Multi-granularity Word Alignment and Decoding for Agglutinative Language Translation

Zhiyang Wang † **Yajuan Lü** † **Qun Liu** †

†Key Lab. of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{wangzhiyang, lvyajuan, liuqun}@ict.ac.cn

Abstract

Lexical sparsity problem is much more serious for agglutinative language translation due to the multitude of inflected variants of lexicons. In this paper, we propose a novel optimization strategy to ease spareness by multi-granularity word alignment and translation for agglutinative language. Multiple alignment results are combined to catch the complementary information for alignments, and rules of different granularities can cooperate effectively to translate more unknown words. Experimental results on Uyghur-Chinese show that our proposed method significantly improves the quality of word alignment and translation, by relative 10.25% of alignment error rate reduction and $+2.46\%$ BLEU increment, respectively.

1 Introduction

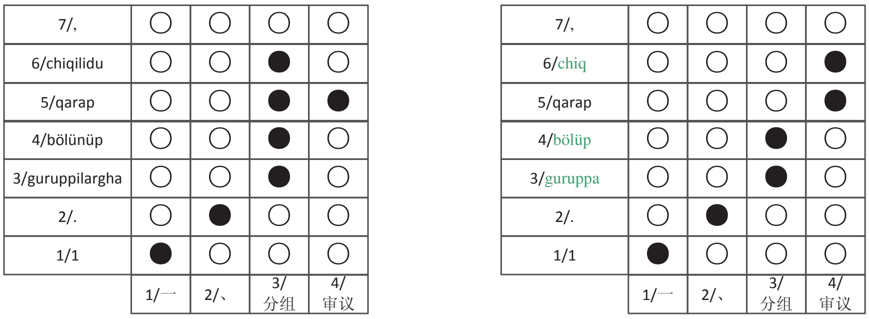
Current Statistical Machine Translation (SMT) (Brown et al., 1993; Koehn et al., 2003; Chiang, 2005; Galley et al., 2006) often suffers much from the lexical sparsity, especially in word alignment and decoding. The widely-used word alignment tool, GIZA++ (Och and Ney, 2003), performs better on large parallel corpora, but much poorer on small ones. Researchers have tried many methods to optimize alignment result such as heuristic methods (Koehn et al., 2003), and also many strategies to extract better translation rules (Liu et al., 2009). While for decoding, a lot of work has been devoted to the alleviation of unknown word problem, such as fuzzy phrase matching (He et al., 2008). These strategies work well for the translation of inflectional and isolating languages.

For the translation of agglutinative languages, however, data sparseness problem becomes much more serious. An inflectional language, such as English, expresses semantics by hierarchical structures of simple words, which explains why the tree-based translation models can alleviate lexical sparsity. But for agglutinative languages, semantics are expressed mainly by concatenation of stem and affixes. Usually, a stem can attach with several prefixes or suffixes, thus leading to tens of hundreds of possible derived words. Different from the Arabic-English translation that has achieved promising progress, such as (Lee, 2004; Habash and Sadat, 2006), most agglutinative languages are less-studied and suffer from resource-scarce problem.

Our work focuses on the translation of Uyghur. According to the agglutinative property of Uyghur, we propose a novel strategy, multi-granularity integration, to optimize word alignment and decoding for translating from Uyghur to Chinese.

To optimize the word alignment, multiple alignment results are combined by assigning each intertranslatable word pair an alignment probability. This procedure is performed for a series of alignment tasks with different lexical granularities, including stem-to-word alignment, and word-to-word alignment. The alignment results produced by all these tasks are integrated together to obtain more accurate word alignments.

While for decoding, translation rules of different granularities are also integrated in a weighted competitive manner. Word-granularity rules generate ac-



(a) (b)

Figure 1: An example of Uyghur-Chinese alignment, which means "*1. group deliberation*" in English. (a) is the word level alignment, (b) represents stem level alignment. The • means the corresponding words are aligned, ○ is not.

curate translation although with smaller coverage, and stem-granularity rules have much larger coverage while maybe with some ambiguities. Rules of different granularities can cooperate effectively in order to utilize the small bi-text as completely as possible.

End-to-end evaluation shows that our proposed method obviously reduces the alignment error rate by relative 10.25%, and also yields a significant increase in BLEU scores (+2.46%) for Uyghur-Chinese translation task.

