Paraphrasing with Bilingual Parallel Corpora

Colin Bannard Chris Callison-Burch
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
{c.j.bannard, callison-burch}@ed.ac.uk

Abstract

Previous work has used monolingual parallel corpora to extract and generate paraphrases. We show that this task can be done using bilingual parallel corpora, a much more commonly available resource. Using alignment techniques from phrase-based statistical machine translation, we show how paraphrases in one language can be identified using a phrase in another language as a pivot. We define a paraphrase probability that allows paraphrases extracted from a bilingual parallel corpus to be ranked using translation probabilities, and show how it can be refined to take contextual information into account. We evaluate our paraphrase extraction and ranking methods using a set of manual word alignments, and contrast the quality with paraphrases extracted from automatic alignments.

1 Introduction

Paraphrases are alternative ways of conveying the same information. Paraphrases are useful in a number of NLP applications. In natural language generation the production of paraphrases allows for the creation of more varied and fluent text (Iordanskaja et al., 1991). In multidocument summarization the identification of paraphrases allows information repeated across documents to be condensed (McKeown et al., 2002). In the automatic evaluation of machine translation, paraphrases may help to alleviate problems presented by the fact that there are often alternative and equally valid ways of translating a text (Pang et al., 2003). In question answering, discovering paraphrased answers may provide additional evidence that an answer is correct (Ibrahim et al., 2003).

In this paper we introduce a novel method for extracting paraphrases that uses bilingual parallel corpora. Past work (Barzilay and McKeown, 2001; Barzilay and Lee, 2003; Pang et al., 2003; Ibrahim et al., 2003) has examined the use of monolingual parallel corpora for paraphrase extraction. Examples of monolingual parallel corpora that have been used are multiple translations of classical French novels into English, and data created for machine translation evaluation methods such as Bleu (Papineni et al., 2002) which use multiple reference translations.

While the results reported for these methods are impressive, their usefulness is limited by the scarcity of monolingual parallel corpora. Small data sets mean a limited number of paraphrases can be extracted. Furthermore, the narrow range of text genres available for monolingual parallel corpora limits the range of contexts in which the paraphrases can be used.

Instead of relying on scarce monolingual parallel data, our method utilizes the abundance of bilingual parallel data that is available. This allows us to create a much larger inventory of phrases that is applicable to a wider range of texts.

Our method for identifying paraphrases is an extension of recent work in phrase-based statistical machine translation (Koehn et al., 2003). The essence of our method is to align phrases in a bilingual parallel corpus, and equate different English phrases that are aligned with the same phrase in the other language. This assumption of similar mean-

Proceedings of the 43rd Annual Meeting of the ACL, pages 597–604,
Ann Arbor, June 2005.

Emma burst into tears **and he tried to** comfort **her,** saying things to make her smile.
Emma cried, **and he tried to** console **her,** adorning his words with puns.

Figure 1: Using a monolingal parallel corpus to extract paraphrases

ing when multiple phrases map onto a single foreign language phrase is the converse of the assumption made in the word sense disambiguation work of Diab and Resnik (2002) which posits different word senses when a single English word maps onto different words in the foreign language (we return to this point in Section 4.4).

The remainder of this paper is as follows: Section 2 contrasts our method for extracting paraphrases with the monolingual case, and describes how we rank the extracted paraphrases with a probability assignment. Section 3 describes our experimental setup and includes information about how phrases were selected, how we manually aligned parts of the bilingual corpus, and how we evaluated the paraphrases. Section 4 gives the results of our evaluation and gives a number of example paraphrases extracted with our technique. Section 5 reviews related work, and Section 6 discusses future directions.

2 Extracting paraphrases

Much previous work on extracting paraphrases (Barzilay and McKeown, 2001; Barzilay and Lee, 2003; Pang et al., 2003) has focused on finding identifying contexts within aligned monolingual sentences from which divergent text can be extracted, and treated as paraphrases. Barzilay and McKeown (2001) gives the example shown in Figure 1 of how identical surrounding substrings can be used to extract the paraphrases of *burst into tears* as *cried* and *comfort* as *console*.

