Using Paraphrases for Parameter Tuning in Statistical Machine Translation

Nitin Madnani, Necip Fazil Ayan, Philip Resnik & Bonnie J. Dorr
Institute for Advanced Computer Studies
University of Maryland
College Park, MD, 20742
{nmadnani,nfa,resnik,bonnie}@umiacs.umd.edu

Abstract

Most state-of-the-art statistical machine translation systems use log-linear models, which are defined in terms of hypothesis features and weights for those features. It is standard to tune the feature weights in order to maximize a translation quality metric, using held-out test sentences and their corresponding reference translations. However, obtaining reference translations is expensive. In this paper, we introduce a new full-sentence paraphrase technique, based on English-to-English decoding with an MT system, and we demonstrate that the resulting paraphrases can be used to drastically reduce the number of human reference translations needed for parameter tuning, without a significant decrease in translation quality.

1 Introduction

Viewed at a very high level, statistical machine translation involves four phases: language and translation model training, parameter tuning, decoding, and evaluation (Lopez, 2007; Koehn et al., 2003). Since their introduction in statistical MT by Och and Ney (2002), log-linear models have been a standard way to combine sub-models in MT systems. Typically such a model takes the form

$$\sum_i \lambda_i \phi_i(\bar{f}, \bar{e}) \tag{1}$$

where ϕ_i are features of the hypothesis e and λ_i are weights associated with those features.

Selecting appropriate weights λ_i is essential in order to obtain good translation performance. Och (2003) introduced minimum error rate training (MERT), a technique for optimizing log-linear model parameters relative to a measure of translation quality. This has become much more standard than optimizing the conditional probability of the training data given the model (i.e., a maximum likelihood criterion), as was common previously. Och showed that system performance is best when parameters are optimized using the same objective function that will be used for evaluation; BLEU (Papineni et al., 2002) remains common for both purposes and is often retained for parameter optimization even when alternative evaluation measures are used, e.g., (Banerjee and Lavie, 2005; Snover et al., 2006).

Minimum error rate training—and more generally, optimization of parameters relative to a translation quality measure—relies on data sets in which source language sentences are paired with (sets of) reference translations. It is widely agreed that, at least for the widely used BLEU criterion, which is based on n-gram overlap between hypotheses and reference translations, the criterion is most accurate when computed with as many distinct reference translations as possible. Intuitively this makes sense: if there are alternative ways to phrase the meaning of the source sentence in the target language, then the translation quality criterion should take as many of those variations into account as possible. To do otherwise is to risk the possibility that the criterion might judge good translations to be poor when they fail to match the exact wording within the reference translations that have been provided.

This reliance on multiple reference translations creates a problem, because reference translations are labor intensive and expensive to obtain. A common source of translated data for MT research is the Linguistic Data Consortium (LDC), where an elaborate process is undertaken that involves translation agencies, detailed translation guidelines, and quality control processes (Strassel et al., 2006). Some

Proceedings of the Second Workshop on Statistical Machine Translation, pages 120–127,
Prague, June 2007.

efforts have been made to develop alternative processes for eliciting translations, e.g., from users on the Web (Oard, 2003) or from informants in low-density languages (Probst et al., 2002). However, reference translations for parameter tuning and evaluation remain a severe data bottleneck for such approaches.

Note, however, one crucial property of reference translations: they are paraphrases, i.e., multiple expressions of the same meaning. Automatic techniques exist for generating paraphrases. Although one would clearly like to retain human translations as the benchmark for *evaluation* of translation, might it be possible to usefully increase the number of reference translations for *tuning* by using automatic paraphrase techniques?

In this paper, we demonstrate that it is, in fact, possible to do so. Section 2 briefly describes our translation framework. Section 3 lays out a novel technique for paraphrasing, designed with the application to parameter tuning in mind. Section 4 presents evaluation results using a state of the art statistical MT system, demonstrating that half the human reference translations in a standard 4-reference tuning set can be replaced with automatically generated paraphrases, with no significant decrease in MT system performance. In Section 5 we discuss related work, and in Section 6 we summarize the results and discuss plans for future research.

