Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources

Bill DOLAN, Chris QUIRK, and Chris BROCKETT
Natural Language Processing Group, Microsoft Research
One Microsoft Way
Redmond, WA 90852, USA
{billdol,chrisq,chrisbkt}@microsoft.com

Abstract

We investigate unsupervised techniques for acquiring monolingual sentence-level paraphrases from a corpus of temporally and topically clustered news articles collected from thousands of web-based news sources. Two techniques are employed: (1) simple string edit distance, and (2) a heuristic strategy that pairs initial (presumably summary) sentences from different news stories in the same cluster. We evaluate both datasets using a word alignment algorithm and a metric borrowed from machine translation. Results show that edit distance data is cleaner and more easily-aligned than the heuristic data, with an overall alignment error rate (AER) of 11.58% on a similarly-extracted test set. On test data extracted by the heuristic strategy, however, performance of the two training sets is similar, with AERs of 13.2% and 14.7% respectively. Analysis of 100 pairs of sentences from each set reveals that the edit distance data lacks many of the complex lexical and syntactic alternations that characterize monolingual paraphrase. The summary sentences, while less readily alignable, retain more of the non-trivial alternations that are of greatest interest learning paraphrase relationships.

1 Introduction

The importance of learning to manipulate monolingual paraphrase relationships for applications like summarization, search, and dialog has been highlighted by a number of recent efforts (Barzilay & McKeown 2001; Shinyama et al. 2002; Lee & Barzilay 2003; Lin & Pantel 2001). While several different learning methods have been applied to this problem, all share a need for large amounts of data in the form of pairs or sets of strings that are likely to exhibit lexical and/or structural paraphrase alternations. One approach[1] that has been successfully used is edit distance, a measure of similarity between strings. The assumption is that strings separated by a small edit distance will tend to be similar in meaning:

The leading **indicators measure** *the economy…*
The leading **index measures** *the economy….*

Lee & Barzilay (2003), for example, use Multi-Sequence Alignment (MSA) to build a corpus of paraphrases involving terrorist acts. Their goal is to extract sentential templates that can be used in high-precision generation of paraphrase alternations within a limited domain.

Our goal here is rather different: our interest lies in constructing a monolingual broad-domain corpus of pairwise aligned sentences. Such data would be amenable to conventional statistical machine translation (SMT) techniques (e.g., those discussed in Och & Ney 2003).[2] In what follows we compare two strategies for unsupervised construction of such a corpus, one employing string similarity and the other associating sentences that may overlap very little at the string level. We measure the relative utility of the two derived monolingual corpora in the context of word alignment techniques developed originally for bilingual text.

We show that although the edit distance corpus is well-suited as training data for the alignment algorithms currently used in SMT, it is an incomplete source of information about paraphrase relations, which exhibit many of the characteristics of comparable bilingual corpora or free translations. Many of the more complex alternations that characterize monolingual paraphrase, such as large-scale lexical alternations and constituent reorderings, are not readily

[1] An alternative approach involves identifying anchor points--pairs of words linked in a known way--and collecting the strings that intervene. (Shinyama, et al. 2002; Lin & Pantel 2001). Since our interest is in learning sentence level paraphrases, including major constituent reorganizations, we do not address this approach here.

[2] Barzilay & McKeown (2001) consider the possibility of using SMT machinery, but reject the idea because of the noisy, comparable nature of their dataset.

captured by edit distance techniques, which conflate semantic similarity with formal similarity. We conclude that paraphrase research would benefit by identifying richer data sources and developing appropriate learning techniques.

2 Data/Methodology

Our two paraphrase datasets are distilled from a corpus of news articles gathered from thousands of news sources over an extended period. While the idea of exploiting multiple news reports for paraphrase acquisition is not new, previous efforts (for example, Shinyama et al. 2002; Barzilay and Lee 2003) have been restricted to at most two news sources. Our work represents what we believe to be the first attempt to exploit the explosion of news coverage on the Web, where a single event can generate scores or hundreds of different articles within a brief period of time. Some of these articles represent minor rewrites of an original AP or Reuters story, while others represent truly distinct descriptions of the same basic facts. The massive redundancy of information conveyed with widely varying surface strings is a resource begging to be exploited.

