Combining Multiple Alignments to Improve Machine Translation

Zhaopeng Tu[1] *Yang Liu*[2] *Yifan He*[3] *Josef van Genabith*[4]
Qun Liu[1,4] *Shouxun Lin*[1]

(1) Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, China
{tuzhaopeng, liuqun, sxlin}@ict.ac.cn

(2) Department of Computer Science and Technology, State Key Lab on Intelligent Technology and
Systems, National Lab for Information Science and Technology, Tsinghua University, China
liuyang2011@tsinghua.edu.cn

(3) Computer Science Department, New York University, USA
yhe@cs.nyu.edu

(4) Centre for Next Generation Localisation, School of Computing, Dublin City University, Ireland
josef@computing.dcu.ie

Abstract

Word alignment is a critical component of machine translation systems. Various methods for word alignment have been proposed, and different models can produce significantly different outputs. To exploit the advantages of different models, we propose three ways to combine multiple alignments for machine translation: (1) *alignment selection*, a novel method to select an alignment with the least expected loss from multiple alignments within the minimum Bayes risk framework; (2) *alignment refinement*, an improved algorithm to refine multiple alignments into a new alignment that favors the consensus of various models; (3) *alignment compaction*, a compact representation that encodes all alignments generated by different methods (including (1) and (2) above) using a novel calculation of link probabilities. Experiments show that our approach not only improves the alignment quality, but also significantly improves translation performance by up to 1.96 BLEU points over single best alignments, and 1.28 points over merging rules extracted from multiple alignments individually.

KEYWORDS: alignment combination, minimum Bayes risk, alignment refinement, weighted alignment matrix.

Alignments	GIZA++	Berkeley	Vigne
GIZA++	–	70.29%	75.17%
Berkeley	70.29%	–	73.25%
Vigne	75.17%	73.25%	–

Table 1: Agreement of alignment links between different alignment models. Here we use three different alignment models: GIZA++ (Och and Ney, 2003), the unsupervised Berkeley aligner (Liang et al., 2006), and a discriminative aligner Vigne (Liu et al., 2010).

1 Introduction

Word alignment is a preliminary step for statistical machine translation (SMT). Most SMT systems, not only phrase-based models (Och and Ney, 2004; Koehn et al., 2003; Chiang, 2005; Xiong et al., 2006), but also syntax-based models (Galley et al., 2006; Shen et al., 2008; Liu et al., 2006; Huang et al., 2006), rely heavily on word-aligned bilingual corpora.

Various methods for word alignment, including generative methods (Brown et al., 1993; Vogel et al., 1996; Liang et al., 2006) and discriminative methods (Moore et al., 2006; Taskar et al., 2005; Blunsom and Cohn, 2006; Liu et al., 2010), have been proposed in the literature. Different models produce significantly different alignments. [1] Table 1 shows the *agreement* between each pair of alignments on 1.5M Chinese-English parallel sentence pairs. Here *agreement* is computed by using one alignment model's output as a gold standard to evaluate the other alignment model's output in terms of F1 score (Xiao et al., 2010). The higher the agreement score is, the more similar two alignments are. Table 1 shows that the agreement scores are always below 76%.

Therefore, it is natural to combine multiple alignments to improve both alignment quality and translation quality. In this paper, we propose three ways to exploit multiple alignments for machine translation: alignment selection, refinement and compaction. Alignment selection chooses high quality alignments while refinement generates new and more reliable alignments. Alignment compaction encodes multiple possible alignments. We show that these methods work well together: alignment refinement e.g. offers high quality alignment choices, that can be exploited by alignment compaction.

2 Related Work

Our research builds on previous work in the field of minimum Bayes risk (MBR) decision, system combination and model compaction. MBR decision aims to find the candidate hypothesis that has the least expected loss under a probability model when the true reference is not known (Brickel and Doksum, 1977). Diverse loss functions have been described by using different evaluation criteria for loss calculation, e.g. edit distance and sentence-level BLEU in SMT (Kumar and Byrne, 2004; Tromble et al., 2008; González-Rubio et al., 2011). In our work, we select an alignment within the MBR framework using a number of loss functions at both alignment and phrase levels.

