Automatically Constructing a Corpus of Sentential Paraphrases

William B. Dolan and Chris Brockett
Natural Language Processing Group
Microsoft Research
Redmond, WA, 98052, USA
{billdol,chrisbkt}@microsoft.com

Abstract

An obstacle to research in automatic paraphrase identification and generation is the lack of large-scale, publicly-available labeled corpora of sentential paraphrases. This paper describes the creation of the recently-released Microsoft Research Paraphrase Corpus, which contains 5801 sentence pairs, each hand-labeled with a binary judgment as to whether the pair constitutes a paraphrase. The corpus was created using heuristic extraction techniques in conjunction with an SVM-based classifier to select likely sentence-level paraphrases from a large corpus of topic-clustered news data. These pairs were then submitted to human judges, who confirmed that 67% were in fact semantically equivalent. In addition to describing the corpus itself, we explore a number of issues that arose in defining guidelines for the human raters.

1 Introduction

The Microsoft Research Paraphrase Corpus (MSRP), available for download at http://research.microsoft.com/research/nlp/msr_paraphrase.htm, consists of 5801 pairs of sentences, each accompanied by a binary judgment indicating whether human raters considered the pair of sentences to be similar enough in meaning to be considered close paraphrases. This data has been published for the purpose of encouraging research in areas relating to paraphrase and sentential synonymy and inference, and to help establish a discourse on the proper construction of paraphrase corpora for training and evaluation. It is hoped that by releasing this corpus, we will stimulate the publication of similar corpora by others and help move the field toward adoption of a shared dataset that will permit useful comparisons of results across research efforts.

2 Motivation

The success of Statistical Machine Translation (SMT) has sparked a successful line of investigation that treats paraphrase acquisition and generation essentially as a monolingual machine translation problem (e.g., Barzilay & Lee, 2003; Pang et al., 2003; Quirk et al., 2004; Finch et al., 2004). However, a lack of standardly-accepted corpora on which to train and evaluate models is a major stumbling block to the successful application of SMT models or other machine learning algorithms to paraphrase tasks. Since paraphrase is not apparently a common "natural" task—under normal circumstances people do not attempt to create extended paraphrase texts—the field lacks a large readily identifiable dataset comparable to, for example, the Canadian Hansard corpus in SMT that can serve as a standard against which algorithms can be trained and evaluated.

What paraphrase data is currently available is usually too small to be viable for either training or testing, or exhibits narrow topic coverage, limiting its broad-domain applicability. One class of paraphrase data that is relatively widely available is multiple translations of sentences in a second language. These, however, tend to be rather restricted in their domain (e.g. the ATR English-Chinese paraphrase corpus, which con-

sists of translations of travel phrases (Zhang & Yamamoto, 2002)), are limited to short handcrafted predicates (e.g. the ATR Japanese-English corpus (Shirai, et al., 2002)), or exhibit quality problems stemming from insufficient command of the target language by the translators of the documents in question, e.g. the Linguistic Data Consortium's Multiple-Translation Chinese Corpus (Huang et al., 2002). Multiple translations of novels, such as those used in (Barzilay & McKeown, 2001) provide a relatively limited dataset to work with, and – since these usually involve works that are out of copyright – usually exhibit older styles of language that have little in common with modern language resources or application requirements.

Likewise, the data made available by (Barzilay & Lee, 2003: http://www.cs.cornell.edu/Info/Projects/NLP/statpar.html), while invaluable in understanding and evaluating their results, is too limited in size and domain coverage to serve as either training or test data.

Attempting to evaluate models of paraphrase acquisition and generation under limitations can thus be an exercise in frustration. Accordingly, we have tried to create a reasonably large corpus of naturally-occurring, non-handcrafted sentence pairs, along with accompanying human judgments, that can be used as a resource for training or testing purposes. Since the search space for identifying any two sentence pairs occurring "in the wild" is huge, and provides far too many negative examples for humans to wade through, clustered news articles were used to constrain the initial search space to data that was likely to yield paraphrase pairs.