Figure 1 shows the flow of our data collection process. We begin with sets of pre-clustered URLs which point to news articles on the Web, representing thousands of different news sources. The clustering algorithm takes into account the full text of each news article, in addition to temporal cues, to produce a set of topically and temporally related articles. Our method is believed to be independent of the specific clustering technology used. The story text is isolated from a sea of advertisements and other miscellaneous text through use of a supervised HMM.

Altogether we collected 11,162 clusters in an 8-month period, assembling 177,095 articles with an average of 15.8 articles per cluster. The clusters are generally coherent in topic and focus. Discrete events like disasters, business announcements, and deaths tend to yield tightly focused clusters, while ongoing stories like the SARS crisis tend to produce less focused clusters. While exact duplicate articles are filtered out of the clusters, many slightly-rewritten variants remain.

2.1 Extracting Sentential Paraphrases

Two separate techniques were employed to extract likely pairs of sentential paraphrases from these clusters. The first used string edit distance, counting the number of lexical deletions and insertions needed to transform one string into another. The second relied on a discourse-based heuristic, specific to the news genre, to identify likely paraphrase pairs even when they have little superficial similarity.

3 Levenshtein Distance

A simple edit distance metric (Levenshtein 1966) was used to identify pairs of sentences within a cluster that are similar at the string level. First, each sentence was normalized to lower case and paired with every other sentence in the cluster. Pairings that were identical or differing only by punctuation were rejected, as were those where the shorter sentence in the pair was less than two thirds the length of the longer, this latter constraint in effect placing an upper bound on edit distance relative to the length of the sentence. Pairs that had been seen before in either order were also rejected. Filtered in this way, our dataset yields 139K non-identical sentence pairs at a Levenshtein distance of $n \leq 12$. [3] Mean Levenshtein distance was 5.17, and mean sentence length was 18.6 words. We will refer to this dataset as L12.

3.1.1 First sentences

The second extraction technique was specifically intended to capture paraphrases which might contain very different sets of content words, word order, and so on. Such pairs are typically used to illustrate the phenomenon of paraphrase, but precisely because their surface dissimilarity renders automatic discovery difficult, they have generally not been the focus of previous computational approaches.

In order to automatically identify sentence pairs of this type, we have attempted to take advantage of some of the unique characteristics of the dataset. The topical clustering is sufficiently precise to ensure that, in general, articles in the same cluster overlap significantly in overall semantic content. Even so, any arbitrary pair of sentences from different articles within a cluster is unlikely to exhibit a paraphrase relationship:

The Phi-X174 genome is short and compact.
This is a robust new step that allows us to make much larger pieces.

To isolate just those sentence pairs that represent likely paraphrases without requiring significant string similarity, we exploited a common journalistic convention: the first sentence or two of

[3] A maximum Levenshtein distance of 12 was selected for the purposes of this paper on the basis of experiments with corpora extracted at various edit distances.

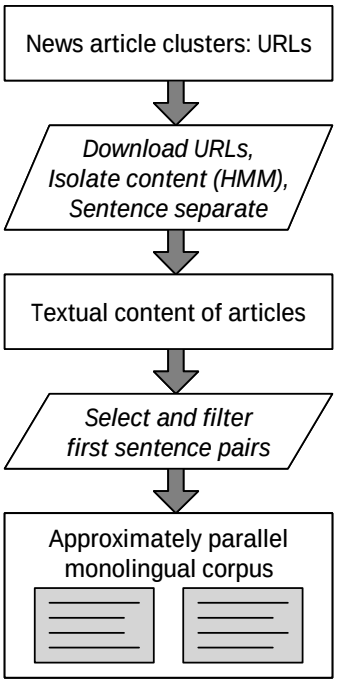

Figure 1. Data collection

a newspaper article typically summarize its content. One might reasonably expect, therefore, that initial sentences from one article in a cluster will be paraphrases of the initial sentences in other articles in that cluster. This heuristic turns out to be a powerful one, often correctly associating sentences that are very different at the string level:

In only 14 days, US researchers have created an artificial bacteria-eating virus from synthetic genes.
An artificial bacteria-eating virus has been made from synthetic genes in the record time of just two weeks.