System combination, the process which integrates fragment outputs from multiple systems, has produced substantial improvements in many natural language processing tasks, including parsing (Henderson and Brill, 1999; Sagae and Lavie, 2006; Fossum and Knight, 2009), word segmentation (Sun and Wan, 2012) and machine translation (Rosti et al., 2007; He et al.,

[1] These alignments have equivalent qualities compared to a true gold standard (see in Table 2).

2008; Feng et al., 2009), just to name a few. Alignment combination has also been explored previously (Och and Ney, 2003; Koehn et al., 2003; Ayan et al., 2005; DeNero and Macherey, 2011). We draw inspiration from (Och and Ney, 2003; Koehn et al., 2003) but our technique differs from previous work in that (1) they require exactly two bidirectional alignments while our approach can use an arbitrary number of alignments; (2) we take into account the occurrences of potential links, which turns out to be important.

Previous research has demonstrated that compact representations can produce improved results by offering more alternatives, e.g. using forests over 1-best trees (Mi and Huang, 2008; Tu et al., 2010), word lattices over 1-best segmentations (Dyer et al., 2008), and weighted alignment matrices (WAMs) over 1-best alignments (Liu et al., 2009; Tu et al., 2011). Instead of using k-best alignments from the same model, as in (Liu et al., 2009; Tu et al., 2011), here we construct WAMs from multiple alignments generated by different models (including MBR-based and refined models). As the alignment probabilities are generally incomparable between different alignment models, we propose a novel calculation of link probabilities in WAMs.

3 Approach

3.1 Alignment Selection

Alignment selection refers to selecting one alignment from multiple alignments using minimum Bayes risk. If the reference alignment a was known, we could measure each alignment a_i using the loss function $\mathscr{L}(a_i, a)$. In the MBR framework, although the true reference alignment is unknown, we assume that the individual alignment models' output forms a reasonable distribution over possible reference alignments. The MBR decision aims to find the candidate alignment that has the least expected loss under the distribution (Brickel and Doksum, 1977).

3.1.1 MBR Decision

MBR decision has the following form:

$$\hat{a} = \arg\min_{a_i \in A} \mathscr{R}(a_i) = \arg\min_{a_i \in A} \sum_{a_j \in A} \mathscr{L}(a_i, a_j) \cdot p(a_j|f, e) \tag{1}$$

where $\mathscr{R}(a_i)$ denotes the Bayes risk of candidate alignment a_i under loss function $\mathscr{L}$, A indicates the set of alignments generated by different models. In general, for different alignment models, the probabilities $p(a|f, e)$ are not directly comparable. For simplicity, in our work below we assume that they are in fact comparable and have the same value. [2]

3.1.2 Loss Functions

The loss function $\mathscr{L}(a_i, a_j)$ is used to measure the quality of alignments. Here we introduce a set of metrics for the evaluation of alignments at both alignment and phrase levels.

AER

Alignment error rate (Och and Ney, 2003) has been used as the official evaluation criterion in most alignment shared tasks (Liu et al., 2009). AER scores are given by:

$$AER(S, P, A) = 1 - (|A \cap S| + |A \cap P|)/(|A| + |S|) \tag{2}$$

[2] Alignment probabilities can be set empirically based on (expected overall) performance (Fossum and Knight, 2009), or uniformly without any bias (Xiao et al., 2010; Duan et al., 2010). We tried a few other settings and found them to be less effective.

