Support Vector Machines for Paraphrase Identification and Corpus Construction

Chris Brockett and William B. Dolan
Natural Language Processing Group
Microsoft Research
One Microsoft Way, Redmond, WA 98502, U.S.A.
{chrisbkt, billdol}@microsoft.com

Abstract

The lack of readily-available large corpora of aligned monolingual sentence pairs is a major obstacle to the development of Statistical Machine Translation-based paraphrase models. In this paper, we describe the use of annotated datasets and Support Vector Machines to induce larger monolingual paraphrase corpora from a comparable corpus of news clusters found on the World Wide Web. Features include: morphological variants; WordNet synonyms and hypernyms; log-likelihood-based word pairings dynamically obtained from baseline sentence alignments; and formal string features such as word-based edit distance. Use of this technique dramatically reduces the Alignment Error Rate of the extracted corpora over heuristic methods based on position of the sentences in the text.

1 Introduction

Paraphrase detection—the ability to determine whether or not two formally distinct strings are similar in meaning—is increasingly recognized as crucial to future applications in multiple fields including Information Retrieval, Question Answering, and Summarization. A growing body of recent research has focused on the problems of identifying and generating paraphrases, e.g., Barzilay & McKeown (2001), Lin & Pantel (2002), Shinyama et al, (2002), Barzilay & Lee (2003), and Pang et al. (2003). One promising approach extends standard Statistical Machine Translation (SMT) techniques (e.g., Brown et al., 1993; Och & Ney, 2000, 2003) to the problems of monolingual paraphrase identification and generation. Finch et al. (2004) have described several MT based paraphrase systems within the context of improving machine translation output. Quirk et al. (2004) describe an end-to-end paraphrase identification and generation system using GIZA++ (Och & Ney, 2003) and a monotone decoder to generate information-preserving paraphrases.

As with conventional SMT systems, SMT-based paraphrase systems require extensive monolingual parallel training corpora. However, while translation is a common human activity, resulting in large corpora of human-translated bilingual sentence pairs being relatively easy to obtain across multiple domains and language pairs, this is not the case in monolingual paraphrase, where naturally-occurring parallel data are hard to come by. The paucity of readily available monolingual parallel training corpora poses a formidable obstacle to the development of SMT-based paraphrase systems.

The present paper describes the extraction of parallel corpora from clustered news articles using annotated seed corpora and an SVM classifier, demonstrating that large parallel corpora can be induced by a classifier that includes morphological and synonymy features derived from both static and dynamic resources.

2 Background

Two broad approaches have dominated the literature on constructing paraphrase corpora. One

Edit Distance ($e \leq 12$)	San Jose Medical Center announced Wednesday that it would close its doors by Dec. 1, 2004.	San Jose Medical Center has announced that it will close its doors by Dec. 1, 2004.
First Two Sentences	The genome of the fungal pathogen that causes Sudden Oak Death has been sequenced by US scientists	Researchers announced Thursday they've completed the genetic blueprint of the blight-causing culprit responsible for Sudden Oak Death

Table 1. Paraphrase Examples Identified by Two Heuristics

approach utilizes multiple translations of a single source language text, where the source language text guarantees semantic equivalence in the target language texts (e.g., Barzilay & McKeown, 2001; Pang et al., 2003). Such corpora are of limited availability, however, since multiple translations of the same document are uncommon in non-literary domains.

The second strain of corpora construction involves mining paraphrase strings or sentences from news articles, with document clustering typically providing the topical coherence necessary to boost the likelihood that any two arbitrary sentences in the cluster are paraphrases. In this vein, Shinyama et al. (2002) use named entity anchors to extract paraphrases within a narrow domain. Barzilay & Lee (2003) employ Multiple Sequence Alignment (MSA, e.g., Durbin et al., 1998) to align strings extracted from closely related news articles. Although the MSA approach can produce dramatic results, it is chiefly effective in extracting highly templatic data, and appears to be of limited extensibility to broad domain application (Quirk et al. 2004).