Also consider the following example, in which related words are obscured by different parts of speech:

Chosun Ilbo, one of South Korea's leading newspapers, said North Korea had **finished developing** *a new ballistic missile last year and was* **planning to deploy it**.
The Chosun Ilbo said **development** *of the new missile, with a range of up to %%number%% kilometres (%%number%% miles),* **had been completed and deployment was imminent**.

A corpus was produced by extracting the first two sentences of each article, then pairing these across documents within each cluster. We will refer to this collection as the F2 corpus. The combination of the first-two sentences heuristic plus topical article clusters allows us to take advantage of meta-information implicit in our corpus, since clustering exploits lexical information from the entire document, not just the few sentences that are our focus. The assumption that two first sentences are semantically related is thus based in part on linguistic information that is external to the sentences themselves.

Sometimes, however, the strategy of pairing sentences based on their cluster and position goes astray. This would lead us to posit a paraphrase relationship where there is none:

Terence Hope should have spent most of yesterday in hospital performing brain surgery.
A leading brain surgeon has been suspended from work following a dispute over a bowl of soup.

To prevent too high an incidence of unrelated sentences, one string-based heuristic filter was found useful: a pair is discarded if the sentences do not share at least 3 words of 4+ characters. This constraint succeeds in filtering out many unrelated pairs, although it can sometimes be too restrictive, excluding completely legitimate paraphrases:

There was no chance it would endanger our planet, astronomers said.
NASA emphasized that there was never danger of a collision.

An additional filter ensured that the word count of the shorter sentence is at least one-half that of the longer sentence. Given the relatively long sentences in our corpus (average length 18.6 words), these filters allowed us to maintain a degree of semantic relatedness between sentences. Accordingly, the dataset encompasses many paraphrases that would have been excluded under a more stringent edit-distance threshold, for example, the following non-paraphrase pair that contain an element of paraphrase:

A staggering %%number%% million Americans **have been victims of identity theft** *in the last five years , according to federal trade commission survey out this week.*
In the last year alone, %%number%% million people **have had their identity purloined**.

Nevertheless, even after filtering in these ways, a significant amount of unfiltered noise remains in the F2 corpus, which consisted of 214K sentence pairs. Out of a sample of 448 held-out sentence pairs, 118 (26.3%) were rated by two independent human evaluators as sentence-level paraphrases, while 151 (33.7%) were rated as partial paraphrases. The remaining ~40% were assessed as

unrelated.[4] Thus, although the F2 data set is nominally larger than the L12 data set, when the noise factor is taken into account, the actual number of full paraphrase sentences in this data set is estimated to be in the region of 56K sentences, with a further estimated 72K sentences containing some paraphrase material that might be a potential source of alignment.

Some of these relations captured in this data can be complex. The following pair, for example, would be unlikely to pass muster on edit distance grounds, but nonetheless contains an inversion of deep semantic roles, employing different lexical items.

The Hartford Courant reported %%day%% that Tony Bryant said **two friends were the killers**.
A lawyer for Skakel says there is a claim that **the murder was carried out by two friends** *of one of Skakel's school classmates, Tony Bryan.*

The F2 data also retains pairs like the following that involve both high-level semantic alternations and long distance dependencies:

Two men who **robbed** *a jeweller's shop* **to raise funds for** *the Bali bombings were each jailed for %%number%% years by Indonesian courts today.*
An Indonesian court today sentenced two men to %%number%% years in prison **for helping finance** *last year's terrorist bombings in Bali* **by robbing** *a jewelry store.*

These examples do not by any means exhaust the inventory of complex paraphrase types that are commonly encountered in the F2 data. We encounter, among other things, polarity alternations, including those involving long-distance dependencies, and a variety of distributed paraphrases, with alignments spanning widely separated elements.