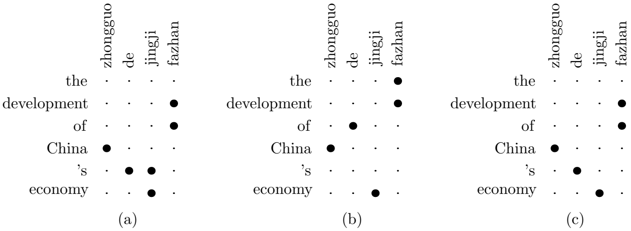


Figure 1: (a) Alignment of a sentence pair generated by GIZA++ (a_1), (b) alignment of the same sentence by Berkeley aligner (a_2), (c) another alignment by Vigne (a_3).

where S and P are sets of sure and possible links in a hand-aligned reference alignment respectively, and A is a candidate alignment. Note that S is a subset of P: $S \subseteq P$. As there is no reference alignment that is hand-aligned by human experts in our work, we cannot distinguish sure links from possible links. Therefore, we regard all links to be sure links: $S = P$. With this, the AER score is calculated by:

$$AER(a_i, a_j) = 1 - (2 \times |a_i \cap a_j|)/(|a_i| + |a_j|) \tag{3}$$

CPER

Although widely used, AER is criticized for correlating poorly with translation performance (Ayan and Dorr, 2006; Fraser and Marcu, 2007). Therefore, Ayan and Dorr (2006) have proposed *constituent phrase error rate* (CPER) for evaluating word alignments at the phrase level instead of the alignment level. CPER can be computed as:

$$CPER(a_i, a_j) = 1 - (2 \times |P_{a_i} \cap P_{a_j}|)/(|P_{a_i}| + |P_{a_j}|) \tag{4}$$

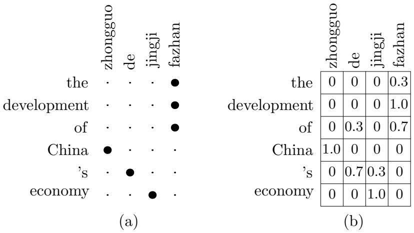
where P_a denotes the set of phrases that are consistent with a given alignment a. Compared with AER, CPER penalizes dissimilar alignment links more heavily. As a dissimilar link reduces the number of intersected links of two alignments by 1 in AER, it might lead to more than one different phrase pair added to or removed from the set of phrases (Ayan and Dorr, 2006).

CHER

As CPER evaluates word alignments in the context of phrase-based MT, we propose a similar metric called *constituent hierarchical-phrase error rate* (CHER) for hierarchical-phrase models. The difference between them is that we use H_a instead of P_a, where H_a denotes the hierarchical phrases extracted. Hierarchical phrases are more sensitive to word alignments because they are sensitive to inside (i.e. subtracted) phrases.

3.2 Alignment Refinement

Alignment refinement refers to extracting parts of multiple alignments and constructing a new alignment instead of selecting the best one from existing alignments. A simple way to refine multiple alignments is to employ their intersection or union. However, using intersection will result in a high-precision but low-recall alignment, while using union will result in a high-recall but low-precision alignment. Koehn et al. (2003) show performance improvements by finding a balance between the intersection and union with the *grow-diag-final* algorithm.



	zhongguo	de	jingji	fazhan
the	·	·	·	●
development	·	·	·	●
of	·	·	·	●
China	●	·	·	·
’s	·	●	·	·
economy	·	·	●	·

(a)

	zhongguo	de	jingji	fazhan
the	0	0	0	0.3
development	0	0	0	1.0
of	0	0.3	0	0.7
China	1.0	0	0	0
’s	0	0.7	0.3	0
economy	0	0	1.0	0

(b)

Figure 2: (a) The refined alignment generated from multiple alignments in Figure 1, (b) the resulting weighted alignment matrix that samples the same alignments, where the number in the cells are the probabilities of the corresponding link.

Unfortunately, this algorithm cannot be applied to our approach. This is because the *grow-diag-final* algorithm requires exactly two bidirectional alignments, while we would use more than two alignments. Therefore, we propose a variation of the *grow-diag-final* algorithm named *grow-diag-final-rank* adapted for multiple alignments. The difference between the two algorithms is that we take into account the occurrences of *conflicting links*. Conflicting links refer to triples <l_i, l_j, l_k>, in which l_i and l_j are the links that share the same source side, and l_j and l_k share the same target side. For example, the triple < (*de*, *’s*), (*de*, *of*), (*fazhan*, *of*)> is conflicting because the first two share the same source side while the latter two share the same target side.