Recent work by Dolan, et al. (2004) describes the construction of broad-domain corpora of aligned paraphrase pairs extracted from news-cluster data on the World Wide Web using two heuristic strategies: 1) pairing sentences based on a word-based edit distance heuristic; and 2) a naive text-feature-based heuristic in which the first two sentences of each article in a cluster are cross-matched with each other, their assumption being that the early sentences of a news article will tend to summarize the whole article and are thus likely to contain the same information as other early sentences of other articles in the cluster. The word-based edit distance heuristic yields pairs that are relatively clean but offer relatively minor rewrites in generation, especially when compared to the MSA model of (Barzilay & Lee, 2003). The text-based heuristic, on the other hand, results in a noisy "comparable" corpus: only 29.7% of sentence pairs are paraphrases, resulting in degraded performance on alignment metrics. This latter technique, however, does afford large numbers of pairings that are widely divergent at the string level; capturing these is of primary interest to paraphrase research. In this paper, we use an annotated corpus and an SVM classifier to refine the output of this second heuristic in an attempt to better identify sentence pairs containing richer paraphrase material, and minimize the noise generated by unwanted and irrelevant data.

3 Constructing a Classifier

3.1 Sequential Minimal Optimization

Although any of a number of machine learning algorithms, including Decision Trees, might be equally applicable here, Support Vector Machines (Vapnik, 1995) have been extensively used in text classification problems and with considerable success (Dumais 1998; Dumais et al., 1998; Joachims 2002). In particular, SVMs are known to be robust in the face of noisy training data. Since they permit solutions in high dimensional space, SVMs lend themselves readily to bulk inclusion of lexical features such as morphological and synonymy information.

For our SVM, we employed an off-the-shelf implementation of the Sequential Minimal Optimization (SMO) algorithm described in Platt (1999).[1] SMO offers the benefit of relatively short training times over very large feature sets, and in particular, appears well suited to handling the sparse features encountered in natural language classification tasks. SMO has been de-

[1] The pseudocode for SMO may be found in the appendix of Platt (1999)

	L12	F2	F3
Corpus size	253,725	51,933	235,061
Levenshtein edit distance	1 < e ≤ 12	e > 12	e > 12
Sentence range in article	All	First two	First three
Length	5 < n < 30	5 < n < 30	5 < n < 30
Length ratio	66%	50%	50%
Shared words	3	3	3

Table 2. Characteristics of L(evenshtein) 12, F(irst) 2, and F(irst) 3 Data

ployed a variety of text classification tasks (e.g., Dumais 1998; Dumais et al., 1998).

3.2 Datasets

To construct our corpus, we collected news articles from news clusters on the World Wide Web. A database of 13,127,938 candidate sentence pairs was assembled from 9,516,684 sentences in 32,408 clusters collected over a 2-year period, using simple heuristics to identify those sentence pairs that were most likely to be paraphrases, and thereby prune the overall search space.

```
Word-based Levenshtein edit distance
   of 1 < e ≤ 20; and a length ratio
   > 66%; OR

Both sentences in the first three
   sentences of each file; and length
   ratio > 50%.
```

From this database, we extracted three datasets. The extraction criteria, and characteristics of these datasets are given in Table 2. The data sets are labled L(evenshtein) 12, F(irst) 2 and F(irst) 3 reflecting their primary selection characteristics. The L12 dataset represents the best case achieved so far, with Alignment Error Rates beginning to approach those reported for alignment of closely parallel bilingual corpora. The F2 dataset was constructed from the first two sentences of the corpus on the same assumptions as those used in Dolan et al. (2004). To avoid conflating the two data types, however, sentence pairs with an edit distance of 12 or less were excluded. Since this resulted in a corpus that was significantly smaller than that desirable for exploring extraction techniques, we also created a third data set, F3 that consisted of the cross-pairings of the first three sentences of each article in each cluster, excluding those where the edit distance is $e \leq 12$.

3.3 Training Data

Our training data consisted of 10,000 sentence pairs extracted from randomly held-out clusters and hand-tagged by two annotators according to whether in their judgment (1 or 0) the sentence pairs constituted paraphrases. The annotators were presented with the sentences pairs in isolation, but were informed that they came from related document sets (clusters). A conservative interpretation of valid paraphrase was adopted: if one sentence was a superstring of the other, e.g., if a clause had no counterpart in the other sentence, the pair was counted as a non-paraphrase. Wherever the two annotators disagreed, the pairs were classed as non-paraphrases. The resultant data set contains 2968 positive and 7032 negative examples.