While monolingual parallel corpora often have identical contexts that can be used for identifying paraphrases, bilingual parallel corpora do not. Instead, we use phrases in the other language as pivots: we look at what foreign language phrases the English translates to, find all occurrences of those foreign phrases, and then look back at what other English phrases they translate to. We treat the other English phrases as potential paraphrases. Figure 2 illustrates how a German phrase can be used as a point of identification for English paraphrases in this way. Section 2.1 explains which statistical machine translation techniques are used to align phrases within sentence pairs in a bilingual corpus.

A significant difference between the present work and that employing monolingual parallel corpora, is that our method frequently extracts more than one possible paraphrase for each phrase. We assign a probability to each of the possible paraphrases. This is a mechanism for ranking paraphrases, which can be utilized when we come to select the correct paraphrase for a given context . Section 2.2 explains how we calculate the probability of a paraphrase.

2.1 Aligning phrase pairs

We use phrase alignments in a parallel corpus as pivots between English paraphrases. We find these alignments using recent *phrase-based* approaches to statistical machine translation.

The original formulation of statistical machine translation (Brown et al., 1993) was defined as a word-based operation. The probability that a foreign sentence is the translation of an English sentence is calculated by summing over the probabilities of all possible word-level alignments, $\mathbf{a}$, between the sentences:

$$p(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} p(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

Thus Brown et al. decompose the problem of determining whether a sentence is a good translation of another into the problem of determining whether there is a sensible mapping between the words in the sentences.

More recent approaches to statistical translation calculate the translation probability using larger blocks of aligned text. Koehn (2004), Tillmann (2003), and Vogel et al. (2003) describe various heuristics for extracting phrase alignments from the Viterbi word-level alignments that are estimated using Brown et al. (1993) models. We use the heuristic for phrase alignment described in Och and Ney (2003) which aligns phrases by incrementally building longer phrases from words and phrases which have adjacent alignment points.[1]

[1]Note that while we induce the translations of phrases from

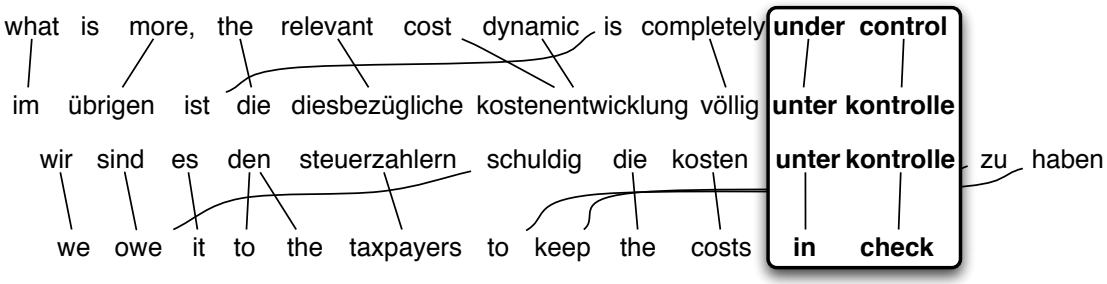


Figure 2: Using a bilingual parallel corpus to extract paraphrases

2.2 Assigning probabilities

We define a paraphrase probability $p(e_2|e_1)$ in terms of the translation model probabilities $p(f|e_1)$, that the original English phrase e_1 translates as a particular phrase f in the other language, and $p(e_2|f)$, that the candidate paraphrase e_2 translates as the foreign language phrase. Since e_1 can translate as multiple foreign language phrases, we sum over f:

$$\hat{e_2} = \arg\max_{e_2 \neq e_1} p(e_2|e_1) \quad (1)$$

$$= \arg\max_{e_2 \neq e_1} \sum_f p(f|e_1)p(e_2|f) \quad (2)$$

The translation model probabilities can be computed using any standard formulation from phrase-based machine translation. For example, $p(e|f)$ can be calculated straightforwardly using maximum likelihood estimation by counting how often the phrases e and f were aligned in the parallel corpus:

$$p(e|f) = \frac{count(e, f)}{\sum_e count(e, f)} \quad (3)$$

Note that the paraphrase probability defined in Equation 2 returns the single best paraphrase, $\hat{e_2}$, irrespective of the context in which e_1 appears. Since the best paraphrase may vary depending on information about the sentence that e_1 appears in, we extend the paraphrase probability to include that sentence S:

$$\hat{e_2} = \arg\max_{e_2 \neq e_1} p(e_2|e_1, S) \quad (4)$$

word-level alignments in this paper, direct estimation of phrasal translations (Marcu and Wong, 2002) would also suffice for extracting paraphrases from bilingual corpora.