3 Source Data

The Microsoft Research Paraphrase Corpus (MSRP) is distilled from a database of 13,127,938 sentence pairs, extracted from 9,516,684 sentences in 32,408 news clusters collected from the World Wide Web over a 2-year period, The methods and assumptions used in building this initial data set are discussed in Quirk et al. (2004) and Dolan et al. (2004). Two heuristics based on shared lexical properties and sentence position in the document were employed to construct the initial database:

Word-based Levenshtein edit distance of $1 < e \leq 20$; and a length ratio > 66%; *OR*

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

Within this initial dataset we were able to automatically identify the names of both authors and copyright holders of 61,618 articles.[1] Limiting ourselves only to sentences found in those articles, we further narrowed the range of candidate pairs using the following criteria:

The number of words in both sentences in words is $5 \geq n \leq 40$;

The two sentences shared at least three words in common;

The length of the shorter of the two sentences, in words, is at least 66.6% that of the longer; and

The two sentences had a bag-of-words lexical distance of $e \geq 8$ edits.

This enabled us extract a set of 49,375 initial candidate sentence pairs whose author was known, The purpose of these heuristics was two-fold: 1) to narrow the search space for subsequent application of the classifier algorithm and human evaluation, and 2) to ensure at least some diversity among the sentences. In particular, we sought to exclude the large number of sentence pairs whose differences might be attributable only to typographical errors, variance between British and American spellings, and minor editorial variations. Lexical distance was computed by constructing an alphabetized list of unique vocabulary items from each of the sentences and measuring the number of insertions and deletions. Note that the number of sentence pairs collected in this first pass was relatively small compared with the overall size of the dataset; the requirement of author identification significantly circumscribed the available dataset.

[1] Author identification was performed on the basis of pattern matching datelines and other textual information. We made a strong effort to ensure correct attribution.

4 Constructing a Classifier

4.1 Sequential Minimal Optimization

To extract candidate pairs from this ~49K list, we used a Support Vector Machine. (Vapnik, 1995), in this case an implementation of the Sequential Minimal Optimization (SMO) algorithm described in Platt (1999),[2] which has been shown to be useful in text classification tasks (Dumais 1998; Dumais et al., 1998).

4.2 Training Set

A separate set of 10,000 sentence pairs had previously been extracted from randomly held-out clusters and hand-tagged by two annotators according to whether the sentence pairs constituted paraphrases. This yielded a set of 2968 positive examples and 7032 negative examples. The sentences represented a random mixture of held out sentences; no attempt was made to match their characteristics to those of the candidate data set.

4.3 Classifiers

In the classifier we restricted the feature set to a small set of feature classes. The main classes are given below. More details can be found in Brockett and Dolan (2005).

String Similarity Features: Absolute and relative length in words, number of shared words, word-based edit distance, and bag-of-words-based lexical distance.

Morphological Variants: A morphological variant lexicon consisting of 95,422 word pairs was created using a hand-crafted stemmer. Each pair is then treated as a feature in the classifier.

WordNet Lexical Mappings: 314,924 word synonyms and hypernym pairs were extracted from WordNet, (Fellbaum, 1998; http://www.cogsci.princeton.edu/~wn/). Only pairs identified as occurring in either training data or the corpus to be classified were included in the final classifier.

Encarta Thesaurus: 125,054 word synonym pairs were extracted from the *Encarta Thesaurus* (Rooney, 2001).

Composite Features: Additional, more abstract features summarized the frequency with which each feature or class of features occurred in the training data, both independently, and in correlation with other features or feature classes.

4.4 Results of Applying the Classifier

Since our purpose was not to evaluate the potential effectiveness of the classifier itself, but to identify a reasonably large set of both positive and plausible "near-miss" negative examples, the classifier was applied with output probabilities deliberately skewed towards over-identification, i.e., towards Type 1 errors, assuming non-paraphrase (0) as null hypothesis. This yielded 20,574 pairs out the initial 49,375-pair data set, from which 5801 pairs were then further randomly selected for human assessment.

5 Human Evaluation

The 5801 sentences selected by the classifier as likely paraphrase pairs were examined by two independent human judges. Each judge was asked whether the two sentences could be considered "semantically equivalent". Disagreements were resolved by a 3rd judge, with the final binary judgment reflecting the majority vote.[3] After resolving differences between raters, 3900 (67%) of the original pairs were judged "semantically equivalent".

5.1 Semantic Divergence

In many instances, the two sentences judged "semantically equivalent" in fact diverge semantically to at least some degree. For instance, both judges considered the following two to be paraphrases:

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

[3] This annotation task was carried out by an independent company, the Butler Hill Group, LLC. Monica Corston-Oliver directed the effort, with Jeff Stevenson, Amy Muia, and David Rojas acting as raters.

Charles O. Prince, 53, was named as Mr. Weill's successor.

Mr. Weill's longtime confidant, Charles O. Prince, 53, was named as his successor.