3.4 Features

Some 264,543 features, including overt lexical pairings, were in theory available to the classifier. In practice, however, the number of dimensions used typically fell to less than 1000 after the lowest frequency features are eliminated (see Table 4.) The main feature classes were:

String Similarity Features: All sentence pairs were assigned string-based features, including absolute and relative length in words, number of shared words, word-based edit distance, and lexical distance, as measured by converting the sentences into alphabetized strings of unique words and applying word based edit distance.

Morphological Variants: Another class of features was co-ocurrence of morphological variants in sentence pairs. Approximately 490,000 sentences in our primary datasets were stemmed using a rule-based stemmer, to yield a lexicon of 95,422 morphologically variant word pairs. Each word pair was treated as a feature. Examples are:

```
orbit|orbital
orbiter|orbiting
```

WordNet Lexical Mappings: Synonyms and hypernyms were extracted from WordNet,

(http://www.cogsci.princeton.edu/~wn/; Fellbaum, 1998), using the morphological variant lexicon from the 490,000 sentences as keywords. The theory here is that as additional paraphrase pairs are identified by the classifier, new information will "come along for the ride," thereby augmenting the range of paraphrases available to be learned. A lexicon of 314,924 word pairs of the following form created. Only those pairs identified as occurring in either training data or the corpus to be classified were included in the final classifier.

```
operation|procedure
operation|work
```

Word Association Pairs: To augment the above resources, we dynamically extracted from the L12 corpus a lexicon of 13001 possibly-synonymous word pairs using a log-likelihood algorithm described in Moore (2001) for machine translation. To minimize the damping effect of the overwhelming number of identical words, these were deleted from each sentence pair prior to processing; the algorithm was then run on the non-identical residue as if it were a bilingual parallel corpus.

To deploy this data in the SVM feature set, a cutoff was arbitrarily selected that yielded 13001 word pairs. Some exemplars (not found in WordNet) include:

```
straight|consecutive
vendors|suppliers
```

Fig. 1 shows the distribution of word pairings obtained by this method on the L12 corpus in comparison with WordNet. Examination of the top-ranked 1500 word pairs reveals that 46.53% are found in WordNet and of the remaining 53.47%, human judges rated 56% as good, yielding an overall "goodness score" of 76.47%. Judgments were by two independent raters. For the purposes of comparison, we automatically eliminated pairs containing trivial substring differences, e.g., spelling errors, British vs. American spellings, singular/plural alternations, and miscellaneous short abbreviations. All pairs on which the

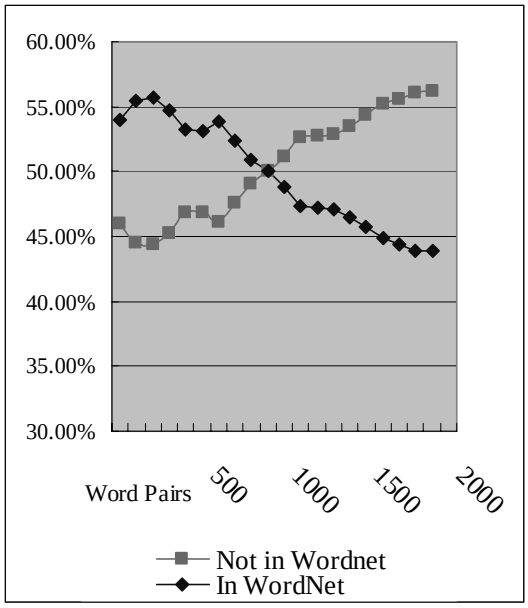

Fig. 1. WordNet Coverage in Word Association Output

raters disagreed were discarded. Also discarded were a large number of partial phrasal matches of the "reported|according" and "where|which" type, where part of a phrase ("according to", "in which") was missing. Although viewed in isolation these do not constitute valid synonym or hypernym pairs, the ability to identify these partial matchings is of central importance within an SMT-framework of paraphrase alignment and generation. These results suggest, among other things, that dynamically-generated lexical data of this kind might be useful in increasing the coverage of hand-built synonymy resources.

