Corpus Statistics Approaches to Discriminating Among Near-Synonyms

Mary Gardiner
Centre for Language Technology
Macquarie University
gardiner@ics.mq.edu.au

Mark Dras
Centre for Language Technology
Macquarie University
madras@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, they can differ from each other subtly in both meaning and usage—consider the pair of near-synonyms *frugal* and *stingy*—and therefore choosing the appropriate near-synonym for a given context is not a trivial problem.

Early work on near-synonyms was that of Edmonds (1997). Edmonds reported an experiment attempting to predict which of a set of near-synonyms would be used in a given context using lexical co-occurrence networks. The conclusion of this work was that corpus statistics approaches did not appear to work well for this type of problem and led instead to the development of machine learning approaches over lexical resources such as *Choose the Right Word* (Hayakawa, 1994).

Our hypothesis is that some kind of corpus statistics approach may still be effective in some situations: particularly if the near-synonyms differ in sentiment from each other. Intuition based on work in sentiment analysis suggests that if the distribution of words embodying some characteristic of sentiment can predict the overall sentiment or attitude of a document, perhaps these same words can predict the choice of an individual 'attitudinal' near-synonym given its context, while this is not necessarily true for other types of near-synonym. This would again open up problems involving this type of near-synonym to corpus statistics methods. As a first step, then, we investigate whether attitudinal near-synonyms are more likely to be successfully predicted by a corpus statistics method than other types. In this paper we present a larger-scale experiment based on Edmonds (1997), and show that attitudinal near-synonyms can in fact be predicted more accurately using corpus statistics methods.

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. Ideally 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.

Edmonds (1997) examined a part of this problem: for 7 sets of near-synonyms trying to choose the most 'typical' among them for any given context based on co-occurrence statistics, where typicality is approximated by being able to predict the author's original word choice. This experiment suggested that context was able to predict an author's word choice to an extent. However, while the results improved on the baseline for most

Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pages 31–39.

cases in the small sample, the results were not considered sufficiently strong to pursue this approach; subsequent work (Inkpen and Hirst, 2006) used machine learning on resources authored by lexicographic experts, such as *Choose the Right Word* (Hayakawa, 1994), to acquire the differences between near-synonyms, although corpus statistics approaches have been used to choose between them (Inkpen et al., 2006).

Very recent work described by Inkpen (2007) has returned to the use of corpus statistics approaches and has discovered that with a sufficiently large amount of training data these approaches are more promising in general. However, neither Edmonds (1997), Edmonds (1999) nor Inkpen (2007) has examined their results in terms of which type of near-synonyms did best, in any case the sample size of 7 was too small to do this.

Differences in nuance between near-synonyms have been categorised in several ways with varying degrees of granularity:

- semantic or denotational variation (*mist* and *fog*) and stylistic or connotational variation (*stingy* and *frugal*) (DiMarco et al., 1993);
- collocational and syntactic variations (*die* and *pass away*), stylistic variations (*house* and *habitation*), expressive variations (*skinny* and *slim*) and denotational variations (*error*, *blunder* and *mistake*) (Edmonds and Hirst, 2002); and
- denotational variations (*invasion* and *incursion*), attitudinal variations (*placid* and *unimaginative*) and stylistic variations (*assistant* and *helper*) (Edmonds and Hirst, 2002; Inkpen and Hirst, 2006).

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 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.

While the performance of the Edmonds's approach in general is modest, and has factors which may worsen the results including not doing word sense disambiguation, we return to it in this paper in order to test whether the thrust of the approach—using corpus statistics approaches to distinguish between near-synonyms—shows signs of being particularly useful for discriminating among near-synonyms that differ in sentiment. Thus, in this paper we apply Edmonds's approach to a much larger sample of near-synonyms to test whether success varies according to near-synonym type. In Section 2 we outline the near-synonym prediction task. In Section 3 we describe the classification sub-task by which we obtained the data, including an annotation experiment to assess the validity of the classification. In Section 4 we describe the approach to near-synonym prediction, the details of the experiment and its results, along with a discussion. In Section 5 we conclude and present some ideas on what this work might lead on to.

2 Task Description

Edmonds (1997) describes an experiment that he designed to test whether or not co-occurrence statistics are sufficient to predict which word in a set of near-synonyms fills a *lexical gap*. He gives this example of asking the system to choose which of *error*, *mistake* or *oversight* fits into the gap in this sentence:

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

Edmonds performed this experiment with 7 sets of near-synonyms:

1. the adjectives *difficult*, *hard* and *tough*;
2. the nouns *error*, *mistake* and *oversight*;
3. the nouns *job*, *task* and *duty*;

4. the nouns *responsibility*, *commitment*, *obligation* and *burden*;

5. the nouns *material*, *stuff* and *substance*;

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

7. the verbs *settle* and *resolve*.

