis techniques aimed at acquiring sentiment differences between near-synonyms.

5.1 Short term plans

In this section, we outline several further tests of the original hypothesis that near-synonyms that differ in attitude are more amenable to corpus statistics techniques when choosing between candidate near-synonyms: Section 5.1.1 describes a methodology for improving the test set; Section 5.1.2 describes a methodology for improving the test words; and Section 5.1.3 describes using a later technique which draws on larger amounts of data.

5.1.1 Larger test set

As described in Section 4, our test set consisted of only 58 word sets, only 7 of which differed in attitude. A stronger version of this experiment, or of any related experiment, would rely on a larger number of wordsets. In this section, we describe a potential source of a larger test set, which could be used to repeat this first experiment, or as test data on any of the following experiments.

The *General Inquirer* wordset (Stone et al., 1966) is a lexicon of words tagged for various attributes. In particular, there are 1046 words tagged as'Pstv' (positive sentiment) and 1165 tagged as 'Ngtv' (negative sentiment). These words have been used both in the evaluation of sentiment analysis systems (Turney and Littman, 2003; Wilson et al., 2005) and for use in generation (Inkpen et al., 2006). Our use of it would be as a source of test words instead.

There are two possible ways that the *General Inquirer* wordset could provide us with a larger test set:

1. we could select sets of closely related words that appear in one word list (for example, *anger* and *fury*, which both appear in the 'Ngtv' list) as our test set; or
2. we could select sets of closely related words where there are members of both the 'Pstv' and 'Ngtv' lists in the set.

Approach 2 is closer to our present experiment, which concerns the hypothesis that words that *differ* in attitude can be predicted more effectively by their context than other closely related words. Approach 1 would test a different hypothesis: that words that merely *possess* some sentiment in their meaning, even if all words in the set have the same polarity of sentiment, can be predicted more effectively by their context. Approach 2 has a disadvantage, however: it is likely to yield a far smaller test set than approach 1.

5.1.2 Word frequency distributions

Informal inspection of the word sets used for the experiment in Section 3, together with the good performance of the most-frequent-word baseline on the 7 word sets with attitude differences compared to the performance of the baseline on the 51 word sets without attitude differences suggest that the distribution of the 7 word sets with attitude differences tended towards having one highly frequent word together with one or more much less frequent words, whereas the 51 word sets without attitude differences tend to have a less dominant most frequent word.

This tendency may affect the comparison of their performance in a number of ways:

1. there will be less evidence for the system to use to choose any of the less frequent words in the attitude word sets; and
2. the evidence that there is for choosing any of the less frequent words in the attitude word sets will be less reliable.

This may affect our ability to directly compare the attitude word sets with the other word sets. Therefore, we propose to investigate the distribution of the word sets in the corpora chosen for future experiments more thoroughly. One possible measure is to compare the entropy (Shannon, 1948) of the relative frequencies of words in the test sets. We can then attempt to choose test word sets with close entropy values.

5.1.3 Inkpen's methodology

Inkpen (2007) describes an alternative approach to the same task as Edmonds (1997) attempted to solve. Instead of estimating the likelihood of a particular near-synonym choice using *t*-scores acquired from bigrams within a window in the training data, she approximated two mutual information scores from the frequency counts of the words in a one-terabyte corpus as estimated by Clarke and Terra's (2003) Waterloo MultiText System: pointwise mutual information (Church and Hanks, 1991) and PMI-IR (Turney, 2001).

Although she did not test the specific question addressed in this paper of whether or not word sets that differ in attitude are more easily distinguished than those that do not, her overall results suggest that a larger amount of data may produce more coherent results, which may allow for a more effective comparison between the performance of word sets differing in attitude to those which do not. It also reduces the number of relatively arbitrary decisions in the Edmonds method, such as the choice of 2.0 and 3.0 for *t*-score and mutual information cut-offs, possibly allowing more effective comparisons with other methods.

5.2 Long term plans

Once the question of the differing performance of near-synonyms differing in sentiment is explored, the main thrust of our research will be in applying sentiment analysis approaches to acquiring the differences between near-synonyms and possibly paraphrases.

The goal is to be able to learn, say, the difference between *stingy* and *frugal* automatically from unstructured text. A possible methodology could be based on Turney and Littman (2003), where the semantic orientation (positive or negative) of a given word is measured by its association with positive words such as *excellent* compared to its association with negative words such as *nasty*, where association can be measured by statistical measures of word association such as Pointwise Mutual Information (Church and Hanks, 1991) or Latent Semantic Analysis (Deerwester et al., 1990). This methodology could be adapted easily to the problem of near synonyms in particular, with the following experimental questions:

1. since near synonyms occur in similar contexts, must the training data be different or larger to reliably distinguish negative words from their positive near-synonyms; and
2. is it possible to use the method to determine ``there is no significant polarity difference between these two near-synonyms'' as well as determining that one near-synonym is more negative or positive than the other.

If near-synonyms and their sentimental differences can be acquired from free text, we intend to test the effectiveness of using these differences as input to word choice decisions in natural language generation, as Inkpen and Hirst (2006) did with their database of near-synonym differences, acquired from lexicographic resources rather than free text.

Acknowledgements

This work was supported by the Australian Research Council's Discovery Grant DP0558852.

References

Satanjeev Banerjee and Ted Pedersen. 2003. The design, implementation and use of the Ngram Statistics Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, February.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple sequence alignment. In *HLT-NAACL 2003: Main Proceedings*, pages 16–23.

Kenneth Church and Patrick Hanks. 1991. Word association norms and mutual information, lexicography. *Computational Linguistics*, 16(1):22–29.

Charles L. A. Clarke 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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Scott Deerwester, Susan T. Dumais, George W. Furnas,

Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6): 391–407.

Philip Edmonds. 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 and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 507–509, July.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, May.

Diana Zaiu Inkpen and Graeme Hirst. 2004. Near-synonym choice in natural language generation. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing III (Selected Papers from RANLP 2003)*, pages 141–152. John Benjamins Publishing Company.

Diana Inkpen and Graeme Hirst. 2006. Building and using a lexical knowledge-base of near-synonym differences. *Computational Linguistics*, 32(2):223–262, June.

Diana Zaiu Inkpen, 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.

Diana Inkpen. 2007. A statistical model of nearsynonym choice. *ACM Transactions of Speech and Language* Processing, 4(1):1–17, January.

Irene Langkilde 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.

Irene Langkilde. 2000. Forest-based statistical sentence-generation. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics and the 6th Conference on Applied Natural Language Processing (NAACL-ANLP 2000)*, pages 170–177, Seattle, USA.

Bo Pang, 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 (EMNLP)*, pages 79–86.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October.

Sidney Siegel, Castellan, Jr., and N. John. 1988. *Non-parametric statistics for the behavioural sciences*. McGraw Hill, Boston.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Peter D. Turney and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Peter Turney. 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 and Germany.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 417–424, Philadelphia, Pennsylvania, USA, July.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *The Proceedings of the The Joint 21st International Conference on Computational Linguistics (COLING) and the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase level sentiment analysis. In *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada.