Improved Statistical Machine Translation Using Paraphrases

Chris Callison-Burch Philipp Koehn Miles Osborne
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
callison-burch@ed.ac.uk

Abstract

Parallel corpora are crucial for training SMT systems. However, for many language pairs they are available only in very limited quantities. For these language pairs a huge portion of phrases encountered at run-time will be unknown. We show how techniques from paraphrasing can be used to deal with these otherwise unknown source language phrases. Our results show that augmenting a state-of-the-art SMT system with paraphrases leads to significantly improved coverage and translation quality. For a training corpus with 10,000 sentence pairs we increase the coverage of unique test set unigrams from 48% to 90%, with more than half of the newly covered items accurately translated, as opposed to none in current approaches.

1 Introduction

As with many other statistical natural language processing tasks, statistical machine translation (Brown et al., 1993) produces high quality results when ample training data is available. This is problematic for so called "low density" language pairs which do not have very large parallel corpora. For example, when words occur infrequently in a parallel corpus parameter estimates for word-level alignments can be inaccurate, which can in turn lead to inaccurate phrase translations. Limited amounts of training data can further lead to a problem of low coverage in that many phrases encountered at run-time are not observed in the training data and therefore their translations will not be learned.

Here we address the problem of unknown phrases. Specifically we show that upon encountering an unknown source phrase, we can substitute a paraphrase for it and then proceed using the translation of that paraphrase. We derive these paraphrases from resources that are external to the parallel corpus that the translation model is trained from, and we are able to exploit (potentially more abundant) parallel corpora from other language pairs to do so.

In this paper we:

- Define a method for incorporating paraphrases of unseen source phrases into the statistical machine translation process.
- Show that by translating paraphrases we achieve a marked improvement in coverage and translation quality, especially in the case of unknown words which to date have been left untranslated.
- Argue that while we observe an improvement in Bleu score, this metric is particularly poorly suited to measuring the sort of improvements that we achieve.
- Present an alternative methodology for targeted manual evaluation that may be useful in other research projects.

2 The Problem of Coverage in SMT

Statistical machine translation made considerable advances in translation quality with the introduction of phrase-based translation (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004). By

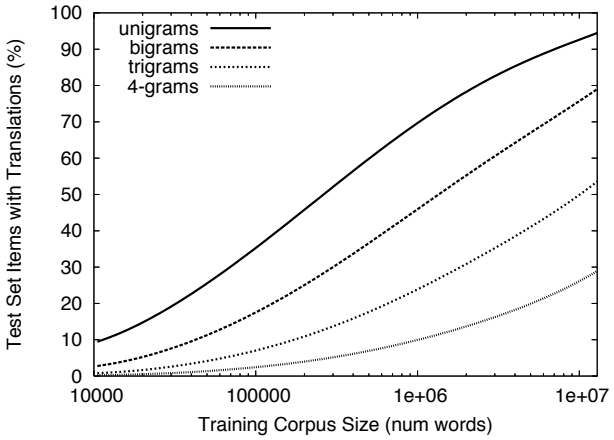

Figure 1: Percent of unique unigrams, bigrams, trigrams, and 4-grams from the Europarl Spanish test sentences for which translations were learned in increasingly large training corpora

increasing the size of the basic unit of translation, phrase-based machine translation does away with many of the problems associated with the original word-based formulation of statistical machine translation (Brown et al., 1993). For instance, with multi-word units less re-ordering needs to occur since local dependencies are frequently captured. For example, common adjective-noun alternations are memorized. However, since this linguistic information is not explicitly and generatively encoded in the model, unseen adjective noun pairs may still be handled incorrectly.