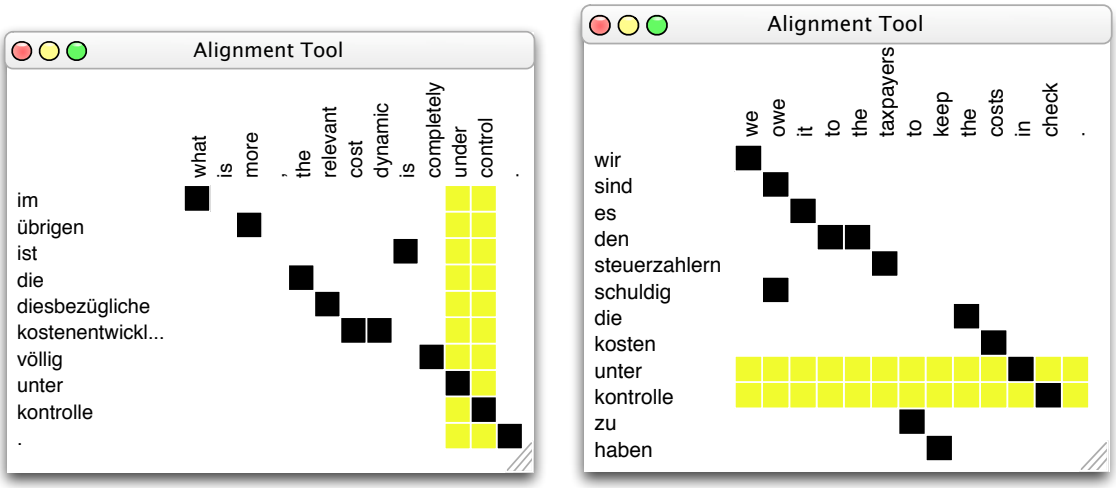
a million, as far as possible, at work, big business, carbon dioxide, central america, close to, concentrate on, crystal clear, do justice to, driving force, first half, for the first time, global warming, great care, green light, hard core, horn of africa, last resort, long ago, long run, military action, military force, moment of truth, new world, noise pollution, not to mention, nuclear power, on average, only too, other than, pick up, president clinton, public transport, quest for, red cross, red tape, socialist party, sooner or later, step up, task force, turn to, under control, vocational training, western sahara, world bank

Table 1: Phrases that were selected to paraphrase

S allows us to re-rank the candidate paraphrases based on additional contextual information. The experiments in this paper employ one variety of contextual information. We include a simple language model probability, which would additionally rank e_2 based on the probability of the sentence formed by substiuting e_2 for e_1 in S. A possible extension which we do not evaluate might be permitting only paraphrases that are the same syntactic type as the original phrase, which we could do by extending the translation model probabilities to count only phrase occurrences of that type.

3 Experimental Design

We extracted 46 English phrases to paraphrase (shown in Table 1), randomly selected from those multi-word phrases in WordNet which also occured multiple times in the first 50,000 sentences of our bilingual corpus. The bilingual corpus that we used



(a) Aligning the English phrase to be paraphrased

(b) Aligning occurrences of its German translation

Figure 3: Phrases highlighted for manual alignment

was the German-English section of the Europarl corpus, version 2 (Koehn, 2002). We produced automatic alignments for it with the Giza++ toolkit (Och and Ney, 2003). Because we wanted to test our method independently of the quality of word alignment algorithms, we also developed a gold standard of word alignments for the set of phrases that we wanted to paraphrase.