Composite Features: From each of the lexical feature classes, we derived a set of more abstract features that summarized the frequency with which each feature or class of features occurred in the training data, both independently, and in correlation with others. These had the effect of performing normalization for sentence length and other factors. Some examples are:

`No_of_List_2_Words` (i.e., the count of Wordnet matches)

	Corpus Size (pairs)	Precision	Recall	AER	Id AER	Non Id AER
L12	~254 K	87.42%	87.66%	**12.46%**	**11.57%**	**21.25%**
F2	~52 K	85.56%	83.31%	15.57%	13.19%	39.08%
F3	~235K	86.53%	81.57%	15.99%	14.24%	33.83%
10K Trained	~24 K	86.93%	87.24%	**12.92%**	**11.69%**	**24.70%**
MSR Trained	~50 K	86.76%	86.39%	13.42%	11.92%	28.31%

Table 3. Precision, Recall and Alignment Error Rates

`External_Matches_2_LED` (i.e,, the ratio of total lexical matches to Levenshtein edit distance.)

4 Evaluation

4.1 Methodology

Evaluation of paraphrase recognition within an SMT framework is highly problematic, since no technique or data set is standardly recognized. Barzilay & Lee (2003) and Quirk et al. (2004) use human evaluations of end-to-end generation, but these are not very useful here, since they add an additional layer of uncertainty into the evaluation, and depend to a significant extent on the quality and functionality of the decoder. Dolan & Brockett (2005) report extraction precision of 67% using a similar classifier, but with the explicit intention of creating a corpus that contained a significant number of naturally-occuring paraphrase-like negative examples.

Since our purpose in the present work is non-application specific corpus construction, we apply an automated technique that is widely used for reporting intermediate results in the SMT community, and is being extended in other fields such as summarization (Daumé and Marcu, forthcoming), namely word-level alignment using an off-the-shelf implementation of the SMT system GIZA++ (Och & Ney, 2003). Below, we use Alignment Error Rate (AER), which is indicative of how far the corpus is from providing a solution under a standard SMT tool. This allows the effective coverage of an extracted corpus to be evaluated efficiently, repeatedly against a single standard, and at little cost after the initial tagging. Further, if used as an objective function, the AER technique offers the prospect of using hillclimbing or other optimization techniques for non-application-specific corpus extraction.

To create the test set, two human annotators created a gold standard word alignment on held out data consisting of 1007 sentences pairs. Following the practice of Och & Ney (2000, 2003), the annotators each created an initial annotation, categorizing alignments as either SURE (necessary) or POSSIBLE (allowed, but not required). In the event of differences, annotators were asked to review their choices. First pass inter-rater agreement was 90.28%, climbing to 94.43% on the second pass. Finally we combined the annotations into a single gold standard as follows: if both annotators agreed that an alignment was SURE, it was tagged as SURE in the gold-standard; otherwise it was tagged as POSSIBLE.

To compute Precision, Recall, and Alignment Error Rate (AER), we adhere to the formulae listed in Och & Ney (2003). Let A be the set of alignments in the comparison, S be the set of SURE alignments in the gold standard, and P be the union of the SURE and POSSIBLE alignments in the gold standard:

$$\text{precision} = \frac{|A \cap P|}{|A|};\ \text{recall} = \frac{|A \cap S|}{|S|}$$

$$\text{AER} = \frac{|A \cap P| + |A \cap S|}{|A + S|}$$

4.2 Baselines

Evaluations were performed on the heuristically-derived L12, F2, and F3 datasets using the above formulation. Results are shown in Table 3. L12 represents the best case, followed respectively by F3 and F2. AERs were also computed separately for identical (Id) and non-identical (Non-Id) word mappings in order to be able to

	Dimensions	Non Id AER
All (fq > 4)	946	**24.70**
No Lexical Pairs	230	25.35
No Word Association	470	25.35
No WordNet	795	25.24
No Morphology	813	25.64

Table 4. Effect of Eliminating Feature Classes on 10K Training Set

drill down on the extent to which new non-identical mappings are being learned from the data. A high Id error rate can be considered indicative of noise in the data. The score that we are most interested in, however, is the Non-Id alignment error rate, which can be considered indicative of coverage as represented by the Giza++ alignment algorithm's ability to learn new mappings from the training data. It will be observed that the F3 dataset non-Id AER is smaller than that of the F2 dataset: it appears that more data is having the desired effect.

Following accepted SMT practice, we added a lexicon of identical word mappings to the training data, since Giza++ does not directly model word identity, and cannot easily capture the fact that many words in paraphrase sentence may translate as themselves. We did not add in word pairs derived from word association data or other supplementary resources that might help resolve matches between unlike but semantically similar words.