Alignment refinement chooses the links with the most occurrences when there are conflicting links. Intuitively, our approach is motivated by the following observation: the links that occur more often in different alignments frequently have a higher confidence than those that occur less often. Our algorithm favors the links that occur frequently. As an example, consider the conflicting links < (*de*, *’s*), (*de*, *of*), (*fazhan*, *of*)>: without considering the number of their occurrences, we would retain the first two links if we run grow-diag-final greedily. In contrast, considering that the links (*de*, *’s*) and (*fazhan*, *of*) occur twice while (*de*, *of*) only occurs once, we prefer to retain (*de*, *’s*) and (*fazhan*, *of*). Figure 2(a) shows the refined alignment generated from the three alignments in Figure 1 using the *grow-diag-final-rank* algorithm.

3.3 Alignment Compaction

Given the original alignments and the alignments generated by alignment refinement, it is quite natural to try to encode them in a compact representation. In this paper, we use weighted alignment matrices for this purpose. A weighted alignment matrix (Liu et al., 2009) is a matrix to encode the probabilities of k-best alignments of the same sentence pair. Each element in the matrix stores a link probability which is estimated from a k-best list.

$$p_m(j,i) = \frac{\sum_{k=1}^{K} p(a_k|f,e) \cdot \delta(a_k,j,i)}{\sum_{k=1}^{K} p(a_k|f,e)} \tag{5}$$

where

$$\delta(a_k,j,i) = \begin{cases} 1 & (j,i) \in a_k \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Here $a_k \in \mathscr{K}$ is a k-best list, $p(a_k|f,e)$ is the probability of an alignment a_k in the k-best list. Intuitively, a higher link probability $p_m(j,i)$ indicates high agreement between different alignments, thereby high quality.

(Liu et al., 2009; Tu et al., 2011) have shown that WAMs yield encouraging results by making good use of k-best alignments from a single alignment model. Unlike in this previous work, in our approach we construct WAMs from alignments generated by different models (including MBR-based and refined models). In a k-best list, each alignment is weighted using their probabilities since they are obtained from the same model, and a higher weight denotes that the alignment model has higher confidence in the output. In contrast, the alignments in our work are generated by different models and their probabilities are generally incomparable. As noted above, we assume that all the alignments have the same probabilities. Then, we obtain:

$$p_m(j,i) = \frac{\sum_{k=1}^{N} \delta(a_k, j, i)}{N} \tag{7}$$

Figure 2(b) shows the WAM that captures the three alignments in Figure 1.[3]

We then follow (Tu et al., 2011) to extract hierarchical phrases from WAM and calculate their translation and lexical probabilities. Instead of extracting phrase pairs that respect the word alignment, Tu et al. (2011) enumerate all potential phrase pairs and calculate their fractional counts. As they soften the alignment consistency constraint, there exists a massive number of phrase pairs extracted from the training corpus. To maintain a reasonable phrase table size, they discard any phrase pair that has a fractional count lower than a threshold t. For further details, see (Tu et al., 2011).

4 Experiments

4.1 Setup

We carry out our experiments using a reimplementation of the hierarchical phrase-based system (Chiang, 2005) on the NIST Chinese-English translation tasks. Our training data contains 1.5M sentence pairs from LDC dataset.[4] We train a 4-gram language model on the Xinhua portion of the GIGAWORD corpus using the SRI Language Toolkit (Stolcke, 2002) with modified Kneser-Ney Smoothing (Kneser and Ney, 1995). We use minimum error rate training (Och, 2003) to optimize the feature weights on the MT02 testset, and test on the MT03/04/05 testsets. For evaluation, case-insensitive NIST BLEU (Papineni et al., 2002) is used to measure translation performance.

Three alignment models are chosen for our experiments with default settings: GIZA++ (Och and Ney, 2003), the unsupervised Berkeley aligner (Liang et al., 2006), and the linear modeling alignment Vigne (Liu et al., 2010). We use the three baseline alignments to select MBR alignments and to generate a refined alignment. We use all three baseline alignments, as well as all of the MBR and refined alignments in the WAM-based compaction approach. When extracting rules from WAM, we follow (Tu et al., 2011) to set the pruning threshold t=0.5.