3.1 Manual alignment

The gold standard alignments were created by highlighting all occurrences of the English phrase to paraphrase and manually aligning it with its German equivalent by correcting the automatic alignment, as shown in Figure 3a. All occurrences of its German equivalents were then highlighted, and aligned with their English translations (Figure 3b). The other words in the sentences were left with their automatic alignments.

3.2 Paraphrase evaluation

We evaluated the accuracy of each of the paraphrases that was extracted from the manually aligned data, as well as the top ranked paraphrases from the experimental conditions detailed below in Section 3.3. Because the acccuracy of paraphrases can vary depending on context, we substituted each

Under control
This situation is **in check** in terms of security.
This situation is **checked** in terms of security.
This situation is **curbed** in terms of security.
This situation is **curb** in terms of security.
This situation is **limit** in terms of security.
This situation is **slow down** in terms of security.

Figure 4: Paraphrases substituted in for the original phrase

set of candidate paraphrases into between 2–10 sentences which contained the original phrase. Figure 4 shows the paraphrases for *under control* substituted into one of the sentences in which it occurred. We created a total of 289 such evaluation sets, with a total of 1366 unique sentences created through substitution.

We had two native English speakers produce judgments as to whether the new sentences preserved the meaning of the original phrase and as to whether they remained grammatical. Paraphrases that were judged to preserve both meaning and grammaticality were considered to be correct, and examples which failed on either judgment were considered to be incorrect.

In Figure 4 *in check*, *checked*, and *curbed* were

under control	checked, curb, curbed, *in check*, limit, slow down
sooner or later	*at some point*, eventually
military force	armed forces, defence, *force*, forces, military forces, peace-keeping personnel
long ago	a little time ago, a long time, *a long time ago*, a lot of time, a while ago, a while back, far, for a long time, for some time, for such a long time, long, long period of time, long term, long time, long while, overdue, some time, some time ago
green light	approval, call, *go-ahead*, indication, message, sign, signal, signals, formal go-ahead
great care	a careful approach, greater emphasis, *particular attention*, special attention, specific attention, very careful
first half	*first six months*
crystal clear	absolutely clear, all clarity, clear, clearly, in great detail, no mistake, no uncertain, obvious, obviously, particularly clear, perfectly clear, quite clear, quite clearly, quite explicitly, quite openly, very clear, *very clear and comprehensive*, very clearly, very sure, very unclear, very well
carbon dioxide	*co2*
at work	at the workplace, employment, held, holding, in the work sphere, operate, organised, taken place, took place, *working*

Table 2: Paraphrases extracted from a manually word-aligned parallel corpus

judged to be correct and *curb*, *limit* and *slow down* were judged to be incorrect. The inter-annotator agreement for these judgements was measured at $\kappa = 0.605$, which is conventionally interpreted as "good" agreement.

3.3 Experiments

We evaluated the accuracy of top ranked paraphrases when the paraphrase probability was calculated using:

1. The manual alignments,
2. The automatic alignments,
3. Automatic alignments produced over multiple corpora in different languages,
4. All of the above with language model re-ranking.
5. All of the above with the candidate paraphrases limited to the same sense as the original phrase.

4 Results

We report the percentage of correct translations (accuracy) for each of these experimental conditions. A summary of these can be seen in Table 3. This section will describe each of the set-ups and the score reported in more detail.

4.1 Manual alignments

Table 2 gives a set of example paraphrases extracted from the gold standard alignments. The italicized paraphrases are those that were assigned the highest probability by Equation 2, which chooses a single best paraphrase without regard for context. The 289 sentences created by substituting the italicized paraphrases in for the original phrase were judged to be correct an average of 74.9% of the time.

Ignoring the constraint that the new sentences remain grammatically correct, these paraphrases were judged to have the correct meaning 84.7% of the time. This suggests that the context plays a more important role with respect to the grammaticality of substituted paraphrases than with respect to their meaning.