If a full paraphrase relationship can be described as "bidirectional entailment", then the majority of the "equivalent" pairs in this dataset exhibit "mostly bidirectional entailments", with one sentence containing information that differs from or is not contained in the other. Our decision to adopt this relatively loose tagging criterion was ultimately a practical one: insisting on complete sets of bidirectional entailments would have limited the dataset to pairs of sentences that are practically identical at the string level, as in the following examples.

The euro rose above US$1.18, the highest price since its January 1999 launch.

The euro rose above $1.18 the highest level since its launch in January 1999.

However, without a carefully controlled study, there was little clear proof that the operation actually improves people's lives.

But without a carefully controlled study, there was little clear proof that the operation improves people's lives.

Such pairs are commonplace in the raw data, reflecting the tendency of news agencies to publish and republish the same articles, with editors introducing small and often inexplicable changes (is "however" really better than "but"?) along the way. The resulting alternations are useful sources of information about synonymy and local syntactic changes, but our goal was to produce a richer type of corpus; one that provides information about the large-scale alternations that typify complex paraphrases.[4]

[4] Recall that in an effort to focus on sentence pairs that are not simply trivial variants of some original single source, we restricted our original dataset by removing all pairs with a minimum word-based Levenshtein distance of ≥ 8.

5.2 Complex Alternations

Some sentence pairs in the news data capture complex and full paraphrase alternations:

Wynn paid $23.5 million for Renoir's "In the Roses (Madame Leon Clapisson)" at a Sotheby auction on Tuesday

Wynn nabbed Renoir's "In the Roses (Madame Leon Clapisson)" for $23.5 on Tuesday at Sotheby's

Far more frequently, however, interesting paraphrases in the data are accompanied by at least minor differences in content:

David Gest has sued his estranged wife Liza Minelli for %MONEY% million for beating him when she was drunk

Liza Minelli's estranged husband is taking her to court for %MONEY% million after saying she threw a lamp at him and beat him in drunken rages

It quickly became clear, that in order to collect significant numbers of sentential paraphrase pairs, our standards for what constitutes "semantic equivalence" would have to be relaxed.

5.3 Rater Instructions

Raters were told to use their best judgment in deciding whether 2 sentences, at a high level, "mean the same thing". Under our relatively loose definition of semantic equivalence, any 2 of the following sentences would have qualified as "paraphrases", despite obvious differences in information content:

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

Scientists have figured out the complete genetic code of a virulent pathogen that has killed tens

of thousands of California native oaks

The East Bay-based Joint Genome Institute said Thursday it has unraveled the genetic blueprint for the diseases that cause the sudden death of oak trees

Several classes of named entities were replaced by generic tags in sentences presented to the raters, so that "Tuesday" became %%DAY%%, "$10,000" became "%%MONEY%%, and so on. In the released version of the dataset, however, these placeholders were replaced by the original strings.

After a good deal of trial-and-error, some specific rating criteria were developed and included in a tagging specification. For the most part, though, the degree of mismatch allowed before the pair was judged "non-equivalent" was left to the discretion of the individual rater: did a particular set of asymmetries alter the meanings of the sentences so much that they could not be regarded as paraphrases? The following sentences, for example, were judged "not equivalent" despite some significant content overlap:

The Gerontology Research Group said Slough was born on %DATE%, making her %NUMBER% years old at the time of her death.

"[Mrs. Slough"] is the oldest living American as of the time she died, L. Stephen Coles, Executive Director of the Gerontology Research Group, said %DATE%.

The tagging task was ill-defined enough that we were surprised at how high inter-rater agreement was (averaging 84%). The Kappa score of 62 is good, but low enough to be indicative of the difficulty of the rating task. We believe that with more practice and discussion between raters, agreement on the task could be improved.

Interestingly, a series of experiments aimed at making the judging task more concrete resulted in uniformly degraded inter-rater agreement. Providing a checkbox to allow judges to specify that one sentence fully entailed another, for instance, left the raters frustrated, slowed down the tagging, and had a negative impact on agreement. Similarly, efforts to identify classes of syntactic alternations that would not count against an "equivalent" judgment resulted, in most cases, in a collapse in inter-rater agreement. After completing hundreds of judgments, the raters themselves were asked for suggestions as to what checkboxes or instructions might improve tagging speed and accuracy. In the end, few generalizations seemed useful in streamlining the task; each pair is sufficiently idiosyncratic that that common sense has to take precedence over formal guidelines.