Thus, having observed phrases in the past dramatically increases the chances that they will be translated correctly in the future. However, for any given test set, a huge amount of training data has to be observed before translations are learned for a reasonable percentage of the test phrases. Figure 1 shows the extent of this problem. For a training corpus containing 10,000 words translations will have been learned for only 10% of the unigrams (*types*, not tokens). For a training corpus containing 100,000 words this increases to 30%. It is not until nearly 10,000,000 words worth of training data have been analyzed that translation for more than 90% of the vocabulary items have been learned. This problem is obviously compounded for higher-order n-grams (longer phrases), and for morphologically richer languages.

encargarnos	to ensure, take care, ensure that
garantizar	guarantee, ensure, guaranteed, assure, provided
velar	ensure, ensuring, safeguard, making sure
procurar	ensure that, try to, ensure, endeavour to
asegurarnos	ensure, secure, make certain
usado	used
utilizado	used, use, spent, utilized
empleado	used, spent, employee
uso	use, used, usage
utiliza	used, uses, used, being used
utilizar	to use, use, used

Table 1: Example of automatically generated paraphrases for the Spanish words *encargarnos* and *usado* along with their English translations which were automatically learned from the Europarl corpus

2.1 Handling unknown words

Currently most statistical machine translation systems are simply unable to handle unknown words. There are two strategies that are generally employed when an unknown source word is encountered. Either the source word is simply omitted when producing the translation, or alternatively it is passed through untranslated, which is a reasonable strategy if the unknown word happens to be a name (assuming that no transliteration need be done). Neither of these strategies is satisfying.

2.2 Using paraphrases in SMT

When a system is trained using 10,000 sentence pairs (roughly 200,000 words) there will be a number of words and phrases in a test sentence which it has not learned the translation of. For example, the Spanish sentence

> *Es positivo llegar a un acuerdo sobre los procedimientos, pero debemos encargarnos de que este sistema no sea susceptible de ser usado como arma política.*

may translate as

> It is good reach an agreement on procedures, but we must *encargarnos* that this system is not susceptible to be *usado* as political weapon.

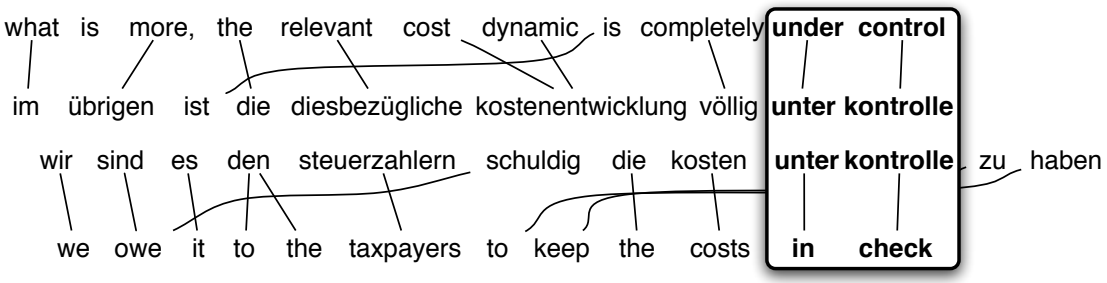

Figure 2: Using a bilingual parallel corpus to extract paraphrases

The strategy that we employ for dealing with unknown source language words is to substitute paraphrases of those words, and then translate the paraphrases. Table 1 gives examples of paraphrases and their translations. If we had learned a translation of *garantizar* we could translate it instead of *encargarnos*, and similarly for *utilizado* instead of *usado*.

3 Acquiring Paraphrases

Paraphrases are alternative ways of expressing the same information within one language. The automatic generation of paraphrases has been the focus of a significant amount of research lately. Many methods for extracting paraphrases (Barzilay and McKeown, 2001; Pang et al., 2003) make use of monolingual parallel corpora, such as multiple translations of classic French novels into English, or the multiple reference translations used by many automatic evaluation metrics for machine translation.

Bannard and Callison-Burch (2005) use bilingual parallel corpora to generate paraphrases. Paraphrases are identified by pivoting through phrases in another language. The foreign language translations of an English phrase are identified, all occurrences of those foreign phrases are found, and all English phrases that they translate back to are treated as potential paraphrases of the original English phrase. Figure 2 illustrates how a German phrase can be used as a point of identification for English paraphrases in this way.

The method defined in Bannard and Callison-Burch (2005) has several features that make it an ideal candidate for incorporation into statistical machine translation system. Firstly, it can easily be applied to any language for which we have one or more parallel corpora. Secondly, it defines a paraphrase probability, $p(e_2|e_1)$, which can be incorporated into the probabilistic framework of SMT.