4.3 Training on the 10K Data

We trained an SVM on the 10 K training set employing 3-fold cross-validation on the training set itself. Validation errors were typically in the region of 16-17%. Linear kernels with default parameters (tolerance=1e-3; margin size computed automatically; error probability=0.5) were employed throughout. Applying the SVM to the F3 data, using 946 features encountered in the training data with frequency > 4, this classifier yielded a set of 24588 sentence pairs, which were then aligned using Giza++.

The alignment result is shown in Table 3. The "10K Trained" row represents the results of applying Giza++ to the data extracted by the SVM. Non-identical word AER, at 24.70%, shows a 36.9% reduction in the non-identical word AER over the F2 dataset (which is approximately double the size), and approximately 28% over the original F3 dataset. This represents a huge improvement in the quality of the data collected by using the SVM and is within striking distance of the score associated with the L12 best case. The difference is especially significant when it is considered that the newly constructed corpus is less than one-tenth the size of the best-case corpus. Table 5 shows sample extracted sentences.

To develop insights into the relative contributions of the different feature classes, we omitted some feature classes from several runs. The results were generally indistinguishable, except for non-Id AER, shown in Table 4, a fact that may be taken to indicate that string-based features such as edit distance still play a major role. Eliminating information about morphological alternations has the largest overall impact, producing a degradation of a 0.94 in on Non-Id AER. Of the three feature classes, removal of WordNet appears to have the least impact, showing the smallest change in Non-Id AER.

When the word association algorithm is applied to the extracted ~24K-sentence-pair set, degradation in word pair quality occurs significantly earlier than observed for the L12 data; after removing "trivial" matches, 22.63% of word pairs in the top ranked 800 were found in Wordnet, while 25.3% of the remainder were judged to be "good" matches. This is equivalent to an overall "goodness score" of 38.25%. The rapid degradation of goodness may be in part attributable to the smaller corpus size yielded by the classifier. Nevertheless, the model learns many valid new word pairs. Given enough data with which to bootstrap, it may be possible to do away with static resources such as Wordnet, and rely entirely on dynamically derived data.

4.4 Training on the MSR Training Set

By way of comparison, we also explored application of the SVM to the training data in the MSR Paraphrase corpus. For this purpose we used the 4076-sentence-pair "training" section of the MSR corpus, comprising 2753 positive and 1323 negative examples. The results at default parameter settings are given in Table 3, with respect to all features that were observed to occur with frequency greater than 4. Although the 49914 sentence pairs yielded by using the

Paraphrase (accepted)	young female chimps learn skills earlier , spend more time studying and tend to do better than young male chimpanzees - at least when it comes to catching termites .	young female chimpanzees are better students than males , at least when it comes to catching termites , according to a study of wild chimps in tanzania 's gombe national park .
	a %%number%% -year-old girl was arrested , handcuffed and taken into custody on charges of stealing a rabbit and a small amount of money from a neighbor 's home .	sheriff 's deputies in pasco county , fla. , this week handcuffed and questioned a %%number%% -year-old girl who was accused of stealing a rabbit and %%money%% from a neighbor 's home .
Non-Paraphrase (rejected)	roy moore , the chief justice of alabama , installed the two-ton sculpture in the rotunda of his courthouse in montgomery , and has refused to remove it .	the eight associate justices of alabama 's supreme court voted unanimously %%day%% to overrule moore and comply with u.s. district judge myron thompson 's order to remove the monument .

Table 5. Sample Pairs Extracted and Rejected by the SVM Trained on the 10K Corpus

MSR Paraphrase Corpus is nearly twice that of the 10K training set, AER performance is measurably degraded. Nevertheless, the MSR-trained corpus outperforms the similar-sized F12, yielding a reduction in Non-Id AER of a not insignificant 16%.

The fact that the MSR training data does not perform as well as the 10 K training set probably reflects its derivative nature, since it was originally constructed with data collected using the 10K training set, as described in Dolan & Brockett (2005). The performance of the MSR corpus is therefore skewed to reflect the biases inherent in its original training, and therefore exhibits the performance degradation commonly associated with bootstrapping. It is also a significantly smaller training set, with a higher proportion of negative examples than in typical in real world data. It will probably be necessary to augment the MSR training corpus with further negative examples before it can be utilized effectively for training classifiers.

5 Discussion and Future Work

These results show that it is possible to use machine learning techniques to induce a corpus of likely sentential paraphrase pairs whose alignment properties measured in terms of AER approach those of a much larger, more homogeneous dataset collected using a string-edit distance heuristic. This result supports the idea that an abstract notion of paraphrase can be captured in a high dimensional model.