In a few cases, firm tagging guidelines were found to be useful. One example was the treatment of pronominal and NP anaphora. Raters were instructed to treat anaphors and their full forms as equivalent, regardless of how great the disparity in length or lexical content between the two sentences. (Often these correspondences are extremely interesting, and in sufficient quantity would provide interesting fodder for learning models of anaphora.)

SCC argued that Lexmark was trying to shield itself from competition…

The company also argued that Lexmark was trying to squash competition…

But Secretary of State Colin Powell brushed off this possibility %%day%%.

Secretary of State Colin Powell last week ruled out a non-aggression treaty.

Note that many of the 33% of sentence pairs judged to be "not equivalent" still overlap significantly in information content and even wording. These pairs reflect a range of relationships, from pairs that are completely unrelated semantically, to those that are partially overlapping, to those that are almost-but-not-quite semantically equivalent.

6 Discussion

Given that MSRP reflects both the initial heuristics and the SVM methodology that was employed to identify paraphrase candidates for human evaluation, it is also limited by that technology. The 67% ratio of positive to negative judgments is a reasonably reliable indicator of the precision of our technique--though it should

be recalled that parameters were deliberately distorted to yield imprecise results that included positive and a large number of "near-miss" negatives. Coverage is hard to estimate reliably. we calculate that fewer than 30% of the pairs in a set of matched first-two sentences extracted from clustered news data, after application of simple heuristics, are paraphrases (Dolan et al., 2004). It seems reasonable to assume that the reduction to 10% seen in the initial data set still leaves many valid paraphrase pairs uncaptured in the corpus. The need to limit the corpus to those sentences for which authorship can be verified, and more specifically. to no more than a single sentence extracted from each article. further constrains the coverage in ways whose consequences are not yet known. In addition, the three-shared-words heuristic further guarantees that an entire class of paraphrases in which no words are shared in common have been excluded from the data. It has been observed that the mean lexical overlap in the corpus is a relatively high 0.7 (Weeds et al, 2005), suggesting that more lexically divergent examples will be needed. In these respects, as Wu (2005) points out, the corpus is far from distributionally neutral. This is a matter that we hope to remedy in the future, since in many ways this excluded set of pairs is the most interesting of all.

The above limitations, together with its relatively small size, perhaps make the MRSP inappropriate for direct use as a training corpus. We show separately that the results of training a classifier on the present corpus may be inferior to other training sets, though better than crude string or text-based heuristics (Brockett & Dolan, 2005). We expect that the utility of the corpus will stem primarily from its use as a tool for evaluating paraphrase recognition algorithms. It has already been applied in this way by Corley & Mihalcea (2005) and Wu (2005).

7 A Virtual Super Corpus?

Although larger than any other non-translation-based labeled paraphrase corpus currently publicly available, MSRP is tiny compared with the huge bilingual parallel corpora publicly available within the Machine Translation community, for example, the Canadian Hansards, the Hong Kong Parliamentary corpus, or the United Nations documents. It is improbable that we will ever encounter a "naturally occurring" paraphrase corpus on the scale of any of these bilingual corpora. Moreover, whatever extraction technique is employed to identify paraphrases in other kinds of data will be apt to reflect the implicit biases of the methodology employed.

Here we would like to put forward a proposal. The paraphrase research community might be able to construct a "virtual paraphrase corpus" that would be adequately large for both training and testing purposes and minimize selectional biases. This could be achieved in something like the following manner. Research groups could compile their own labeled paraphrase corpora, applying whatever learning techniques they choose to select their initial data. If enough interested groups were to release a sufficiently large number of reasonably-sized corpora, it might be possible to achieve some sort consensus, in a manner analogous to the division of the Penn Treebank into sections, whereby classifiers and other tools are conventionally trained on one subset of corpora, and tested against another subset. Though this would present issues of its own, it would obviate many of the problems of extraction bias inherent in automated extraction, and allow better cross comparison across systems.

8 Future Directions

For our part we plan to expand the MSRP, both by extending the number of sentence pairs, and also improving the balance of positive and negative examples. We anticipate using multiple classifiers to reduce inherent biases in candidate corpus selection, and with better author identification to ensure proper attribution, to be able to draw on a larger dataset for consideration by our judges.

In future releases we expect to make available more information about individual evaluator judgments. Burger & Ferro (2005) have suggested that this data may allow researchers greater freedom to construct models based on the judgments of specific judges or combinations of judges, permitting more fine-grained use of the corpus.

