A Comparative Study of Hypothesis Alignment and its Improvement for Machine Translation System Combination

Boxing Chen*, Min Zhang, Haizhou Li and Aiti Aw

Institute for Infocomm Research
1 Fusionopolis Way, 138632 Singapore
{bxchen, mzhang, hli, aaiti}@i2r.a-star.edu.sg

Abstract

Recently confusion network decoding shows the best performance in combining outputs from multiple machine translation (MT) systems. However, overcoming different word orders presented in multiple MT systems during hypothesis alignment still remains the biggest challenge to confusion network-based MT system combination. In this paper, we compare four commonly used word alignment methods, namely GIZA++, TER, CLA and IHMM, for hypothesis alignment. Then we propose a method to build the confusion network from intersection word alignment, which utilizes both direct and inverse word alignment between the backbone and hypothesis to improve the reliability of hypothesis alignment. Experimental results demonstrate that the intersection word alignment yields consistent performance improvement for all four word alignment methods on both Chinese-to-English spoken and written language tasks.

1 Introduction

Machine translation (MT) system combination technique leverages on multiple MT systems to achieve better performance by combining their outputs. Confusion network based system combination for machine translation has shown promising advantage compared with other techniques based system combination, such as sentence level hypothesis selection by voting and source sentence re-decoding using the phrases or translation models that are learned from the source sentences and target hypotheses pairs (Rosti et al., 2007a; Huang and Papineni, 2007).

In general, the confusion network based system combination method for MT consists of four steps: 1) Backbone selection: to select a backbone (also called "skeleton") from all hypotheses. The backbone defines the word orders of the final translation. 2) Hypothesis alignment: to build word-alignment between backbone and each hypothesis. 3) Confusion network construction: to build a confusion network based on hypothesis alignments. 4) Confusion network decoding: to decode the best translation from a confusion network. Among the four steps, the hypothesis alignment presents the biggest challenge to the method due to the varying word orders between outputs from different MT systems (Rosti et al, 2007). Many techniques have been studied to address this issue. Bangalore et al. (2001) used the edit distance alignment algorithm which is extended to multiple strings to build confusion network, it only allows monotonic alignment. Jayaraman and Lavie (2005) proposed a heuristic-based matching algorithm which allows non-monotonic alignments to align the words between the hypotheses. More recently, Matusov et al. (2006, 2008) used GIZA++ to produce word alignment for hypotheses pairs. Sim et al. (2007), Rosti et al. (2007a), and Rosti et al. (2007b) used minimum Translation Error Rate (TER) (Snover et al., 2006) alignment to build the confusion network. Rosti et al. (2008) extended TER algorithm which allows a confusion network as the reference to compute word alignment. Karakos et al. (2008) used ITG-based method for hypothesis alignment. Chen et al. (2008) used Competitive Linking Algorithm (CLA) (Melamed, 2000) to align the words to construct confusion network. Ayan et al. (2008) proposed to improve alignment of hypotheses using synonyms as found in WordNet (Fellbaum, 1998) and a two-pass alignment strategy based on TER word alignment approach. He et al. (2008) proposed an IHMM-based word alignment method which the parameters are estimated indirectly from a variety of sources.

Although many methods have been attempted, no systematic comparison among them has been reported. A through and fair comparison among them would be of great meaning to the MT sys-

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 941–948,
Suntec, Singapore, 2-7 August 2009.

tem combination research. In this paper, we implement a confusion network-based decoder. Based on this decoder, we compare four commonly used word alignment methods (GIZA++, TER, CLA and IHMM) for hypothesis alignment using the same experimental data and the same multiple MT system outputs with similar features in terms of translation performance. We conduct the comparison study and other experiments in this paper on both spoken and newswire domains: Chinese-to-English spoken and written language translation tasks. Our comparison shows that although the performance differences between the four methods are not significant, IHMM consistently show slightly better performance than other methods. This is mainly due to the fact the IHMM is able to explore more knowledge sources and Viterbi decoding used in IHMM allows more thorough search for the best alignment while other methods has to use less optimal greedy search.

In addition, for better performance, instead of only using one direction word alignment (n-to-1 from hypothesis to backbone) as in previous work, we propose to use more reliable word alignments which are derived from the intersection of two-direction hypothesis alignment to construct confusion network. Experimental results show that the intersection word alignment-based method consistently improves the performance for all four methods on both spoken and written language tasks.