The remaining part of this paper is organized as follows: we first give a simple introduction of Uyghur in section 2. Then the methods of multi-granularity alignment and translation are presented in section 3 and 4. Section 5 describes our experiments. Section 6 reviews the related work. Conclusions are drawn in section 7 finally.

2 Uyghur Language Issues

Uyghur belongs to the Altaic language family. Its alphabet consists of 32 letters, including 8 vowels and 24 consonants. Each letter may have different shapes in different positions of a word. As an agglutinative language, a Uyghur word can be decomposed, to the maximum extent, into a stem, a prefix, and a sequence of suffixes. There are 6 prefixes and 532 suffixes, including 243 derivational suffixes and 289 inflectional suffixes. One word in such language can generate tens of hundreds of inflected variants.

Uyghur is spoken by more than ten million people dwelling in Central Asia. This nation attracts much attentions of the world in recent years, but its

Items	#sent	#token	#type
Uy-word	114,579	2,675,101	**83,212**
Uy-stem	114,579	2,675,101	**22,783**
Ch-word	114,579	2,483,465	36,804

Table 1: Statistics of the training corpus. *Uy* means the bi-text of Uyghur side, and *Ch* represents Chinese side. "-*word*" means before stemming, "-*stem*" after stemming.

language is resource-poor and less-studied. We only collected about $120K$ Uyghur-Chinese sentence pairs currently, and such a small bi-text is apt to result in poor translation performance. To mitigate this problem, morphological analysis is necessary. More precisely, we will do stemming on Uyghur.

Stemming is a shallow morphological analysis, which uses a lexical entry to replace inflected words. In order to get the stem of Uyghur, we developed a generative statistical model for Uyghur morphological analysis according to (Jiang et al., 2010). This model describes the result of morphological analysis as a directed graph, where the nodes represent the stems, affixes and their tags, while the edges represent the transition or generation relationships between nodes. Let's take the collected parallel corpus as an example. The number of unique words is $83,212$ before stemming and $22,783$ after stemming. About 72.8% reduction in vocabulary is obtained by stemming.

Figure 1 gives a simple example of Uyghur-Chinese alignment which is produced by GIZA++. While employing the surface word to align, we get a really bad alignment result owing to word sparsi-

7/,	0	0. 13	0	0
6/chiqilidu	0	0	0. 98	0. 16
5/qarap	0	0	0. 07	1
4/bölünüp	0	0	1	0
3/guruppilargha	0	0. 07	1	0
2/.	0	0. 49	0. 41	0
1/1	1	0	0	0
	1/一	2/、	3/分组	4/审议

(a)

7/,	0	0. 01	0	0
6/chiq	0	0	0	0. 99
5/qarap	0	0	0	1
4/bölüp	0	0	1	0
3/guruppa	0	0. 01	0. 99	0
2/.	0	0. 96	0	0
1/1	1	0	0	0
	1/一	2/、	3/分组	4/审议

(b)

Figure 2: Uyghur-Chinese alignment with weighted matrix. (a) is the word level matrix, (b) means the stem level matrix.

ty. In Figure 1a, there are even four Uyghur words align to one Chinese word. After stemming, the alignment model can make a better estimation for word's co-occurrence, and then generates superior alignment results. In fact, the stem level alignment in Figure 1b gets the same result of gold standard in this example.

3 Multi-granularity Word Alignment

Formally, given a source sentence $\mathbf{f} = f_1^J = f_1, \ldots, f_j, \ldots, f_J$ and a target sentence $\mathbf{e} = e_1^I = e_1, \ldots, e_i, \ldots, e_I$, we define a link $l = (j, i)$ to exist if f_j and e_i are translations of each other. Therefore, an alignment $\mathbf{a}$ is a subset of the product of word positions:

$$\mathbf{a} \subseteq \{(j,i) : j = 1, \ldots, J; i = 1, \ldots, I\} \quad (1)$$

Traditional alignment models (Brown et al., 1993; Vogel et al., 1996) treat word alignment as a hidden process and maximize the likelihood of bi-text using the Expectation Maximization (EM) algorithm. Usually, the SMT system takes 1-best result as the final alignment result to generate translation tables, phrase tables, or syntactic transformation rules. In this work, we employ existing alignment models and produce multiple sets of alignment with different granularities, then combine them together to catch complementary information to get a better alignment result.

Specifically, we feed aligners with two granularities: surface word and stem, and stemming is only done on the source side. The former uses Uyghur surface word and Chinese surface word as the alignment factor, and the latter employs Uyghur stem and Chinese surface word.

We usually take the 1-best result from GIZA++ as the final alignment result. However, it seems not appropriate for resource-scarce language pairs due to lexical sparsity problem. And we believe that offering each link a probability might help to distinguish ”good” alignment links from ”bad” ones. Since GIZA++ could provide n-best lists alignment results with probabilities, we can estimate the link probability from it.

Formally, a weighted matrix $\mathbf{A}$ is a $J \times I$ matrix, in which each element stores a link probability $p(j, i)$ to indicate how well f_j and e_i are aligned. The link probability is estimated from n-best lists by calculating relative frequencies,

$$p(j,i) = \frac{\sum_{k=1}^{N} p(\mathbf{a}_k) \times \delta(\mathbf{a}_k, j, i)}{\sum_{k=1}^{N} p(\mathbf{a}_k)} \quad (2)$$

where

$$\delta(\mathbf{a}_k, j, i) = \begin{cases} 1 & (j,i) \in \mathbf{a}_k \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Note that N is the number of n-best list size, $p(\mathbf{a}_k)$ is the probability of kth best in the n-best lists, $\delta(\mathbf{a}_k, j, i)$ indicates whether a link (j, i) occurs in the alignment $\mathbf{a}_k$ or not. We assign 0 to any unseen alignment.

Figure 2 is an example of alignment represented as weighted matrix, which is generated on 10-best alignment lists. We can see that the link probability makes a rough corresponding to the results revealed in Figure 1. Intuitively, if combining the two matrices together, we can catch the complementary information of each other and at least enhance the link probabilities of real aligned word-pairs.

Similarly, let $\mathbf{A}_{word}$ and $\mathbf{A}_{stem}$ denote the word level and stem level weighted matrix, and they are combined to generate the final weighted matrix by linear interpolation:

$$\mathbf{A} = \lambda * \mathbf{A}_{word} + (1 - \lambda) * \mathbf{A}_{stem} \tag{4}$$

Here λ can be optimized on a tuning set. Therefore, we can use the method proposed by (Liu et al., 2009) to extract phrase pairs, or just use the weighted matrices to store alignments of different granularities, and combine them together by pruning "bad" ones with a threshold. That's to say, for each element in matrix $\mathbf{A}$, we only retain the ones whose link probability $p(j, i)$ is above a prespecified threshold.

4 Multi-granularity SMT

Since one word in agglutinative languages can generate tens of hundreds of inflected variants, standard statistical approaches, especially SMT, are apt to suffer from data spareness issues. There may exists many unknown words owing to the rich morphological features.

Previous section has revealed that stemming can reduce vocabulary size notably. Intuitively, by employing the stem-granularity rules, we may alleviate the number of unknown words. That's to say, except the use of word-granularity rules, we employ stem-granularity rules simultaneously. Perhaps it is a feasible way to incorporate the two different rule sets into log-linear model with different weights similar to (Koehn and Hoang, 2007).

5 Experiments

In this section, we first describe the experimental settings, and then verify the effect of multi-granularity word alignment and machine translation.

5.1 Experimental Setup

We conduct our experiments on Uyghur-Chinese language pair. We have collected about $120K$ bi-text sentences of Uyghur-Chinese[1] which come from government document. Table 1 shows the statistics of the training corpus. In addition, there are $1,000$ sentences in the tuning set and $1,000$ sentences in the test set, both with one reference. For the language model, we use the SRI Language Modeling Toolkit (Stolcke, 2002) to train a 4-gram model with the target side of training corpus. To measure the quality of word alignment, we manually aligned 100 parallel sentences from the training corpus of Uyghur-Chinese.