Future work will revolve around optimizing classifiers for different domains, corpus types and training sets. It seems probable that the effect of the 10K training corpus can be greatly augmented by adding sentence pairs that have been aligned from multiple translations using the techniques described in, e.g., Barzilay & McKeown (2001) and Pang et al. (2003).

6 Conclusions

We have shown that supervised machine learning techniques such as SVMs can significantly expand available paraphrase corpora, and achieve a reduction of noise as measured by AER on non-identical words.

Although from the present research has focused on “ready-made” news clusters found on the web, nothing in this paper depends on the availability of such clusters. Given standard clustering techniques, the approach that we have described for inductive classifier learning should in principle be applicable to any flat corpus which contains multiple sentences expressing similar content. We expect also that the techniques described here could be extended to identify bilingual sentence pairs in comparable corpora, helping automate the construction of corpora for machine translation.

The ultimate test of paraphrase identification technologies lies in applications. These are likely to be in fields such as extractive multi-document summarization where paraphrase detection might eliminate sentences with comparable content and Question Answering, for both identifying sentence pairs with comparable content and generating unique new text. Such prac-

tical applications will only be possible once large corpora are available to permit the development of robust paraphrase models on the scale of the best SMT models. We believe that the corpus construction techniques that we have described here represent an important contribution to this goal.

Acknowledgements

We would like to thank Monica Corston-Oliver, Jeff Stevenson, Amy Muia and Margaret Salome of Butler Hill Group LLC for their assistance in annotating and evaluating our data. This paper has also benefited from feedback from several anonymous reviewers. All errors and omissions are our own.

References

Regina Barzilay and Katherine. R. McKeown. 2001. Extracting Paraphrases from a parallel corpus. In *Proceedings of the ACL/EACL*.

Regina Barzilay and Lillian Lee. 2003. Learning to Paraphrase; an unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL 2003*.

P. Brown, S. A. Della Pietra, V.J. Della Pietra and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation. *Computational Linguistics*, Vol. 19(2): 263-311.

Hal Daumé III and Daniel Marcu. (forthcoming) Induction of Word and Phrase Alignments for Automatic Document Summarization. To appear in *Computational Linguistics*.

William. B. Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. *Proceedings of COLING 2004*, Geneva, Switzerland.

William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of The Third International Workshop on Paraphrasing (IWP2005), Jeju, Republic of Korea*.

Susan Dumais. 1998. Using SVMs for Text Categorization. *IEEE Intelligent Systems*, Jul.-Aug. 1998: 21-23

Susan Dumais, John Platt, David Heckerman, Mehran Sahami. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management.*

Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge University Press.

Christiane Fellbaum (ed.). 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Andrew Finch, Taro Watanabe, Yasuhiro Akiba and Eiichiro Sumita. 2004. Paraphrasing as Machine Translation. *Journal of Natural Language Processing*, 11(5), pp 87-111.

Thorsten Joachims. 2002. *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms.* Kluwer Academic Publishers, Boston/Dordrecht/London.

Microsoft Research Paraphrase Corpus. http://research.microsoft.com/research/downloads/default.aspx

Robert C. Moore. 2001. Towards a Simple and Accurate Statistical Approach to Learning Translation Relationships among Words. In *Proceedings of the Workshop on Data-Driven Machine Translation*, ACL 2001.

Franz Joseph Och and H. Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the ACL*, Hong Kong, China, pp 440-447.

Franz Joseph Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29 (1): 19-52.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In *Proceedings of NAACL-HLT.*

John C. Platt. 1999. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In Bernhard Schölkopf, Christopher J. C. Burges and Alexander J. Smola (eds.). 1999. *Advances in Kernel Methods: Support Vector Learning*. The MIT Press, Cambridge, MA. 185-208.

Quirk, Chris, Chris Brockett, and William B. Dolan. 2004. Monolingual Machine Translation for Paraphrase Generation, In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 25-26 July 2004, Barcelona Spain, pp. 142-149.

Bernhard Schölkopf and Alexander J. Smola. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* The MIT Press, Cambridge, MA.

Y. Shinyama, S. Sekine and K. Sudo 2002. *Automatic Paraphrase Acquisition from News Articles.* In Proceedings of NAACL-HLT.

Vladimir N. Vapnik. 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.