3.1 Paraphrase probabilities

The paraphrase probability $p(e_2|e_1)$ is defined in terms of two translation model probabilities: $p(f|e_1)$, the probability that the original English phrase e_1 translates as a particular phrase f in the other language, and $p(e_2|f)$, the probability that the candidate paraphrase e_2 translates as the foreign language phrase. Since e_1 can translate as multiple foreign language phrases, we marginalize f out:

$$p(e_2|e_1) = \sum_f p(f|e_1)p(e_2|f) \quad (1)$$

The translation model probabilities can be computed using any standard formulation from phrase-based machine translation. For example, $p(e_2|f)$ can be calculated straightforwardly using maximum likelihood estimation by counting how often the phrases e and f were aligned in the parallel corpus:

$$p(e_2|f) \approx \frac{count(e_2, f)}{\sum_{e_2} count(e_2, f)} \quad (2)$$

There is nothing that limits us to estimating paraphrases probabilities from a single parallel corpus. We can extend the definition of the paraphrase probability to include multiple corpora, as follows:

$$p(e_2|e_1) \approx \frac{\sum_{c \in C} \sum_{f \text{ in } c} p(f|e_1)p(e_2|f)}{|C|} \quad (3)$$

where c is a parallel corpus from a set of parallel corpora C. Thus multiple corpora may be used

by summing over all paraphrase probabilities calculated from a single corpus (as in Equation 1) and normalized by the number of parallel corpora.

4 Experimental Design

We examined the application of paraphrases to deal with unknown phrases when translating from Spanish and French into English. We used the publicly available Europarl multilingual parallel corpus (Koehn, 2005) to create six training corpora for the two language pairs, and used the standard Europarl development and test sets.

4.1 Baseline

For a baseline system we produced a phrase-based statistical machine translation system based on the log-linear formulation described in (Och and Ney, 2002)

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) \quad (4)$$

$$= \arg\max_{\mathbf{e}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{f}) \quad (5)$$

The baseline model had a total of eight feature functions, $h_m(\mathbf{e}, \mathbf{f})$: a language model probability, a phrase translation probability, a reverse phrase translation probability, lexical translation probability, a reverse lexical translation probability, a word penalty, a phrase penalty, and a distortion cost. To set the weights, λ_m, we performed minimum error rate training (Och, 2003) on the development set using Bleu (Papineni et al., 2002) as the objective function.

The phrase translation probabilities were determined using maximum likelihood estimation over phrases induced from word-level alignments produced by performing Giza++ training on each of the three training corpora. We used the Pharaoh beam-search decoder (Koehn, 2004) to produce the translations after all of the model parameters had been set.

When the baseline system encountered unknown words in the test set, its behavior was simply to reproduce the foreign word in the translated output. This is the default behavior for many systems, as noted in Section 2.1.

4.2 Translation with paraphrases

We extracted all source language (Spanish and French) phrases up to length 10 from the test and development sets which did not have translations in phrase tables that were generated for the three training corpora. For each of these phrases we generated a list of paraphrases using all of the parallel corpora from Europarl aside from the Spanish-English and French-English corpora. We used bitexts between Spanish and Danish, Dutch, Finnish, French, German, Italian, Portuguese, and Swedish to generate our Spanish paraphrases, and did similarly for the French paraphrases. We manage the parallel corpora with a suffix array -based data structure (Callison-Burch et al., 2005). We calculated paraphrase probabilities using the Bannard and Callison-Burch (2005) method, summarized in Equation 3. Source language phrases that included names and numbers were not paraphrased.

For each paraphrase that had translations in the phrase table, we added additional entries in the phrase table containing the original phrase and the paraphrase's translations. We augmented the baseline model by incorporating the paraphrase probability into an additional feature function which assigns values as follows:

$$h(\mathbf{e}, \mathbf{f_1}) = \begin{cases} p(\mathbf{f_2}|\mathbf{f_1}) & \text{If phrase table entry } (\mathbf{e}, \mathbf{f_1}) \\ & \text{is generated from } (\mathbf{e}, \mathbf{f_2}) \\ 1 & \text{Otherwise} \end{cases}$$

Just as we did in the baseline system, we performed minimum error rate training to set the weights of the nine feature functions in our translation model that exploits paraphrases.

We tested the usefulness of the paraphrase feature function by performing an additional experiment where the phrase table was expanded but the paraphrase probability was omitted.