	Recall(%)	Precision(%)	AER(%)
Baseline	85.90	79.27	17.55
Replace	86.99	81.27	15.75
0.5	81.56	85.34	16.58
0.6	77.93	89.80	16.54
0.7	77.41	90.57	16.52
0.8	76.97	90.85	16.66
0.9	76.14	91.72	16.79
1.0	68.21	94.19	20.87

Table 2: Word alignment results by threshold pruning on combined weighted alignment matrix. *Baseline* is the alignment produced by GIZA++ on word level. *Replace* is on stem level.

And Moses[2] is used as our baseline SMT system. The decoding weights are optimized with minimum error rate training (MERT) (Och, 2003) to maximum word level BLEU (Papineni et al., 2002) scores.

5.2 Word Alignment

The quality of alignment is computed as appropriately refined precision and recall measures. Additionally, we also use the Alignment Error Rate (AER) (Och and Ney, 2000) which is derived from the well-known F-measure. AER requires gold alignments that are marked as 'sure' or 'probable'. Here, we don't distinguish them, so the AER is computed as:

$$Pr = \frac{|A \cap S|}{|A|}, Rc = \frac{|A \cap S|}{|S|} \tag{5}$$

$$AER = 1 - \frac{2PrRc}{Pr + Rc} \tag{6}$$

[1] http://mt.xmu.edu.cn/cwmt2011/en/index.html. Part of the corpus will be published for SMT evaluation in CWMT 2011.

[2] http://www.statmt.org/moses/

where A links are proposed and S links are gold. NULL links are not included in this evaluation.

In order to detect the performance of multi-granularity alignment, we first generate word level and stem level weighted alignment matrices respectively, then combine them by just setting $\lambda = 0.5$ as default. According to the combined weighted matrix, we choose the threshold 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 to generate word-to-word alignment without probability.

Table 2 shows the effect of threshold pruning. Because we only get 100 human-aligned sentence pairs, it is somehow hard to reflect the effect of multi-granularity alignment comprehensively. However, there are still some trends we can get from the results. The stem level alignment performs best, by relative 10.25% of alignment error rate reduction compared to baseline system. This illustrates that stemming is useful for Uyghur language processing, and stem may be used to replace corresponding word for translation to some extent.

Except for the threshold 1.0, all the others perform better than baseline. Besides, as the threshold becomes larger, the recall decreases, but the precision improves consistently. And we observe two interesting side effects on threshold pruning.

- When the threshold is up to the maximum value, say, 1.0, the precision is about 94.19%. Although the recall is low, it is really a good resource to extract bilingual dictionary.

- As we all know, high-quality bi-text seldom exists and noisy data is always there. When using the threshold pruning, some sentence pairs don't get any candidate alignments. We can see in Figure 2a that the combined score of word pair ("*qarap*", "分组") is only 0.035, which will be pruned without doubt in our way. In fact, we have checked that none of these sentences is bi-text. Hence, maybe this is a way to filter noisy data from parallel corpus.

5.3 Statistical Machine Translation

According to Table 3, we can see that it always performs better when using stem as the basic translation unit. In order to generate the weighted alignment matrix (WAM), we conduct on different N-best lists produced by GIZA++. And the 100-best is better than 10-best both on word level and stem level, but the difference is not significant. When we take the WAM produced by 100-best of GIZA++ on stem level, it gains +1.42% BLEU score compared to baseline.

	Tuning(%)	Test(%)
word-level	41.50	39.62(+0.0)
stem-level	42.67	40.24(+0.62)
wam-10-word-level	41.42	39.68(+0.02)
wam-100-word-level	41.60	39.96(+0.34)
wam-10-stem-level	42.81	40.94(+1.32)
wam-100-stem-level	42.84	41.04(+1.42)

Table 3: Some baseline SMT results. *wam*-based methods are similar to (Liu et al., 2009).

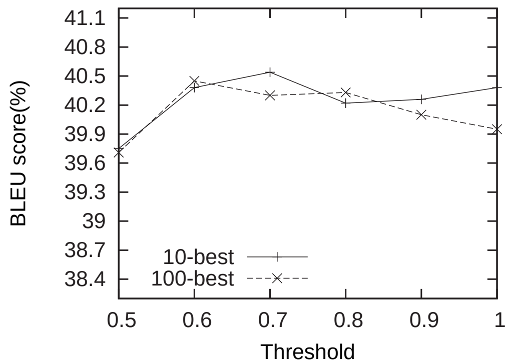


Figure 3: Translation results of Uyghur-Chinese based on pruning method.