2 Translation Framework

The work described in this paper makes use of the Hiero statistical MT framework (Chiang, 2007). Hiero is formally based on a weighted synchronous context-free grammar (CFG), containing synchronous rules of the form

$$X \rightarrow \langle \bar{e}, \bar{f}, \phi_1^k(\bar{f}, \bar{e}, X) \rangle \tag{2}$$

where X is a symbol from the nonterminal alphabet, and $\bar{e}$ and $\bar{f}$ can contain both words (terminals) and variables (nonterminals) that serve as placeholders for other phrases. In the context of statistical MT, where phrase-based models are frequently used, these synchronous rules can be interpreted as pairs of *hierarchical phrases*. The underlying strength of a hierarchical phrase is that it allows for effective learning of not only the lexical re-orderings, but phrasal re-orderings, as well. Each $\phi(\bar{e}, \bar{f}, X)$ denotes a feature function defined on the pair of hierarchical phrases.[1] Feature functions represent conditional and joint co-occurrence probabilities over the hierarchical paraphrase pair.

The Hiero framework includes methods to learn grammars and feature values from unannotated parallel corpora, without requiring syntactic annotation of the data. Briefly, training a Hiero model proceeds as follows:

- GIZA++ (Och and Ney, 2000) is run on the parallel corpus in both directions, followed by an alignment refinement heuristic that yields a many-to-many alignment for each parallel sentence.
- Initial phrase pairs are identified following the procedure typically employed in phrase based systems (Koehn et al., 2003; Och and Ney, 2004).
- Grammar rules in the form of equation (2) are induced by "subtracting" out hierarchical phrase pairs from these initial phrase pairs.
- Fractional counts are assigned to each produced rule:

$$c(X \rightarrow \langle \bar{e}, \bar{f} \rangle) = \sum_{j=1}^{m} \frac{1}{n_{jr}} \tag{3}$$

  where m is the number of initial phrase pairs that give rise to this grammar rule and n_{jr} is the number of grammar rules produced by the j^{th} initial phrase pair.
- Feature functions $\phi_1^k(\bar{f}, \bar{e}, X)$ are calculated for each rule using the accumulated counts.

Once training has taken place, minimum error rate training (Och, 2003) is used to tune the parameters λ_i.

Finally, decoding in Hiero takes place using a CKY synchronous parser with beam search, augmented to permit efficient incorporation of language model scores (Chiang, 2007). Given a source language sentence *f*, the decoder parses the source language sentence using the grammar it has learned

[1] Currently only one nonterminal symbol is used in Hiero productions.

during training, with parser search guided by the model; a target-language hypothesis is generated simultaneously via the synchronous rules, and the yield of that hypothesized analysis represents the hypothesized string e in the target language.

3 Generating Paraphrases

As discussed in Section 1, our goal is to make it possible to accomplish the parameter-tuning phase using fewer human reference translations. We accomplish this by beginning with a small set of human reference translations for each sentence in the development set, and expanding that set by automatically paraphrasing each member of the set rather than by acquiring more human translations.

Most previous work on paraphrase has focused on high quality rather than coverage (Barzilay and Lee, 2003; Quirk et al., 2004), but generating artificial references for MT parameter tuning in our setting has two unique properties compared to other paraphrase applications. First, we would like to obtain 100% coverage, in order to avoid modifications to our minimum error rate training infrastructure.[2] Second, we prefer that paraphrases be as distinct as possible from the original sentences, while retaining as much of the original meaning as possible.

In order to satisfy these two properties, we approach sentence-level paraphrase for English as a problem of English-to-English translation, constructing the model using English-F translation, for a second language F, as a pivot. Following Bannard and Callison-Burch (2005), we first identify English-to-F correspondences, then map from English to English by following translation units from English to F and back. Then, generalizing their approach, we use those mappings to create a well defined English-to-English translation model. The parameters of this model are tuned using MERT, and then the model is used in an the (unmodified) statistical MT system, yielding sentence-level English paraphrases by means of decoding input English sentences. The remainder of this section presents this process in detail.

[2]Strictly speaking, this was not a requirement of the approach, but rather a concession to practical considerations.

3.1 Mapping and Backmapping

We employ the following strategy for the induction of the required monolingual grammar. First, we train the Hiero system in standard fashion on a bilingual English-F training corpus. Then, for each existing production in the resulting Hiero grammar, we create multiple new English-to-English productions by pivoting on the foreign hierarchical phrase in the rule. For example, assume that we have the following toy grammar for English-F, as produced by Hiero:

$$\begin{aligned} X &\rightarrow \langle \bar{e1}, \bar{f1} \rangle \\ X &\rightarrow \langle \bar{e3}, \bar{f1} \rangle \\ X &\rightarrow \langle \bar{e1}, \bar{f2} \rangle \\ X &\rightarrow \langle \bar{e2}, \bar{f2} \rangle \\ X &\rightarrow \langle \bar{e4}, \bar{f2} \rangle \end{aligned}$$