One further issue that we will also be attempting to address is the need to provide a better metric for corpus coverage and quality. Until reliable metrics can be established for end-to-

end paraphrase tasks—these will probably need to be application specific—the Alignment Error Rate strategy that was successfully applied in early development of machine translation systems (Och & Ney, 2000, 2003) offers a useful intermediate representation of the coverage and precision of a corpus and extraction techniques. Though fullscale reliability studies have yet to be performed, the AER technique is already finding application in other fields such as summarization (Daumé & Marcu, forthcoming). We expect to be able to provide a reasonably large corpus of word-aligned paraphrase sentences in the near future that we hope will serve as some sort of standard by which corpus extraction techniques can be measured and compared in a uniform fashion.

One other path that we are concurrently exploring is collection and validation of paraphrase data by volunteers on the web. Some initial efforts using game formats for elicitation are presented in Chklovski (2005) and Brockett & Dolan (2005). It is our hope that web volunteers will prove a useful source of colloquial paraphrases of written text, and–if paraphrase identification can be effectively embedded in the game–of paraphrase judgments.

9 Conclusion

We have used heuristic techniques and a classifier to automatically create a corpus of 5801 "naturally occurring" (non-constructed) sentence pairs, labeled according to whether, in the judgment of our evaluators, the sentences "mean the same thing" or not. To our knowledge, MSRP constitutes the largest currently-available broad-domain corpus of paraphrase pairs that does not have its origins in translations from another language. We hope that others will utilize it, find it useful, and provide feedback when it is not.

The methodology that we have described for extracting this corpus is readily adaptable by others, and is not limited to news clusters, but can be readily extended to any flat corpus containing a large number of semantically similar sentences on which topic-based document clustering is possible. We have shown that by allowing a statistical learning algorithm to constrain the search space, it is possible to identify a manageable-sized candidate corpus on the basis of which human judges can label sentence pairs for paraphrase content quickly and in a cost effective manner. We hope that others will follow our example.

Acknowledgements

We would like to thank Monica Corston-Oliver, Jeff Stevenson, Amy Muia and David Rojas of Butler Hill Group LLC for their assistance in annotating the Microsoft Research Paraphrase Corpus and in preparing the seed data used for training. 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*.

Chris Brockett and William B. Dolan. 2005. Support Vector Machines for Paraphrase Identification and Corpus Construction. In *Proceedings of The Third International Workshop on Paraphrasing (IWP2005)*, Jeju, Republic of Korea.

Chris Brockett and William B. Dolan. 2005. Echo Chamber: A Game for Eliciting a Colloquial Paraphrase Corpus. *AAAI 2005 Spring Symposium, Knowledge Collection from Volunteer Contributors (KCVC05).* Stanford, CA. March 21-23, 2005.

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.

John Burger and Lisa Ferro. 2005. Generating an Entailment Corpus from News Headlines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. pp 49-54.

Timothy Chklovski. 2005 1001 Paraphrases: Incenting Responsible Contributions in Collecting Paraphrases from Volunteers. *AAAI 2005 Spring Symposium, Knowledge Collection from Volunteer Contributors* (KCVC05). Stanford, CA. March 21-23, 2005.

Courtney Courley and Rada Mihalcea. 2005. Measuring the Semantic Similarity of Texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. Pp 13-18.

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.

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.*

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.

Pascale Fung and Percy Cheung. 2004. Multi-level Bootstrapping for Extracting Parallel Sentences from a Quasi-Comparable Corpus. In *Proceedings of Coling 2004*, 1051-1057.

Shudong Huang, David Graff, and George Doddington (eds.) 2002. *Multiple-Translation Chinese Corpus.* Linguistic Data Consortium.

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

V. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physice-Doklady 10: 707-710.

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

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.

Chris Quirk, 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.

Kathy Rooney (ed.) 2001. *Encarta Thesaurus.* Bloomsbury Publishing.

Satoshi Shirai, Kazuhide Yamamoto, Francis Bond & Hozumi Tanaka. 2002. Towards a thesaurus of predicates. In Proceedings of LREC 2002 (Third International Conference on Language Resources and Evaluation), (May 29-31, 2002). Vol.6, pp. 1965-1972.

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

Julie Weeds, David Weir and Bill Keller. 2005. The Distributional Similarity of Subparses. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. pp 7-12.

Dekai Wu. 2005. Recognizing Paraphrases and Textual Entailment using Inversion Transduction Grammars. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. Pp 25-30.

Yujie Zhang and Kazuhide Yamamoto. 2002 Paraphrasing of Chinese Utterances. Proceedings of Coling 2002, pp.1163-1169.