[3] In practice, alignment compaction encodes both baseline alignments and the new alignments in Section 3.1 and 3.2.

[4] The corpus includes LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06.

Alignments	AER	BAER	CPER	CHER
GIZA++	22.50	27.92	24.11	33.23
Berkeley	21.11	26.41	23.35	34.44
Vigne	19.13	24.05	23.54	34.02
Selection$_{AER}$	**17.93**	**23.29**	22.10	31.47
Selection$_{CPER}$	18.32	23.72	**21.53**	30.56
Selection$_{CHER}$	18.52	23.93	21.68	30.84
Refinement	18.79	24.43	**21.50**	**30.31**

Table 2: Evaluation of alignment quality. Here "Selection$_{\mathscr{L}}$" indicates the alignment selected from multiple single alignments using MBR decision under the loss function $\mathscr{L}$ (e.g. AER, CPER and CHER). For all metrics, the lower the score is, the better the alignment quality is.

4.2 Evaluation of Alignment Quality

In this section, we investigate the quality of different alignments on the Chinese-English language pair data. We annotated 1007 sentences with annotations that distinguish between sure and possible links.[5] We used 502 sentences as the tuning set, and 505 sentences as the test set. We run GIZA++ and the Berkeley aligner on the training corpus as well as the test set. We tune the feature weights of Vigne on the tuning set using AER as the optimization criterion. We evaluate alignments in terms of AER, CPER and CHER as described in Section 3.1.2. Inspired by Fraser and Marcu (2007), we also employ a new metric called **balanced AER** (BAER) that considers only the sure links in the reference alignments:

$$BAER(S,A) = 1 - (2 \times |A \cap S|)/(|A| + |S|) \tag{8}$$

For all metrics, lower score indicates better alignment quality.

Table 2 lists the alignment quality results for different alignment strategies. We find that both selection and refinement methods outperform single alignments at all metrics, indicating that our methods improve the quality of alignment in a certain way. One finding is that the selection method usually achieves the best score at the metric it uses as loss function. For example, the selection method using AER as loss function outperforms other alignments at the AER and BAER metrics while underperforming at other metrics. This is intuitive, since the method always selects the alignment with the minimum expected loss under the metric.

4.3 Evaluation of Translation Quality

Table 3 summaries the results of translation performance with different alignment methods.

- **Baseline results**. We have three baseline systems: GIZA++, Berkeley and Vigne. The results show that GIZA++ achieves the best performance among the baseline systems. Therefore, we compare our methods with GIZA++ system in the following analysis.
- **Rule Merging**. Different alignments generally result in very different sets of hierarchical rules. As one would expect, merging them outperforms using any of them individually through enlarging the rule coverage. Experimental results show that merging rules indeed outperforms using single best alignments, at the cost of a much larger rule table.

[5] available at http://nlp.ict.ac.cn/~tuzhaopeng/.

Alignments	Links	Rules	DEV	MT03	MT04	MT05	Avg.
GIZA++	45.4M	143M	35.07	33.11	35.06	32.98	33.72
Berkeley	33.7M	270M	34.72	32.64	34.93	32.58	33.38
Vigne	35.6M	140M	34.64	33.16	34.29	32.45	33.30
Rule Merging	–	553M	35.55	34.12**	35.88**	33.66*	34.55
Inter	24.5M	178M	34.10	32.35	34.17	32.47	33.00
Union	55.6M	94M	34.83	33.42	35.04	33.05	33.84
Selection$_{AER}$	37.9M	175M	35.35	33.65**	35.82**	33.56*	34.34
Selection$_{CPER}$	38.9M	187M	35.36	34.21**	36.05**	33.71**	34.66
Selection$_{CHER}$	39.1M	182M	35.71	34.16**	35.88**	33.94**	34.66
Refinement	45.5M	210M	35.44	33.81**	35.98**	33.95**	34.58
Compaction	55.6M	319M	**36.64**	**35.01****	**36.81****	**34.94****	**35.59**

Table 3: Evaluation of translation quality. "Links" denotes the number of links in the alignment and "Rules" denotes the number of rules (Chiang, 2005) extracted from the corresponding alignment. "Avg." is the average BLEU score on the three test sets. Significance tests are done against GIZA++ on test sets following the *sign-test* approach (Collins et al., 2005), and "**" and "*" denote p-value less than 0.01 and 0.05, respectively. Furthermore, Compaction is significantly better than Rule Merging for p-value less than 0.01 on all test sets.