If we use the foreign phrase $\bar{f}1$ as a pivot and backmap, we can extract the two English-to-English rules: $X \rightarrow \langle \bar{e1}, \bar{e3} \rangle$ and $X \rightarrow \langle \bar{e3}, \bar{e1} \rangle$. Backmapping using both $\bar{f}1$ and $\bar{f}2$ produces the following new rules (ignoring duplicates and rules that map any English phrase to itself):

$$\begin{aligned} X &\rightarrow \langle \bar{e1}, \bar{e2} \rangle \\ X &\rightarrow \langle \bar{e1}, \bar{e3} \rangle \\ X &\rightarrow \langle \bar{e1}, \bar{e4} \rangle \\ X &\rightarrow \langle \bar{e2}, \bar{e1} \rangle \\ X &\rightarrow \langle \bar{e2}, \bar{e4} \rangle \end{aligned}$$

3.2 Feature values

Each rule production in a Hiero grammar is weighted by several feature values defined on the rule themselves. In order to perform accurate backmapping, we must recompute these feature functions for the newly created English-to-English grammar. Rather than computing approximations based on feature values already existing in the bilingual Hiero grammar, we calculate these features in a more principled manner, by computing maximum likelihood estimates directly from the fractional counts that Hiero accumulates in the penultimate training step.

We use the following features in our induced English-to-English grammar:[3]

[3]Hiero also uses lexical weights (Koehn et al., 2003) in both

- The joint probability of the two English hierarchical paraphrases, conditioned on the nonterminal symbol, as defined by this formula:

$$p(\bar{e_1}, \bar{e_2}|x) = \frac{c(X \rightarrow \langle \bar{e_1}, \bar{e_2} \rangle)}{\sum_{\bar{e_1}', \bar{e_2}'} c(X \rightarrow \langle \bar{e_1}', \bar{e_2}' \rangle)} = \frac{c(X \rightarrow \langle \bar{e_1}, \bar{e_2} \rangle)}{c(X)} \quad (4)$$

  where the numerator is the fractional count of the rule under consideration and the denominator represents the marginal count over all the English hierarchical phrase pairs.

- The conditionals $p(\bar{e_1}, x|\bar{e_2})$ and $p(\bar{e_2}, x|\bar{e_1})$ defined as follows:

$$p(\bar{e_1}, x|\bar{e_2}) = \frac{c(X \rightarrow \langle \bar{e_1}, \bar{e_2} \rangle)}{\sum_{\bar{e_1}'} c(X \rightarrow \langle \bar{e_1}', \bar{e_2} \rangle)} \quad (5)$$

$$p(\bar{e_2}, x|\bar{e_1}) = \frac{c(X \rightarrow \langle \bar{e_1}, \bar{e_2} \rangle)}{\sum_{\bar{e_2}'} c(X \rightarrow \langle \bar{e_1}, \bar{e_2}' \rangle)} \quad (6)$$

Finally, for all induced rules, we calculate a word penalty $exp(-T(\bar{e_2}))$, where $T(\bar{e_2})$ just counts the number of terminal symbols in $\bar{e_2}$. This feature allows the model to learn whether it should produce shorter or longer paraphrases.

In addition to the features above that are estimated from the training data, we also use a trigram language model. Since we are decoding to produce English sentences, we can use the same language model employed in a standard statistical MT setting.

Calculating the proposed features is complicated by the fact that we don't actually have the counts for English-to-English rules because there is no English-to-English parallel corpus. This is where the counts provided by Hiero come into the picture. We estimate the counts that we need as follows:

$$c(X \rightarrow \langle \bar{e_1}, \bar{e_2} \rangle) = \sum_{\bar{f}} c(X \rightarrow \langle \bar{e_1}, \bar{f} \rangle) c(X \rightarrow \langle \bar{e_2}, \bar{f} \rangle) \quad (7)$$

An intuitive way to think about the formula above is by using an example at the corpus level. Assume that, in the given bilingual parallel corpus, there are m sentences in which the English phrase $\bar{e_1}$ co-occurs with the foreign phrase $\bar{f}$ and n sentences in which the same foreign phrase $\bar{f}$ co-occurs with the English phrase $\bar{e_2}$. The problem can then be thought of as defining a function $g(m, n)$ which computes the number of sentences in a hypothetical English-to-English parallel corpus wherein the phrases $\bar{e_1}$ and $\bar{e_1}$ co-occur. For this paper, we define $g(m, n)$ to be the upper bound mn.

directions as features but we don't use them for our grammar.