5.3.1 Wam-based method

We conduct the threshold pruning method on the combined weighted alignment matrix. Here we only take the surface word as the basic translation unit. The results in Figure 3 show that: whatever the threshold is, it always gets improvements compared to the baseline system. And the combined WAM performs similar with different N-best lists generated by GIZA++.

Besides, we also checked the influence on translation of different value λ. Intuitively, λ is related to the quality of alignment. Therefore, we set

$$\lambda = 10^{\alpha * \frac{1-AER_{word}}{2-AER_{word}-AER_{stem}}}, \tag{7}$$

which can clearly distinguish the quality of alignment with different α.

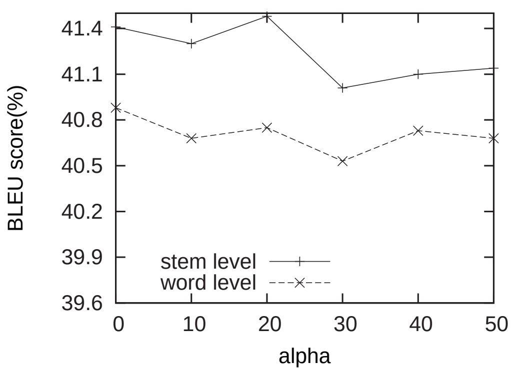


Figure 4: Translation results of Uyghur-Chinese with different λ.

As we can see from Figure 4, stem level translation is always better than word level, and they have similar changes with different λ. With different λ, we get different translation results. When we set α = 20, which means λ is 0.304, we get the best translation performance on stem level. The BLEU score is 0.4148, which is $+1.86\%$ compared to baseline system.

5.3.2 Multi-granularity Decoding

For the multi-granularity SMT, we employ the frame of factored models (Koehn and Hoang, 2007) of Moses. When we translate from Uyghur to Chinese, we just keep two factors (surface word and stem) on the source side. And we get 0.4115 BLEU score with Moses' factored models. Here if we extract the word level and stem level phrase pairs from the combined weighted alignment matrices, we obtain obvious better performance with BLEU score 0.4172, which is gained $+2.46\%$ compared to baseline. And it is also better than the original factored models.

6 Related Work

Word alignment, as an essential step in most phrase and syntax based statistical machine translation systems, has received a significant amount of research over the years. Some are focused on the improvement of word alignment models, notably in (Moore, 2005; Liu et al., 2010; Riesa and Marcu, 2010; Saers et al., 2010). Others are trying to incorporate morpho-syntactic knowledge (Xiang et al., 2010; Carpuat et al., 2010; Luong and Kan, 2010) which makes it easier to determine corresponding words directly.

As for the translation task with respect to agglutinative languages, we often subject to the problem of word sparsity. The reduction of sparsity can be achieved by increasing the number of training data or via morphological analysis. Nakov and Ng (2009) propose a method for improving SMT of resource-scarce languages by exploiting their similarity to resource-rich ones. Nießen and Ney (2004) use morphological decomposition to get better alignments. (Lee, 2004) changes the word segmentation of Arabic to induce morphological and syntactic symmetry between parallel sentences. Some are take the morphological analysis as a preprocessing step for word alignments. Elming and Habash (2007) first tokenize words into smaller units to align, and then the alignments are mapped back to the original word form. Similarly, Carpuat et al. (2010) propose to reorder post-verbal subject (VS) constructions of Arabic sentences into SV order for word alignment only, and the phrase extraction and decoding are performed on the original word order.

Moreover, there is still some work on alignment combination. Koehn et al. (2003) combine the alignments from two different directions, source-to-target and target-to-source. Ayan and Dorr (2006) propose a maximum entropy approach to combine multiple alignment from different models based on a set of linguistic and alignment features. Xiang et al. (2010) generate multiple sets of diversified alignments based on different motivations and then combine them according to confidence scores (Huang, 2009). Zhang and Sumita (2007) use English lemmas in training which improves the quality of word alignment and yield better translation performance.

The difference between our work and above is: we generate multi-granularity word alignments, and propose different methods to combine them; and to the best of our knowledge, this is the first time that statistical method of translating Uyghur into Chinese is reported.