This paper is organized as follows. Section 2 presents a standard framework of confusion network based machine translation system combination. Section 3 introduces four word alignment methods, and the algorithm of computing intersection word alignment for all four word alignment methods. Section 4 describes the experiments setting and results on two translation tasks. Section 5 concludes the paper.

2 Confusion network based system combination

In order to compare different hypothesis alignment methods, we implement a confusion network decoding system as follows:

Backbone selection: in the previous work, Matusov et al. (2006, 2008) let every hypothesis play the role of the backbone (also called "skeleton" or "alignment reference") once. We follow the work of (Sim et al., 2007; Rosti et al., 2007a; Rosti et al., 2007b; He et al., 2008) and choose the hypothesis that best agrees with other hypotheses on average as the backbone by applying Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004). TER score (Snover et al, 2006) is used as the loss function in MBR decoding. Given a hypothesis set *H*, the backbone can be computed using the following equation, where $TER(\bullet,\bullet)$ returns the TER score of two hypotheses.

$$E_b = \arg\min_{\hat{E}\in H} \sum_{E\in H} TER(\hat{E}, E) \qquad (1)$$

Hypothesis alignment: all hypotheses are word-aligned to the corresponding backbone in a many-to-one manner. We apply four word alignment methods: GIZA++-based, TER-based, CLA-based, and IHMM-based word alignment algorithm. For each method, we will give details in the next section.

Confusion network construction: confusion network is built from one-to-one word alignment; therefore, we need to normalize the word alignment before constructing the confusion network.

The first normalization operation is removing duplicated links, since GIZA++ and IHMM-based word alignments could be n-to-1 mappings between the hypothesis and backbone. Similar to the work of (He et al., 2008), we keep the link which has the highest similarity measure $S(e'_j, e_i)$ based on surface matching score, such as the length of maximum common subsequence (*MCS*) of the considered word pair.

$$S(e'_j, e_i) = \frac{2\times len(MCS(e'_j, e_i))}{len(e'_j) + len(e_i)} \qquad (2)$$

where $MCS(e'_j, e_i)$ is the maximum common subsequence of word e'_j and e_i ; $len(.)$ is a function to compute the length of letter sequence. The other hypothesis words are set to align to the *null* word. For example, in Figure 1, e'_1 and e'_3 are aligned to the same backbone word e_2, we remove the link between e_2 and e'_3 if $S(e'_3, e_2) < S(e'_1, e_2)$, as shown in Figure 1 (b).

The second normalization operation is reordering the hypothesis words to match the word order of the backbone. The aligned words are reordered according to their alignment indices. To reorder the null-aligned words, we need to first insert the *null* words into the proper position in the backbone and then reorder the null-aligned hypothesis words to match the *nulls* on the backbone side. Reordering null-aligned words varies based to the word alignment method in the pre-

vious work. We reorder the null-aligned word following the approach of Chen et al. (2008) with some extension. The null-aligned words are reordered with its adjacent word: moving with its left word (as Figure 1 (c)) or right word (as Figure 1 (d)). However, to reduce the possibility of breaking a syntactic phrase, we extend to choose one of the two above operations depending on which one has the higher likelihood with the current null-aligned word. It is implemented by comparing two association scores based on co-occurrence frequencies. They are association score of the null-aligned word and its left word, or the null-aligned word and its right word. We use point-wise mutual information (*MI*) as Equation 3 to estimate the likelihood.

$$MI(e'_i, e'_{i+1}) = \log \frac{p(e'_i e'_{i+1})}{p(e'_i)\,p(e'_{i+1})} \quad (3)$$

where $p(e'_i e'_{i+1})$ is the occurrence probability of bigram $e'_i e'_{i+1}$ observed in the hypothesis list; $p(e'_i)$ and $p(e'_{i+1})$ are probabilities of hypothesis word e'_i and e'_{i+1} respectively.

In example of Figure 1, we choose (c) if $MI(e'_2, e'_3) > MI(e'_3, e'_4)$, otherwise, word is reordered as (d).