4.3 Evaluation

We evaluated the efficacy of using paraphrases in three ways: by calculating the Bleu score for the translated output, by measuring the increase in coverage when including paraphrases, and through a targeted manual evaluation of the phrasal translations of unseen phrases to determine how many of the newly covered phrases were accurately translated.

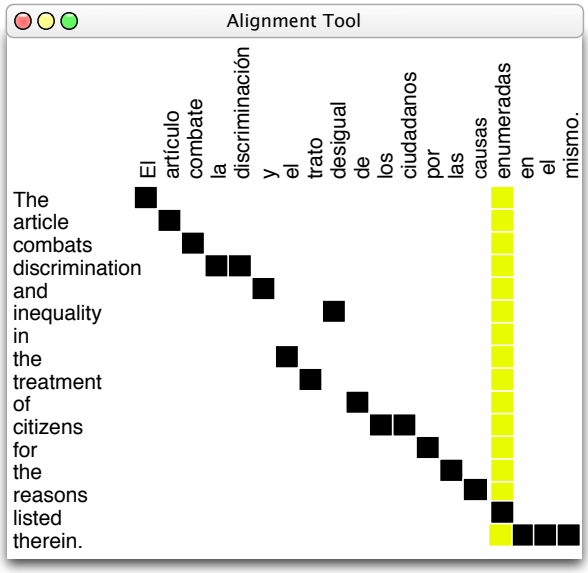

Figure 3: Test sentences and reference translations were manually word-aligned. This allowed us to equate unseen phrases with their corresponding English phrase. In this case *enumeradas* with *listed*.

The article combats discrimination and inequality in the treatment of citizens for the reasons **listed** therein.
The article combats discrimination and the different treatment of citizens for the reasons **mentioned** in the same.
The article fights against uneven and the treatment of citizens for the reasons **enshrined** in the same.
The article is countering discrimination and the unequal treatment of citizens for the reasons **that** in the same.

Figure 4: Judges were asked whether the highlighted phrase retained the same meaning as the highlighted phrase in the reference translation (top)

Although Bleu is currently the standard metric for MT evaluation, we believe that it may not meaningfully measure translation improvements in our setup. By substituting a paraphrase for an unknown source phrase there is a strong chance that its translation may also be a paraphrase of the equivalent target language phrase. Bleu relies on exact matches of n-grams in a reference translation. Thus if our translation is a paraphrase of the reference, Bleu will fail to score it correctly.

Because Bleu is potentially insensitive to the type of changes that we were making to the translations, we additionally performed a focused manual evaluation (Callison-Burch et al., 2006). To do this, had bilingual speakers create word-level alignments for the first 150 and 250 sentence in the Spanish-English and French-English test corpora, as shown in Figure 3. We were able to use these alignments to extract the translations of the Spanish and French words that we were applying our paraphrase method to.

Knowing this correspondence between foreign phrases and their English counterparts allowed us to directly analyze whether translations that were being produced from paraphrases remained faithful to the meaning of the reference translation. When producing our translations using the Pharaoh decoder we employed its "trace" facility, which tells which source sentence span each target phrase was derived from. This allowed us to identify which elements in the machine translated output corresponded to the paraphrased foreign phrase. We asked a monolingual judge whether the phrases in the machine translated output had the same meaning as of the reference phrase. This is illustrated in Figure 4.

In addition to judging the accuracy of 100 phrases for each of the translated sets, we measured how much our paraphrase method increased the coverage of the translation system. Because we focus on words that the system was previously unable to translate, the increase in coverage and the translation quality of the newly covered phrases are the two most relevant indicators as to the efficacy of the method.

5 Results

We produced translations under five conditions for each of our training corpora: a set of baseline translations without any additional entries in the phrase table, a condition where we added the translations of paraphrases for unseen source words along with paraphrase probabilities, a condition where we added the translations of paraphrases of multi-word phrases along with paraphrase probabilities, and two additional conditions where we added the translations of paraphrases of single and multi-word paraphrase without paraphrase probabilities.