3.2 Word Error Alignment Rate

An objective scoring function was needed to compare the relative success of the two data collection strategies sketched in 2.1.1 and 2.1.2. Which technique produces more data? Are the types of data significantly different in character or utility? In order to address such questions, we used word Alignment Error Rate (AER), a metric borrowed from the field of statistical machine translation (Och & Ney 2003). AER measures how accurately an automatic algorithm can align words in corpus of parallel sentence pairs, with a human-tagged corpus of alignments serving as the gold standard. Paraphrase data is of course monolingual, but otherwise the task is very similar to the MT alignment problem, posing the same issues with one-to-many, many-to-many, and one/many-to-null word mappings. Our a priori assumption was that the lower the AER for a corpus, the more likely it would be to yield learnable information about paraphrase alternations.

We closely followed the evaluation standards established in Melamed (2001) and Och & Ney (2000, 2003). Following Och & Ney's methodology, two annotators each created an initial annotation for each dataset, subcategorizing alignments as either SURE (necessary) or POSSIBLE (allowed, but not required). Differences were then highlighted and the annotators were asked to review these cases. Finally we combined the two annotations into a single gold standard in the following manner: if both annotators agreed that an alignment should be SURE, then the alignment was marked as sure in the gold-standard; otherwise the alignment was marked as POSSIBLE.

To compute Precision, Recall, and Alignment Error Rate (AER) for the twin datasets, we used exactly 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. Then we have:

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

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

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

We held out a set of news clusters from our training data and randomly extracted two sets of sentence pairs for blind evaluation. The first is a set of 250 sentence pairs extracted on the basis of an edit distance of $5 \leq n \leq 20$, arbitrarily chosen to allow a range of reasonably divergent candidate pairs. These sentence pairs were checked by an independent human evaluator to ensure that they contained paraphrases before they were tagged for alignments. The second set comprised 116 sentence pairs randomly selected from the set of first-two sentence pairs. These were likewise hand-vetted by independent human evaluators. After an initial training pass and refinement of the linking

[4] This contrasts with 16.7% pairs assessed as unrelated in a 10,000 pair sampling of the L12 data.

Training Data Type:	L12	F2	L12	F2
Test Data Type:	**250 Edit Dist**	**250 Edit Dist**	**116 F2 Heuristic**	**116 F2 Heuristic**
Precision	87.46%	86.44%	85.07%	84.16%
Recall	89.52%	82.64%	88.70%	86.55%
AER	**11.58%**	**15.41%**	**13.24%**	**14.71%**
Identical word precision	89.36%	88.79%	92.92%	93.41%
Identical word recall	89.50%	83.10%	93.49%	92.47%
Identical word AER	**10.57%**	**14.14%**	**6.80%**	**7.06%**
Non-Identical word precision	76.99%	71.86%	60.54%	53.69%
Non-Identical word recall	90.22%	69.57%	59.50%	50.41%
Non-Identical word AER	**20.88%**	**28.57%**	**39.81%**	**47.46%**

Table 1. Precision, recall, and alignment error rates (AER) for F2 and L12

specification, interrater agreement measured in terms of AER[5] was 93.1% for the edit distance test set versus 83.7% for the F2 test set, suggestive of the greater variability in the latter data set.

3.3 Data Alignment

Each corpus was used as input to the word alignment algorithms available in Giza++ (Och & Ney 2000). Giza++ is a freely available implementation of IBM Models 1-5 (Brown et al. 1993) and the HMM alignment (Vogel et al. 1996), along with various improvements and modifications motivated by experimentation by Och & Ney (2000). Giza++ accepts as input a corpus of sentence pairs and produces as output a Viterbi alignment of that corpus as well as the parameters for the model that produced those alignments.

While these models have proven effective at the word alignment task (Mihalcea & Pedersen 2003), there are significant practical limitations in their output. Most fundamentally, all alignments have either zero or one connection to each target word. Hence they are unable to produce the many-to-many alignments required to identify correspondences with idioms and other phrasal chunks.

To mitigate this limitation on final mappings, we follow the approach of Och (2000): we align once in the forward direction and again in the backward direction. These alignments can subsequently be recombined in a variety of ways, such as union to maximize recall or intersection to maximize precision. Och also documents a method for heuristically recombining the unidirectional alignments intended to balance precision and recall. In our experience, many alignment errors are present in one side but not the other, hence this recombination also serves to filter noise from the process.