This small sample size does not allow for any analysis of whether there is any pattern to the different performances of each set and whether or not these differences in performance relate to any particular properties of those sets. Edmonds (1999) repeated the experiment using all of WordNet's synonym sets, but did not break down performance based on any properties of the synsets.

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 certain 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 correct part of speech in the 1989 Wall Street Journal:

$$\text{frequency}_{synset} = \sum_{i=1}^{n} \text{count}(w_i) \tag{2}$$

Synsets were then manually excluded from this set by the authors if they:

1. contained only one word (for example *commercial* with the meaning "of the kind or quality used in commerce");

2. 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*);

3. only occurred in a frozen idiom (for example *question* and *head* as in "the subject matter at issue");

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

5. 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 interchangable (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 the authors 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:

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 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.

Table 1: Break-down of categories assigned in the annotation experiment

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 2: Inter-annotator agreement and κ scores for 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[a]

[a] This figure for inter-annotator agreement 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

3.2 Results

Inter-annotator agreement for the annotation experiment is shown in Table 1 both:

- individually for annotations that the annotators felt were *definitely* correct and those that 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:

1. 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
2. the agreement between annotators on *only* the annotations they were definitely sure about, as per Wiebe and Mihalcea (2006).

In fact, we calculated two difference 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 et al., 1988) as recommended by Di Eugenio and Glass (2004). 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 2. Raw inter-annotator agreement is also shown.

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

4 Predicting the most typical word

4.1 Method

In this experiment we replicate the Edmonds (1997) experiment for a larger set of near-synonyms which have been categorised as differing from each other either in attitude or not in attitude, as described in section 3.

Each candidate token c, where a token is a part-of-speech tagged word, such as *(JJ arduous)* or *(NN fight)*, for the gap in sentence S is assigned a score, score(c, S), which is the sum of its significance score with each individual remaining token w in that sentence:

$$\text{score}(c, S) = \sum_{w \in S} \text{sig}(c, w) \quad (3)$$

The candidate c which maximises score(c, S) is chosen as the word fitting the lexical gap in sentence S. Where there is more than one candidate c with an equal maximum value of score(c, 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 sig(c, w) by connecting words in a *collocation network*. The principle is that if word w_0 co-occurs significantly with word w_1 which in turn co-occurs significantly with word w_2, then the presence of w_0 should weakly predict the appearance of w_2 even if they do not significantly co-occur in the training corpus. That is, he assumes that if, for example, *task* co-occurs significantly with *difficult*, and *difficult* co-occurs significantly 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 $\text{sig}(c, w)$ we take the shortest path of significance between the tokens c and w, which is either c, w where c and w significantly co-occur, or c, w_0, w where c and w both significantly co-occur with a third word, w_0.

Where tokens c and w significantly co-occur together, their significance score is their t-score (Church et al., 1991) as calculated by the Ngram Statistics Package (Banerjee and Pedersen, 2003):

$$\text{sig}(c, w) = t(c, w) \tag{4}$$

The t-score is calculated by comparing the likelihood of both words c and w occuring within a certain window of each other. The size of the window is either a 4 word window surrounding c, that is, c and w were found at most 2 words apart, or a 10 word window surrounding c, that is, c and w were found at most 5 words apart.

Where tokens c and w both significantly co-occur with token w_0, their significance score is a combination of their t-scores, with a bias factor devised by Edmonds to account for their weaker connection.

$$\text{sig}(c, w) = \frac{1}{8}(t(c, w_0) + \frac{t(w_0, w)}{2}) \tag{5}$$

If there is more than one candidate word w_0 co-occuring significantly with both c and w, the word w_0 is chosen so that the value of $\text{sig}(c, w)$ in equation 5 is maximised.

In the above, we have used "significantly co-occur" without definition. The test we are using is that from the description by Edmonds (1999) of the same experiment: any two words w_0 and w_1 significantly co-occur if their t-scores are greater than 2.0 and their mutual information score is greater than 3.0, as suggested by the observation of Church et al. (1991) that t-scores and mutual information scores emphasise different kinds of co-occurrence.

Input to the t-score and mutual information systems was the part-of-speech tagged 1989 *Wall Street Journal*. Stop words were those used by Edmonds, defined as any token with a raw frequency of over 800 in the corpus, and all punctuation, numbers, symbols and proper nouns. Per Edmonds we did not perform lemmatisation or word sense disambiguation.

As a baseline, also as per Edmonds (1997), we choose the most frequent element of the synset.