	Spanish-English						French-English					
Corpus size	10k	20k	40k	80k	160k	320k	10k	20k	40k	80k	160k	320k
Baseline	22.6	25.0	26.5	26.5	28.7	**30.0**	21.9	24.3	26.3	27.8	28.8	29.5
Single word	23.1	25.2	26.6	**28.0**	**29.0**	**30.0**	22.7	24.2	26.9	27.7	28.9	**29.8**
Multi-word	**23.3**	**26.0**	**27.2**	**28.0**	28.8	29.7	**23.7**	**25.1**	**27.1**	**28.5**	**29.1**	**29.8**

Table 2: Bleu scores for the various training corpora, including baseline results without paraphrasing, results for only paraphrasing unknown words, and results for paraphrasing any unseen phrase. Corpus size is measured in sentences.

Corpus size	10k	20k	40k	80k	160k	320k	10k	20k	40k	80k	160k	320k
Single w/o-ff	23.0	25.1	26.7	28.0	**29.0**	29.9	22.5	24.1	26.0	27.6	28.8	29.6
Multi w/o-ff	20.6	22.6	21.9	24.0	25.4	27.5	19.7	22.1	24.3	25.6	26.0	28.1

Table 3: Bleu scores for the various training corpora, when the paraphrase feature function *is not* included

5.1 Bleu scores

Table 2 gives the Bleu scores for each of these conditions. We were able to measure a translation improvement for all sizes of training corpora, under both the single word and multi-word conditions, except for the largest Spanish-English corpus. For the single word condition, it would have been surprising if we had seen a decrease in Bleu score. Because we are translating words that were previously untranslatable it would be unlikely that we could do any worse. In the worst case we would be replacing one word that did not occur in the reference translation with another, and thus have no effect on Bleu.

More interesting is the fact that by paraphrasing unseen multi-word units we get an increase in quality above and beyond the single word paraphrases. These multi-word units may not have been observed in the training data as a unit, but each of the component words may have been. In this case translating a paraphrase would not be guaranteed to received an improved or identical Bleu score, as in the single word case. Thus the improved Bleu score is notable.

Table 3 shows that incorporating the paraphrase probability into the model's feature functions plays a critical role. Without it, the multi-word paraphrases harm translation performance when compared to the baseline.

5.2 Manual evaluation

We performed a manual evaluation by judging the accuracy of phrases for 100 paraphrased translations from each of the sets using the manual word alignments.[1] Table 4 gives the percentage of time that each of the translations of paraphrases were judged to have the same meaning as the equivalent target phrase. In the case of the translations of single word paraphrases for the Spanish accuracy ranged from just below 50% to just below 70%. This number is impressive in light of the fact that none of those items are correctly translated in the baseline model, which simply inserts the foreign language word. As with the Bleu scores, the translations of multi-word paraphrases were judged to be more accurate than the translations of single word paraphrases.

In performing the manual evaluation we were additionally able to determine how often Bleu was capable of measuring an actual improvement in translation. For those items judged to have the same meaning as the gold standard phrases we could track how many would have contributed to a higher Bleu score (that is, which of them were exactly the same as the reference translation phrase, or had some words in common with the reference translation phrase). By counting how often a correct phrase would have contributed to an increased Bleu score, and how often it would fail to increase the Bleu score we were able to determine with what frequency Bleu was sensitive to our improvements. We found that Bleu was insensitive to our translation improvements between 60-75% of the time, thus re-

[1] Note that for the larger training corpora fewer than 100 paraphrases occurred in the first 150 and 250 sentence pairs.

	Spanish-English						French-English					
Corpus size	10k	20k	40k	80k	160k	320k	10k	20k	40k	80k	160k	320k
Single word	48%	53%	57%	67%*	33%*	50%*	54%	49%	45%	50%	39%*	21%*
Multi-word	64%	65%	66%	71%	76%*	71%*	60%	67%	63%	58%	65%	42%*

Table 4: Percent of time that the translation of a paraphrase was judged to retain the same meaning as the corresponding phrase in the gold standard. Starred items had fewer than 100 judgments and should not be taken as reliable estimates.

Size	1-gram	2-gram	3-gram	4-gram
10k	48%	25%	10%	3%
20k	60%	35%	15%	6%
40k	71%	45%	22%	9%
80k	80%	55%	29%	12%
160k	86%	64%	37%	17%
320k	91%	71%	45%	22%

Table 5: The percent of the unique test set phrases which have translations in each of the Spanish-English training corpora prior to paraphrasing

Size	1-gram	2-gram	3-gram	4-gram
10k	90%	67%	37%	16%
20k	90%	69%	39%	17%
40k	91%	71%	41%	18%
80k	92%	73%	44%	20%
160k	92%	75%	46%	22%
320k	93%	77%	50%	25%

Table 6: The percent of the unique test set phrases which have translations in each of the Spanish-English training corpora after paraphrasing

inforcing our belief that it is not an appropriate measure for translation improvements of this sort.