7 Conclusions and Future Work

This paper proposes a novel strategy, multi-granularity integration, to optimize the word alignment and decoding for agglutinative language translation. For word alignment, we perform a series of alignment tasks with different lexical granularities. In each task, multiple alignment results are combined by assigning each inter-translatable lexical pair an alignment probability. While for decoding, the translation rules of different granularities are also integrated in a weighted competitive manner, so as to integrate the coverage of fine-grained stem/affix rules and the accuracy of large-grained word rules. End-to-end evaluation shows that the proposed methods can improve not only the quality of statistical word alignment but also the performance of statistical machine translation.

However, there is still some future work to do. The combination approach itself is not limited to any specific language. It provides a general framework for improving quality of word alignment, especially for agglutinative languages. We plan to extend our methods to other agglutinative languages, such as Korean, Japanese.

Acknowledgments

This work was supported by National Natural Science Foundation of China Contract 60736014 and 60873167. We thank the anonymous reviewers for their insightful comments.

References

Necip Fazil Ayan and Bonnie J. Dorr. 2006. A maximum entropy approach to combining word alignments. In *NAACL '06*, pages 96–103, June.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.

Marine Carpuat, Yuval Marton, and Nizar Habash. 2010. Improving arabic-to-english statistical machine translation by reordering post-verbal subjects for alignment. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 178–183, July.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *ACL '05*, pages 263–270.

Jakob Elming and Nizar Habash. 2007. Combination of statistical word alignments based on multiple preprocessing schemes. In *HLT-NAACL '07, Short Papers*, pages 25–28, April.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of COLING/ACL 2006*, pages 961–968.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 49–52, New York City, USA, June. Association for Computational Linguistics.

Zhongjun He, Qun Liu, and Shouxun Lin. 2008. Partial matching strategy for phrase-based statistical machine translation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 161–164, Columbus, Ohio, June.

Fei Huang. 2009. Confidence measure for word alignment. In *ACL-AFNLP '09*, pages 932–940, August.

Wenbin Jiang, Jinxing Wu, Qing Chang, Lili Zhao, Qun Liu, and Nasanurtu. 2010. Generative statistical modeling for mongolian lexical analysis. In *3 (rd).CN NMYS-NLP, 2 (nd).CN MLKD*.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *EMNLP-CoNLL '07*, pages 868–876, June.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03*, pages 48–54.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *HLT-NAACL 2004: Short Papers*, pages 57–60, May 2 - May 7.

Yang Liu, Tian Xia, Xinyan Xiao, and Qun Liu. 2009. Weighted alignment matrices for statistical machine translation. In *EMNLP '09*, pages 1017–1026, August.

Yang Liu, Qun Liu, and Shouxun Lin. 2010. Discriminative word alignment by linear modeling. *Comput. Linguist.*, 36:303–339.

Minh-Thang Luong and Min-Yen Kan. 2010. Enhancing morphological alignment for translating highly inflected languages. In *Coling 2010*, pages 743–751, August.

Robert C. Moore. 2005. A discriminative framework for bilingual word alignment. In *EMNLP '05*, pages 81–88, October.

Preslav Nakov and Hwee Tou Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *EMNLP '09*, pages 1358–1367, August.

Sonja Nießen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Comput. Linguist.*, 30:181–204, June.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *ACL '00*, pages 440–447.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL '02*, pages 311–318.

Jason Riesa and Daniel Marcu. 2010. Hierarchical search for word alignment. In *ACL '10*, pages 157–166, July.

Markus Saers, Joakim Nivre, and Dekai Wu. 2010. Word alignment with stochastic bracketing linear inversion transduction grammar. In *HLT-NAACL '10*, pages 341–344, June.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *ICSLP 2002*, pages 311–318.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *COLING 1996 - Volume 2*, pages 836–841.

Bing Xiang, Yonggang Deng, and Bowen Zhou. 2010. Diversify and combine: Improving word alignment for machine translation on low-resource languages. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 22–26, July.

Ruiqiang Zhang and Eiichiro Sumita. 2007. Boosting statistical machine translation by lemmatization and linear interpolation. In *ACL '07*, pages 181–184, June.