In order to allow the surrounding words in the sentence to have an influence on which paraphrase was selected, we re-ranked the paraphrase probabilities based on a trigram language model trained on the entire English portion of the Europarl corpus. Paraphrases were selected from among all those in Table 2, and not constrained to the italicized phrases. In the case of the paraphrases extracted from the manual word alignments, the language model re-ranking had virtually no influence, and resulted in a slight dip in accuracy to 71.7%

	Paraphrase Prob	Paraphrase Prob & LM	Correct Meaning
Manual Alignments	74.9	71.7	84.7
Automatic Alignments	48.9	55.3	64.5
Using Multiple Corpora	55.0	57.4	65.4
Word Sense Controlled	57.0	61.9	70.4

Table 3: Paraphrase accuracy and correct meaning for the different data conditions

4.2 Automatic alignments

In this experimental condition paraphrases were extracted from a set of automatic alignments produced by running Giza++ over a set of 1,036,000 German-English sentence pairs (roughly 28,000,000 words in each language). When the single best paraphrase (irrespective of context) was used in place of the original phrase in the evaluation sentence the accuracy reached 48.9% which is quite low compared to the 74.9% of the manually aligned set.

As with the manual alignments it seems that we are selecting phrases which have the correct meaning but are not grammatical in context. Indeed our judges thought the meaning of the paraphrases to be correct in 64.5% of cases. Using a language model to select the best paraphrase given the context reduces the number of ungrammatical examples and gives an improvement in quality from 48.9% to 55.3% correct.

These results suggest two things: that improving the quality of automatic alignments would lead to more accurate paraphrases, and that there is room for improvement in limiting the paraphrases by their context. We address these points below.

4.3 Using multiple corpora

Work in statistical machine translation suggests that, like many other machine learning problems, performance increases as the amount of training data increases. Och and Ney (2003) show that the accuracy of alignments produced by Giza++ improve as the size of the training corpus increases.

Since we used the whole of the German-English section of the Europarl corpus, we could not try improving the alignments by simply adding more German-English training data. However, there is nothing that limits our paraphrase extraction method to drawing on candidate paraphrases from a single target language. We therefore re-formulated the paraphrase probability to include multiple corpora, as follows:

$$\hat{e_2} = \arg\max_{e_2 \neq e_1} \sum_{C} \sum_{f \ in \ C} p(f|e_1)p(e_2|f) \quad (5)$$

where C is a parallel corpus from a set of parallel corpora.

For this condition we used Giza++ to align the French-English, Spanish-English, and Italian-English portions of the Europarl corpus in addition to the German-English portion, for a total of around 4,000,000 sentence pairs in the training data.

The accuracy of paraphrases extracted over multiple corpora increased to 55%, and further to 57.4% when the language model re-ranking was included.

4.4 Controlling for word sense

As mentioned in Section 1, the way that we extract paraphrases is the converse of the methodology employed in word sense disambiguation work that uses parallel corpora (Diab and Resnik, 2002). The assumption made in the word sense disambiguation work is that if a source language word aligns with different target language words then those words may represent different word senses. This can be observed in the paraphrases for *at work* in Table 2. The paraphrases *at the workplace, employment,* and *in the work sphere* are a different sense of the phrase than *operate, held,* and *holding,* and they are aligned with different German phrases.

When we calculate the paraphrase probability we sum over different target language phrases. Therefore the English phrases that are aligned with the different German phrases (which themselves maybe indicative of different word senses) are mingled. Performance may be degraded since paraphrases that reflect different senses of the original phrase, and which therefore have a different meaning, are included in the same candidate set.

We therefore performed an experiment to see whether improvement could be had by limiting the candidate paraphrases to be the same sense as the original phrase in each test sentence. To do this, we used the fact that our test sentences were drawn from a parallel corpus. We limited phrases to the same word sense by constraining the candidate paraphrases to those that aligned with the same target language phrase. Our basic paraphrase calculation was therefore:

$$p(e_2|e_1, f) = p(f|e_1)p(e_2|f) \quad (6)$$

Using the foreign language phrase to identify the word sense is obviously not applicable in monolingual settings, but acts as a convenient stand-in for a proper word sense disambiguation algorithm here.

When word sense is controlled in this way, the accuracy of the paraphrases extracted from the automatic alignments raises dramatically from 48.9% to 57% without language model re-ranking, and further to 61.9% when language model re-ranking was included.