5.3 Increase in coverage

As illustrated in Figure 1, translation models suffer from sparse data. When only a very small parallel corpus is available for training, translations are learned for very few of the unique phrases in a test set. If we exclude 451 words worth of names, numbers, and foreign language text in 2,000 sentences that comprise the Spanish portion of the Europarl test set, then the number of unique n-grams in text are: 7,331 unigrams, 28,890 bigrams, 44,194 trigrams, and 48,259 4-grams. Table 5 gives the percentage of these which have translations in each of the three training corpora, if we do not use paraphrasing.

In contrast after expanding the phrase table using the translations of paraphrases, the coverage of the unique test set phrases goes up dramatically (shown in Table 6). For the first training corpus with 10,000 sentence pairs and roughly 200,000 words of text in each language, the coverage goes up from less than 50% of the vocabulary items being covered to 90%. The coverage of unique 4-grams jumps from 3% to 16% – a level reached only after observing more than 100,000 sentence pairs, or roughly three million words of text, without using paraphrases.

6 Related Work

Previous research on trying to overcome data sparsity issues in statistical machine translation has largely focused on introducing morphological analysis as a way of reducing the number of types observed in a training text. For example, Nissen and Ney (2004) apply morphological analyzers to English and German and are able to reduce the amount of training data needed to reach a certain level of translation quality. Goldwater and McClosky (2005) find that stemming Czech and using lemmas improves the word-to-word correspondences when training Czech-English alignment models. Koehn and Knight (2003) show how monolingual texts and parallel corpora can be used to figure out appropriate places to split German compounds.

Still other approaches focus on ways of acquiring data. Resnik and Smith (2003) develop a method for gathering parallel corpora from the web. Oard et al. (2003) describe various methods employed for quickly gathering resources to create a machine translation system for a language with no initial resources.

7 Discussion

In this paper we have shown that significant gains in coverage and translation quality can be had by integrating paraphrases into statistical machine translation. In effect, paraphrases introduce some amount of *generalization* into statistical machine translation. Whereas before we relied on having observed a particular word or phrase in the training set in order to produce a translation of it, we are no longer tied to having seen every word in advance. We can exploit knowledge that is external to the translation model about what words have similar meanings and use that in the process of translation. This method is particularly pertinent to small data conditions, which are plagued by sparse data problems.

In future work, we plan to determine how much data is required to learn useful paraphrases. The scenario described in this paper was very favorable to creating high quality paraphrases. The large number of parallel corpora between Spanish and the other languages present in the Europarl corpus allowed us to generate high quality, in domain data. While this is a realistic scenario, in that many new official languages have been added to the European Union, some of which do not yet have extensive parallel corpora, we realize that this may be a slightly idealized scenario.

Finally, we plan to formalize our targeted manual evaluation method, in the hopes of creating a evaluation methodology for machine translation that is more thorough and elucidating than Bleu.

Acknowledgments

Thank you to Alexandra Birch and Stephanie Vandamme for creating the word alignments.

References

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

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

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, Colin Bannard, and Josh Schroeder. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proceedings of ACL*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation. In *Proceedings of EACL*.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of EMNLP*.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of EACL*.

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

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

Philipp Koehn. 2005. A parallel corpus for statistical machine translation. In *Proceedings of MT-Summit*.

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

Sonja Nissen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntatic analysis. *Computational Linguistics*, 30(2):181–204.

Doug Oard, David Doermann, Bonnie Dorr, Daqing He, Phillip Resnik, William Byrne, Sanjeeve Khudanpur, David Yarowsky, Anton Leuski, Philipp Koehn, and Kevin Knight. 2003. Desperately seeking Cebuano. In *Proceedings of HLT-NAACL*.

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

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*.

Franz Josef Och. 2003. Minimum error rate training for 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*.

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

Philip Resnik and Noah Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380, September.