- **Alignment Selection**. Concerning selection methods, the results show that using loss functions at phrase level (i.e. CPER and CHER) outperforms loss function at alignment level (i.e AER). One possible reason is that CPER and CHER relate more tightly to the translation performance, because they care about the phrases which are used directly in machine translation. In brief, using selection methods with different loss functions improves translation performance in BLEU score by up to 0.92 points on average.

- **Alignment Refinement**. Table 3 shows that simply using the intersection (Inter) or union (Union) does not achieve any improvement. This is in accord with intuition, because intersection discards many useful links while union includes many incorrect links. By contrast, alignment refinement finds a good balance between them, and achieves significant improvement in BLEU score ranging from 0.70 to 0.97 points.

- **Alignment Compaction**. Alignment compaction encodes all alignments and achieves the best result, which improves BLEU scores by between 1.75 and 1.96 points. Compared with rule merging, alignment combination produces substantial improvements in both translation performance and rule table size.

5 Conclusion

In this paper, we have presented three simple and effective methods to make use of multiple alignments. First, we select the alignments with minimum Bayes risk using different loss functions at both alignment and phrase levels. Then, we refine multiple alignments using an improved *grow-diag-final-rank* algorithm that considers the occurrences of alignment links. Finally, we use a compact representation named weighted alignment matrix to represent all alignments (including MBR-based and refined alignments) and propose a novel calculation of link probabilities. Experimental results show that our method not only improves the alignment quality, but also significantly improves translation performance over both single best alignments and merging rules extracted from different single alignments individually.

Acknowledgement

The authors were supported by 863 State Key Project No. 2011AA01A207, 863 Project No. 2012AA011102 and Science Foundation Ireland (Grant No. 07/CE/I1142). We thank the anonymous reviewers for their insightful comments.

References

Ayan, N. F. and Dorr, B. J. (2006). Going beyond aer: An extensive analysis of word alignments and their impact on mt. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Sydney, Australia. Association for Computational Linguistics.

Ayan, N. F., Dorr, B. J., and Monz, C. (2005). Neuralign: Combining word alignments using neural networks. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 65–72, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Blunsom, P. and Cohn, T. (2006). Discriminative word alignment with conditional random fields. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 65–72, Sydney, Australia. Association for Computational Linguistics.

Brickel, P. J. and Doksum, K. A. (1977). Mathematical statistics: basic ideas and selected topics.

Brown, P. E., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics.

Collins, M., Koehn, P., and Kučerová, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 531–540. Association for Computational Linguistics.

DeNero, J. and Macherey, K. (2011). Model-based aligner combination using dual decomposition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 420–429, Portland, Oregon, USA. Association for Computational Linguistics.

Duan, N., Li, M., Zhang, D., and Zhou, M. (2010). Mixture model-based minimum bayes risk decoding using multiple machine translation systems. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 313–321, Beijing, China. International Committee on Computational Linguistics.

Dyer, C., Muresan, S., and Resnik, P. (2008). Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*, pages 1012–1020, Columbus, Ohio. Association for Computational Linguistics.

Feng, Y., Liu, Y., Mi, H., Liu, Q., and Lü, Y. (2009). Lattice-based system combination for statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1105–1113, Singapore. Association for Computational Linguistics.

Fossum, V. and Knight, K. (2009). Combining constituent parsers. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 253–256, Boulder, Colorado. Association for Computational Linguistics.

Fraser, A. and Marcu, D. (2007). Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W., and Thayer, I. (2006). Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 961–968, Sydney, Australia. Association for Computational Linguistics.

González-Rubio, J., Juan, A., and Casacuberta, F. (2011). Minimum bayes-risk system combination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1268–1277, Portland, Oregon, USA. Association for Computational Linguistics.