Tables 1 and 2 show some examples of paraphrases generated by our system across a range of paraphrase quality for two different pivot languages.

3.3 Tuning Model Parameters

Although the goal of the paraphrasing approach is to make it less data-intensive to tune log-linear model parameters for translation, our paraphrasing approach, since it is based on an English-to-English log-linear model, also requires its own parameter tuning. This, however, is straightforward: regardless of how the paraphrasing model will be used in statistical MT, e.g., irrespective of source language, it is possible to use any existing set of English paraphrases as the tuning set for English-to-English translation. We used the 2002 NIST MT evaluation test set reference translations. For every item in the set, we randomly chose one sentence as the source sentence, and the remainder as the "reference translations" for purposes of minimum error rate training.

4 Evaluation

Having developed a paraphrasing approach based on English-to-English translation, we evaluated its use in improving minimum error rate training for translation from a second language into English.

Generating paraphrases via English-to-English translation makes use of a parallel corpus, from which a weighted synchronous grammar is automatically acquired. Although nothing about our approach requires that the paraphrase system's training bitext be the same one used in the translation experiments (see Section 6), doing so is not precluded, either, and it is a particularly convenient choice when the paraphrasing is being done in support of MT.[4] The training bitext comprised of Chinese-English

[4]The choice of the foreign language used as the pivot should not really matter but it is worth exploring this using other language pairs as our bitext.

O:	we must bear in mind the community as a whole .
P:	we must remember the wider community .
O:	thirdly , the implications of enlargement for the union 's regional policy cannot be overlooked .
P:	finally , the impact of enlargement for eu regional policy cannot be ignored .
O:	how this works in practice will become clear when the authority has to act .
P:	how this operate in practice will emerge when the government has to play .
O:	this is an ill-advised policy .
P:	this is an unwelcome in europe .

Table 1: Example paraphrases with French as the pivot language. **O** = Original Sentence, **P** = Paraphrase.

O:	alcatel added that the company's whole year earnings would be announced on february 4 .
P:	alcatel said that the company's total annual revenues would be released on february 4 .
O:	he was now preparing a speech concerning the us policy for the upcoming world economic forum .
P:	he was now ready to talk with regard to the us policies for the forthcoming international economic forum .
O:	tibet has entered an excellent phase of political stability, ethnic unity and people living in peace .
P:	tibetans have come to cordial political stability, national unity and lived in harmony .
O:	its ocean and blue-sky scenery and the mediterranean climate make it world's famous scenic spot .
P:	its harbour and blue-sky appearance and the border situation decided it world's renowned tourist attraction .

Table 2: Example paraphrases with Chinese as the pivot language. **O** = Original Sentence, **P** = Paraphrase.

Corpus	# Sentences	# Words
HK News	542540	11171933
FBIS	240996	9121210
Xinhua	54022	1497562
News1	9916	314121
Treebank	3963	125848
Total	851437	22230674

Table 3: Chinese-English corpora used as training bitext both for paraphrasing and for evaluation.

parallel corpora containing $850,000$ sentence pairs – approx. 22 million words (details shown in Table 3).

As the source of development data for minimum error rate training, we used the 919 source sentences and human reference translations from the 2003 NIST Chinese-English MT evaluation exercise. As raw material for experimentation, we generated a paraphrase for each reference sentence via 1-best decoding using the English-to-English translation approach of Section 3.

As our test data, we used the 1082 source sentences and human reference translations from the 2005 NIST Chinese-English MT evaluation.

Our core experiment involved three conditions where the only difference was the set of references for the development set used for tuning feature weights. For each condition, once the weights were tuned, they were used to decode the test set. Note that for all the conditions, the decoded test set was always scored against the *same* four high-quality human reference translations included with the set.

The three experimental conditions were designed around the constraint that our development set contains a total of four human reference translations per sentence, and therefore a maximum of four human references with which to compute an upper bound:

- **Baseline (2H):** For each item in the development set, we randomly chose two of the four human-constructed reference translations as references for minimum error rate training.