5 Related Work

Barzilay and McKeown (2001) extract both single- and multiple-word paraphrases from a monolingual parallel corpus. They co-train a classifier to identify whether two phrases were paraphrases of each other based on their surrounding context. Two disadvantages of this method are that it requires identical bounding substrings, and has bias towards single words. For an evaluation set of 500 paraphrases, they report an average precision of 86% at identifying paraphrases out of context, and of 91% when the paraphrases are substituted into the original context of the aligned sentence. The results of our systems are not directly comparable, since Barzilay and McKeown (2001) evaluated their paraphrases with a different set of criteria (they asked judges whether to judge paraphrases based on "approximate conceptual equivalence"). Furthermore, their evaluation was carried out only by substituting the paraphrase in for the phrase with the identical context, and not in for arbitrary occurrences of the original phrase, as we have done.

Lin and Pantel (2001) use a standard (non-parallel) monolingual corpus to generate paraphrases, based on dependancy graphs and distributional similarity. One strong disadvantage of this method is that their paraphrases can also have opposite meanings.

Ibrahim et al. (2003) combine the two approaches: aligned monolingual corpora and parsing. They evaluated their system with human judges who were asked whether the paraphrases were "roughly interchangeable given the genre", scored an average of 41% on a set of 130 paraphrases, with the judges all agreeing 75% of the time, and a correlation of 0.66. The shortcomings of this method are that it is dependent upon parse quality, and is limited by the rareness of the data.

Pang et al. (2003) use parse trees over sentences in monolingual parallel corpus to identify paraphrases by grouping similar syntactic constituents. They use heuristics such as keyword checking to limit the over-application of this method. Our alignment method might be an improvement of their heuristics for choosing which constituents ought to be grouped.

6 Discussion and Future Work

In this paper we have introduced a novel method for extracting paraphrases, which we believe greatly increases the usefulness of paraphrasing in NLP applications. The advantages of our method are that it:

- Produces a ranked list of high quality paraphrases with associated probabilities, from which the best paraphrase can be chosen according to the target context. We have shown how a language model can be used to select the best paraphrase for a particular context from this list.
- Straightforwardly handles multi-word units. Whereas for previous approaches the evaluation has been performed over mostly single word paraphrases, our results are reported exclusively over units of between 2 and 4 words.
- Because we use a much more abundant source of data, our method can be used for a much wider range of text genres than previous approaches, namely any for which parallel data is available.

One crucial thing to note is that we have demonstrated our paraphrases to be of higher quality when the alignments used to produce them are improved. This means that our method will reap the benefits of research that improvements to automatic alignment techniques (Callison-Burch et al., 2004), and will further improve as more parallel data becomes available.

In the future we plan to:

- Investigate whether our re-ranking can be further improved by using a syntax-based language model.
- Formulate a paraphrase probability for sentential paraphrases, and use this to try to identify paraphrases across documents in order to condense information for multi-document summarization.
- See whether paraphrases can be used to increase coverage for statistical machine translation when translating into "low-density" languages which have small parallel corpora.

Acknowledgments

The authors would like to thank Beatrice Alex, Marco Kuhlmann, and Josh Schroeder for their valuable input as well as their time spent annotating and contributing to the software.

References

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL*.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of ACL*.

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Chris Callison-Burch, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of ACL*.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of ACL*.

Ali Ibrahim, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the Second International Workshop on Paraphrasing (ACL 2003)*.

Lidija Iordanskaja, Richard Kittredge, and Alain Polgére. 1991. Lexical selection and paraphrase in a meaning-text generation model. In Cécile L. Paris, William R. Swartout, and William C. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*. Kluwer Academic.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*.

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Unpublished Draft.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA*.

Dekang Lin and Patrick Pantel. 2001. DIRT - discovery of inference rules from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of EMNLP*.

Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. 2002. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of the Human Language Technology Conference*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of HLT/NAACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Christoph Tillmann. 2003. A projection extension algorithm for statistical machine translation. In *Proceedings of EMNLP*.

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, and Alex Waibel. 2003. The CMU statistical machine translation system. In *Proceedings of MT Summit 9*.