4 Evaluation

Table 1 shows the results of training translation models on data extracted by both methods and then tested on the blind data. The best overall performance, irrespective of test data type, is achieved by the L12 training set, with an 11.58% overall AER on the 250 sentence pair edit distance test set (20.88% AER for non-identical words). The F2 training data is probably too sparse and, with 40% unrelated sentence pairs, too noisy to achieve equally good results; nevertheless the gap between the results for the two training data types is dramatically narrower on the F2 test data. The nearly comparable numbers for the two training data sets, at 13.2% and 14.7% respectively, suggest that the L12 training corpus provides no substantive advantage over the F2 data when tested on the more complex test data. This is particularly striking given the noise inherent in the F2 training data.

5 Analysis/Discussion

To explore some of the differences between the training sets, we hand-examined a random sample of sentence pairs from each corpus type. The most common paraphrase alternations that we observed fell into the following broad categories:

- **Elaboration**: Sentence pairs can differ in total information content, with an added word, phrase or clause in one sentence that has no

[5] The formula for AER given here and in Och & Ney (2003) is intended to compare an automatic alignment against a gold standard alignment. However, when comparing one human against another, both comparison and reference distinguish between SURE and POSSIBLE links. Because the AER is asymmetric (though each direction differs by less than 5%), we have presented the average of the directional AERs.

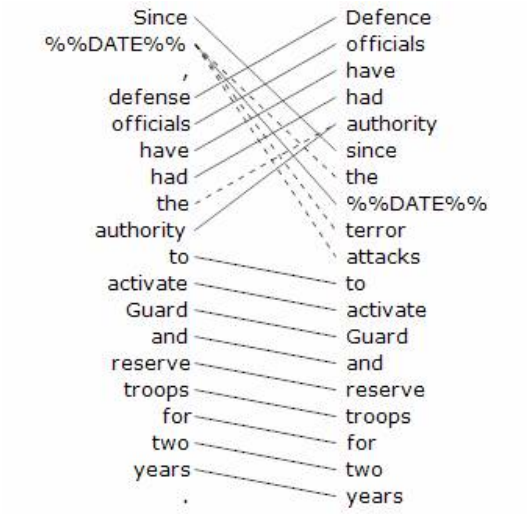

Figure 2. Sample human-aligned paraphrase

counterpart in the other (e.g. *the NASDAQ / the tech-heavy NASDAQ*).

- **Phrasal**: An entire group of words in one sentence alternates with one word or a phrase in the other. Some are non-compositional idioms (*has pulled the plug on* / *is dropping plans for*); others involve different phrasing (*electronically / in electronic form, more than a million people / a massive crowd*).
- **Spelling**: British/American sources systematically differ in spellings of common words (*colour* / *color*); other variants also appear (*email / e-mail*).
- **Synonymy**: Sentence pairs differ only in one or two words (e.g. *charges* / *accusations*), suggesting an editor's hand in modifying a single source sentence.
- **Anaphora**: A full NP in one sentence corresponds to an anaphor in the other (*Prime Minister Blair* / *He*). Cases of NP anaphora (*ISS* / *the Atlanta-based security company*) are also common in the data, but in quantifying paraphrase types we restricted our attention to the simpler case of pronominal anaphora.
- **Reordering**: Words, phrases, or entire constituents occur in different order in two related sentences, either because of major syntactic differences (e.g. topicalization, voice alternations) or more local pragmatic choices (e.g. adverb or prepositional phrase placement).

These categories do not cover all possible alternations between pairs of paraphrased sentences; moreover, categories often overlap in the same sequence of words. It is common, for example, to find instances of clausal Reordering combined with Synonymy.

	L12	**F2**
Elaboration	**0.83**	**1.3**
Phrasal	**0.14**	**0.69**
Spelling	0.12	0.01
Synonym	0.18	0.25
Anaphora	0.1	0.13
Reordering	**0.02**	**0.41**

Table 2. Mean number of instances of paraphrase phenomena per sentence

Figure 2 shows a hand-aligned paraphrase pair taken from the F2 data. This pair displays one Spelling alternation (*defence* / *defense*), one Reordering (position of the "since" phrase), and one example of Elaboration (*terror attacks* occurs in only one sentence).