- **Expanded (2H + 2P):** For each of the two human references in the baseline tuning set, we automatically generated a corresponding paraphrase using (1-best) English-to-English translation, decoding using the model developed in Section 3. This condition represents the critical case in which you have a limited number of human references (two, in this case) and augment them with artificially generated reference translations. This yields a set of four references for minimum error rate training (two human, two paraphrased), which permits a direct comparison against the upper bound of four human-generated reference translations.

- **Upper bound: 4H**: We performed minimum error rate training using the four human references from the development set.

In addition to these core experimental conditions, we added a fourth condition to assess the effect on performance when all four human reference translations are used in expanding the reference set via paraphrase:

- **Expanded (4H + 4P):** This is the same as Condition 2, but using all four human references.

Note that since we have only four human references per item, this fourth condition does not permit comparison with an upper bound of eight human references.

Table 4 shows BLEU and TER scores on the test set for all four conditions.[5] If only two human references were available (simulated by using only two of the available four), expanding to four using paraphrases would yield a clear improvement. Using bootstrap resampling to compute confidence intervals (Koehn, 2004), we find that the improvement in BLEU score is statistically significant at $p < .01$.

Equally interesting, expanding the number of reference translations from two to four using paraphrases yields performance that approaches the upper bound obtained by doing MERT using all four human reference translations. The difference in BLEU between conditions 2 and 3 is *not* significant.

Finally, our fourth condition asks whether it is possible to improve MT performance given the typical four human reference translations used for MERT in most statistical MT systems, by adding a paraphrase to each one for a total eight references per translation. There is indeed further improvement, although the difference in BLEU score does not reach significance.

[5]We plan to include METEOR scores in future experiments.

Condition	References used	BLEU	TER
1	2 H	30.43	59.82
2	2 H + 2 P	31.10	58.79
3	4 H	31.26	58.66
4	4 H + 4 P	31.68	58.24

Table 4: BLEU and TER scores showing utility of paraphrased reference translations. **H** = human references, **P** = paraphrased references.

We also evaluated our test set using TER (Snover et al., 2006) and observed that the TER scores follow the same trend as the BLEU scores. Specifically, the TER scores demonstrate that using paraphrases to artificially expand the reference set is better than using only 2 human reference translations and as good as using 4 human reference translations.[6]

5 Related Work

The approach we have taken here arises from a typical situation in NLP systems: the lack of sufficient data to accurately estimate a model based on supervised training data. In a structured prediction problem such as MT, we have an example input and a single labeled, correct output. However, this output is chosen from a space in which the number of possible outputs is exponential in the input size, and in which there are many good outputs in this space (although they are vastly outnumbered by the bad outputs). Various discriminative learning methods have attempted to deal with the first of these issues, often by restricting the space of examples. For instance, some max-margin methods restrict their computations to a set of examples from a "feasible set," where they are expected to be maximally discriminative (Tillmann and Zhang, 2006). The present approach deals with the second issue: in a learning problem where the use of a single positive example is likely to be highly biased, how can we produce a set of positive examples that is more representative of the space of correct outcomes? Our method exploits alternative sources of information to produce new positive examples that are, we hope, reasonably likely to represent a consensus of good examples.

Quite a bit of work has been done on paraphrase,

[6]We anticipate doing significance tests for differences in TER in future work.

some clearly related to our technique, although in general previous work has been focused on human readability rather than high coverage, noisy paraphrases for use downstream in an automatic process.

At the sentence level, (Barzilay and Lee, 2003) employed an unsupervised learning approach to cluster sentences and extract *lattice pairs* from comparable monolingual corpora. Their technique produces a paraphrase *only* if the input sentence matches any of the extracted lattice pairs, leading to a bias strongly favoring quality over coverage. They were able to generate paraphrases for 59 sentences (12%) out of a 484-sentence test set, generating no paraphrases at all for the remainder.

Quirk et al. (2004) also generate sentential paraphrases using a monolingual corpus. They use IBM Model-1 scores as the only feature, and employ a monotone decoder (i.e., one that cannot produce phrase-level reordering). This approach emphasizes very simple "substitutions of words and short phrases," and, in fact, almost a third of their best sentential "paraphrases" are identical to the input sentence.

A number of other approaches rely on parallel monolingual data and, additionally, require parsing of the training sentences (Ibrahim et al., 2003; Pang et al., 2003). Lin and Pantel (2001) use a non-parallel corpus and employ a dependency parser and computation of distributional similarity to learn paraphrases.