4.2 Test synsets and test sentences

Two types of test data were used:

1. lists of WordNet synsets divided into attitudinal and non-attitudinal synsets; and
2. sentences containing words from those synsets.

The lists of synsets is drawn from the annotation experiment described in Section 3. Synsets were chosen where both annotators are certain of their label, and where both annotators have the same label. As shown in Table 1, 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.

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

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

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

(7) *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*, *grueling*, *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*, *arduous*, *punishing* and *backbreaking* are the four most frequent words in the {*arduous*, *backbreaking*, *grueling*, *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.

The sentences for each experiment were selected from one of two corpora: the 1987 Wall Street Journal corpus or the 1988 Wall Street Journal corpus. (Recall that the 1989 Wall Street Journal was used as training data.)

Table 3: Test runs for the Edmonds experiment

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

For each n length synset, all sentences containing one of the words in that synset are found. For example, when the chosen 4-synset is {*hard*, *arduous*, *punishing* and *backbreaking*}, the selected sentences would include:

(8) With closing arguments completed in the first-ever racketeering trial of securities-firm officials, defense lawyers repaired to a downtown bar to celebrate the end of an *arduous* trial.

The word *arduous* is then removed, and the system is asked to predict which of {*hard*, *arduous*, *punishing* and *backbreaking*} goes into the gap using the method described in Section 4.

Table 3 shows the complete test set for the experiment. The same 58 synsets described above are used for all the sets of test data.

- the test sentences may be drawn from the complete 1987 Wall Street Journal, or the complete 1988 Wall Street Journal; and
- the training data is drawn from the 1989 Wall Street Journal, but may have consisted of bigrams drawn from 4 word windows around the target word (the sets called *4win-*), or from bigrams drawn from 10 word windows around the target word (the sets called *10win-*).

4.3 Results and Discussion

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.

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.

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 as is standard for comparing two proportions (Moore and McCabe, 2003). 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 attitu-

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

	Sentences containing test word		Baseline correctness (%)		Edmonds predictions (%)		Edmonds precision (%)	
Test set	Att.	Non-att.	Att.	Non-att.	Att.	Non-att.	Att.	Non-att.
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	86.7	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

dinal 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 Conclusion and Future Work

In this paper we have shown that human annotators can divide near-synonym sets into two classes, those that have members that differ from each other in attitude, and those which do not. We have also investigated whether these two different types of near-synonym sets perform differently when corpus statistics approaches are used to try and discriminate between them.

The data suggests that corpus statistics approaches may perform better on synsets whose members differ in attitude from one another than they do on synsets whose members do not differ in attitude, since the Edmonds methods improves upon the baseline more often when the differences are in attitude than when they are not in attitude.

This preliminary result suggests several further approaches. The first is increasing the size of the size of the word sets used to test any new methods beyond 7 attitudinal words. The second is applying the method of Inkpen (2007) to near-synonyms that differ in attitude rather than that of Edmonds (1997). Edmonds method was the only existing method to use for this comparison at the time of the experiments; but the recent work by Inkpen showed that a larger set of training data greatly improves over Edmonds's original results and suggests that corpus statistics measures are appropriate for discriminating between near-synonyms in general. Her method may confirm whether or near corpus statistics methods apply particularly well to near-synonyms differing in attitude.

Another possibility suggested by a preliminary inspection of the sets of near-synonyms is that attitudinal near-synonyms are distributed differently within the test data. Specifically, the most common word in any attitudinal near-synonym set is particularly common compared to the other words in that set when that word is compared the most common word in a typical non-attitudinal sets. This should be tested and, if confirmed, test data where the word frequencies in attitudinal near-synonym sets more closely match those in non-attitudinal sets may allow for more conclusive tests of the relative performance of corpus statistics measures on the two different types of near-synonym sets.

In the longer term, we hope to use this result to develop new corpus statistics methods to acquire and predict usage of attitudinal near-synonyms, drawing on methods from sentiment analysis.

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.

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, 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.

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

Barbara Di Eugenio and Michael Glass. 2004. The

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

	All words		Att. words		Non-att. words	
Test data	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

kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101.

Chrysanne DiMarco, 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.

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.

Philip Edmonds. 1999. *Semantic representations of near-synonyms for automatic lexical choice*. Ph.D. thesis, University of Toronto.

Philip Edmonds and Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

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

Samuel I. Hayakawa, editor. 1994. *Choose the Right Word*. Harper Collins Publishers. Second Edition, revised by Eugene Ehrlich.

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

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

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.

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.

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.

David S. Moore and George P. McCabe. 2003. *Introduction to the Practice of Statistics*. W. H. Freeman and Company, fourth edition.

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.

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

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.

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).