To quantify the differences between L12 and F2, we randomly chose 100 sentence pairs from each dataset and counted the number of times each phenomenon was encountered. A given sentence pair might exhibit multiple instances of a single phenomenon, such as two phrasal paraphrase changes or two synonym replacements. In this case all instances were counted. Lower-frequency changes that fell outside of the above categories were not tallied: for example, the presence or absence of a definite article (*had authority / had the authority*) in Figure 2 was ignored. After summing all alternations in each sentence pair, we calculated the average number of occurrences of each paraphrase type in each data set. The results are shown in Table 2.

Several major differences stand out between the two data sets. First, the F2 data is less parallel, as evidenced by the higher percentage of Elaborations found in those sentence pairs. Loss of parallelism, however, is offset by greater diversity of paraphrase types encountered in the F2 data. Phrasal alternations are more than 4x more common, and Reorderings occur over 20x more frequently. Thus while string difference methods may produce relatively clean training data, this is achieved at the cost of filtering out common (and interesting) paraphrase relationships.

6 Conclusions and Future Work

Edit distance identifies sentence pairs that exhibit lexical and short phrasal alternations that can be aligned with considerable success. Given a large dataset and a well-motivated clustering of documents, useful datasets can be gleaned even without resorting to more sophisticated techniques

(such as Multiple Sequence Alignment, as employed by Barzilay & Lee 2003).

However, there is a disparity between the kinds of paraphrase alternations that we need to be able to align and those that we can already align well using current SMT techniques. Based solely on the criterion of word AER, the L12 data would seem to be superior to the F2 data as a source of paraphrase knowledge. Hand evaluation, though, indicates that many of the phenomena that we are interested in learning may be absent from this L12 data. String edit distance extraction techniques involve assumptions about the data that are inadequate, but achieve high precision. Techniques like our F2 extraction strategies appear to extract a more diverse variety of data, but yield more noise. We believe that an approach with the strengths of both methods would lead to significant improvement in paraphrase identification and generation.

In the near term, however, the relatively similar performances of F2 and L12-trained models on the F2 test data suggest that with further refinements, this more complex type of data can achieve good results. More data will surely help.

One focus of future work is to build a classifier to predict whether two sentences are related through paraphrase. Features might include edit distance, temporal/topical clustering information, information about cross-document discourse structure, relative sentence length, and synonymy information. We believe that this work has potential impact on the fields of summarization, information retrieval, and question answering.

Our ultimate goal is to apply current SMT techniques to the problems of paraphrase recognition and generation. We feel that this is a natural extension of the body of recent developments in SMT; perhaps explorations in monolingual data may have a reciprocal impact. The field of SMT, long focused on closely aligned data, is only now beginning to address the kinds of problems immediately encountered in monolingual paraphrase (including phrasal translations and large scale reorderings). Algorithms to address these phenomena will be equally applicable to both fields. Of course a broad-domain SMT-influenced paraphrase solution will require very large corpora of sentential paraphrases. In this paper we have described just one example of a class of data extraction techniques that we hope will scale to this task.

Acknowledgements

We are grateful to the Mo Corston-Oliver, Jeff Stevenson and Amy Muia of the Butler Hill Group for their work in annotating the data used in the experiments. We have also benefited from discussions with Ken Church, Mark Johnson, Daniel Marcu and Franz Och. We remain, however, responsible for all content.

References

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

R. Barzilay and L. Lee. 2003. Learning to Paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL.*

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

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

D. Lin and P. Pantel. 2001. DIRT - Discovery of Inference Rules from Text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining.*

I. D. Melamed. 2001. *Empirical Methods for Exploiting Parallel Texts.* MIT Press.

R. Mihalcea and T. Pedersen. 2003 An Evaluation Exercise for Word Alignment. In *Proceedings of the Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond.* May 31, 2003. Edmonton, Canada.

F. Och and H. Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the ACL*, Hong Kong, China.

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

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

S. Vogel, H. Ney and C. Tillmann. 1996. HMM-Based Word Alignment in Statistical Translation. In *Proceedings of the Annual Meeting of the ACL*, Copenhagen, Denmark.