He, X., Yang, M., Gao, J., Nguyen, P., and Moore, R. (2008). Indirect-hmm-based hypothesis alignment for computing outputs from machine translation systems. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 98–107, Honolulu, Hawaii. Association for Computational Linguistics.

Henderson, J. C. and Brill, E. (1999). Exploiting diversity in natural language processing: Combining parsers. In *Proceedings of the Fourth Conference on Empirical Methods in Natural Language Processing*, pages 187–194.

Huang, L., Knight, K., and Joshi, A. (2006). Statistical syntax-directed translation with extended domain of locality. In *Proceedings of AMTA*, pages 66–73. Citeseer.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *ICASSP IEEE INT CONF ACOUST SPEECH SIGNAL PROCESS PROC*, volume 1, pages 181–184.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Kumar, S. and Byrne, W. (2004). Minimum bayes-risk decoding for statistical machine translation. In Susan Dumais, D. M. and Roukos, S., editors, *HLT-NAACL 2004: Main Proceedings*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Liang, P., Taskar, B., and Klein, D. (2006). Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York City, USA. Association for Computational Linguistics.

Liu, Y., Liu, Q., and Lin, S. (2006). Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 609–616, Sydney, Australia. Association for Computational Linguistics.

Liu, Y., Liu, Q., and Lin, S. (2010). Discriminative word alignment by linear modeling. *Computational Linguistics*, 36(3):303–339.

Liu, Y., Xia, T., Xiao, X., and Liu, Q. (2009). Weighted alignment matrices for statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1017–1026, Singapore. Association for Computational Linguistics.

Mi, H. and Huang, L. (2008). Forest-based translation rule extraction. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 206–214, Honolulu, Hawaii. Association for Computational Linguistics.

Moore, R. C., Yih, W.-t., and Bode, A. (2006). Improved discriminative bilingual word alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 513–520, Sydney, Australia. Association for Computational Linguistics.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

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

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rosti, A.-V., Ayan, N. F., Xiang, B., Matsoukas, S., Schwartz, R., and Dorr, B. (2007). Combining outputs from multiple machine translation systems. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 228–235, Rochester, New York. Association for Computational Linguistics.

Sagae, K. and Lavie, A. (2006). Parser combination by reparsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 129–132, New York City, USA. Association for Computational Linguistics.

Shen, L., Xu, J., and Weischedel, R. (2008). A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of ACL-08: HLT*, pages 577–585, Columbus, Ohio. Association for Computational Linguistics.

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. In *Proceedings of Seventh International Conference on Spoken Language Processing*, volume 3, pages 901–904. Citeseer.

Sun, W. and Wan, X. (2012). Reducing approximation and estimation errors for chinese lexical processing with heterogeneous annotations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 232–241, Jeju Island, Korea. Association for Computational Linguistics.

Taskar, B., Simon, L.-J., and Dan, K. (2005). A discriminative matching approach to word alignment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 73–80, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Tromble, R., Kumar, S., Och, F., and Macherey, W. (2008). Lattice Minimum Bayes-Risk decoding for statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 620–629, Honolulu, Hawaii. Association for Computational Linguistics.

Tu, Z., Liu, Y., Hwang, Y.-S., Liu, Q., and Lin, S. (2010). Dependency forest for statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1092–1100, Beijing, China. International Committee on Computational Linguistics.

Tu, Z., Liu, Y., Liu, Q., and Lin, S. (2011). Extracting Hierarchical Rules from a Weighted Alignment Matrix. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1294–1303, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Vogel, S., Ney, H., and Tillmann, C. (1996). Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, pages 836–841. Association for Computational Linguistics.

Xiao, T., Zhu, J., Zhang, H., and Zhu, M. (2010). An empirical study of translation rule extraction with multiple parsers. In *Coling 2010: Posters*, pages 1345–1353, Beijing, China. International Committee on Computational Linguistics.

Xiong, D., Liu, Q., and Lin, S. (2006). Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 521–528, Sydney, Australia. Association for Computational Linguistics.