There has also been recent work on using paraphrases to improve statistical machine translation. Callison-Burch et al. (2006) extract phrase-level paraphrases by mapping input phrases into a phrase table and then mapping back to the source language. However, they do not generate paraphrases of entire sentences, but instead employ paraphrases to add entries to an existing phrase table solely for the purpose of increasing source-language coverage.

Other work has incorporated paraphrases into MT evaluation: Russo-Lassner et al. (2005) use a combination of paraphrase-based features to evaluate translation output; Zhou et al. (2006) propose a new metric that extends n-gram matching to include synonyms and paraphrases; and Lavie's METEOR metric (Banerjee and Lavie, 2005) can be used with additional knowledge such as WordNet in order to support inexact lexical matches.

6 Conclusions and Future Work

We introduced an automatic paraphrasing technique based on English-to-English translation of full sentences using a statistical MT system, and demonstrated that, using this technique, it is possible to cut in half the usual number of reference translations used for minimum error rate training with no significant loss in translation quality. Our method enables the generation of paraphrases for thousands of sentences in a very short amount of time (much shorter than creating other low-cost human references). This might prove beneficial for various discriminative training methods (Tillmann and Zhang, 2006).

This has important implications for data acquisition strategies For example, it suggests that rather than obtaining four reference translations per sentence for development sets, it may be more worthwhile to obtain fewer translations for a wider range of sentences, e.g., expanding into new topics and genres. In addition, this approach can significantly increase the utility of datasets which include only a single reference translation.

A number of future research directions are possible. First, since we have already demonstrated that noisy paraphrases can nonetheless add value, it would be straightforward to explore the quantity/quality tradeoff by expanding the MERT reference translations with n-best paraphrases for $n > 1$.

We also plan to conduct an intrinsic evaluation of the quality of paraphrases that our technique generates. It is important to note that a different tradeoff ratio may lead to even better results, e.g, using *only* the paraphrased references when they pass some goodness threshold, as used in Ueffing's (2006) self-training MT approach.

We have also observed that named entities are usually paraphrased incorrectly if there is a genre mismatch between the training and the test data. The Hiero decoder allows spans of source text to be annotated with inline translations using XML. We plan to identify and annotate named entities in the English source so that they are left unchanged.

Also, since the language F for English-F pivoting is arbitrary, we plan to investigate using English-to-English grammars created using *multiple* English-F grammars based on different languages, both individually and in combination, in order to improve paraphrase quality.

We also plan to explore a wider range of paraphrase-creation techniques, ranging from simple word substitutions (e.g., based on WordNet) to using the pivot technique with other translations systems.

7 Acknowledgments

We are indebted to David Chiang, Adam Lopez and Smaranda Muresan for insights and comments. This work has been supported under the GALE program of the Defense Advaned Research Projects Agency, Contract No. HR0011-06-2-001. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the view of DARPA.

References

S. Banerjee and A. Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at ACL.*

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL.*

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

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of HLT-NAACL.*

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2).

A. Ibrahim, B. Katz, and J. Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings the Second International Workshop on Paraphrasing (ACL 2003).*

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

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*.

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*.

A. Lopez. 2007. A survey of statistical machine translation. Technical Report 2006-47, University of Maryland, College Park.

D. W. Oard. 2003. The surprise langauge exercises. *ACM Transactions on Asian Language Information Processing*, 2(3).

Franz J. Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of ACL.*

Franz J. Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL.*

Franz Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4).

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL.*

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.*

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL.*

K. Probst, L. Levin, E. Peterson, A. Lavie, and J. Carbonell. 2002. Mt for minority languages using elicitation-based learning of syntactic transfer rules. *Machine Translation*, 17(4).

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP 2004*.

Grazia Russo-Lassner, Jimmy Lin, and Philip Resnik. 2005. A paraphrase-based approach to machine translation evaluation. Technical Report UMIACS-TR-2005-57, University of Maryland, College Park.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA.*

S. Strassel, C. Cieri, A. Cole, D. DiPersio, M. Liberman, X. Ma, M. Maamouri, and K. Maeda. 2006. Integrated linguistic resources for language exploitation technologies. In *Proceedings of LREC.*

Christoph Tillmann and Tong Zhang. 2006. A discriminative global training algorithm for statistical MT. In *Proceedings of ACL.*

Nicola Ueffing. 2006. Using monolingual source-language data to improve MT performance. In *Proceedings of IWSLT.*

L. Zhou, C.-Y. Lin, D. Muntenau, and E. Hovy. 2006. ParaEval: Using paraphrases to evaluate summaries automatically. In *Proceedings of HLT-NAACL.*