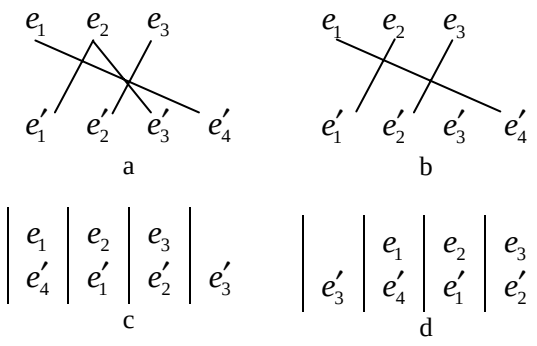

Figure 1: Example of alignment normalization.

Confusion network decoding: the output translations for a given source sentence are extracted from the confusion network through a beam-search algorithm with a log-linear combination of a set of feature functions. The feature functions which are employed in the search process are:

- Language model(s),
- Direct and inverse IBM model-1,
- Position-based word posterior probabilities (arc scores of the confusion network),
- Word penalty,
- N-gram frequencies (Chen et al., 2005),
- N-gram posterior probabilities (Zens and Ney, 2006).

The n-grams used in the last two feature functions are collected from the original hypotheses list from each single system. The weights of feature functions are optimized to maximize the scoring measure (Och, 2003).

3 Word alignment algorithms

We compare four word alignment methods which are widely used in confusion network based system combination or bilingual parallel corpora word alignment.

3.1 Hypothesis-to-backbone word alignment

GIZA++: Matusov et al. (2006, 2008) proposed using GIZA++ (Och and Ney, 2003) to align words between the backbone and hypothesis. This method uses enhanced HMM model bootstrapped from IBM Model-1 to estimate the alignment model. All hypotheses of the whole test set are collected to create sentence pairs for GIZA++ training. GIZA++ produces hypothesis-backbone many-to-1 word alignments.

TER-based: TER-based word alignment method (Sim et al., 2007; Rosti et al., 2007a; Rosti et al., 2007b) is an extension of multiple string matching algorithm based on Levenshtein edit distance (Bangalore et al., 2001). The TER (translation error rate) score (Snover et al., 2006) measures the ratio of minimum number of string edits between a hypothesis and reference where the edits include insertions, deletions, substitutions and phrase shifts. The hypothesis is modified to match the reference, where a greedy search is used to select the set of shifts because an optimal sequence of edits (with shifts) is very expensive to find. The best alignment is the one that gives the minimum number of translation edits. TER-based method produces 1-to-1 word alignments.

CLA-based: Chen et al. (2008) used competitive linking algorithm (CLA) (Melamed, 2000) to build confusion network for hypothesis regeneration. Firstly, an association score is computed for every possible word pair from the backbone and hypothesis to be aligned. Then a greedy algorithm is applied to select the best word alignment. We compute the association score from a linear combination of two clues:

surface similarity computed as Equation (2) and position difference based distortion score by following (He et al., 2008). CLA works under a 1-to-1 assumption, so it produces 1-to-1 word alignments.

IHMM-based: He et al. (2008) propose an indirect hidden Markov model (IHMM) for hypothesis alignment. Different from traditional HMM, this model estimates the parameters indirectly from various sources, such as word semantic similarity, surface similarity and distortion penalty, etc. For fair comparison reason, we also use the surface similarity computed as Equation (2) and position difference based distortion score which are used for CLA-based word alignment. IHMM-based method produces many-to-1 word alignments.

3.2 Intersection word alignment and its expansion

In previous work, Matusov et al. (2006, 2008) used both direction word alignments to compute so-called state occupation probabilities and then compute the final word alignment. The other work usually used only one direction word alignment (many/1-to-1 from hypothesis to backbone). In this paper, we use more reliable word alignments which are derived from the intersection of both direct (hypothesis-to-backbone) and inverse (backbone-to-hypothesis) word alignments with heuristic-based expansion which is widely used in bilingual word alignment. The algorithm includes two steps:

1) Generate bi-directional word alignments. It is straightforward for GIZA++ and IHMM to generate bi-directional word alignments. This is simply achieved by switching the parameters of source and target sentences. Due to the nature of greedy search in TER, the bi-directional TER-based word alignments by switching the parameters of source and target sentences are not necessary exactly the same. For example, in Figure 2, the word "shot" can be aligned to either "shoot" or "the" as the edit cost of word pair (shot, shoot) and (shot, the) are the same when compute the minimum-edit-distance for TER score.

I	shot		killer		I	shoot	the	killer
I	shoot	the	killer		I		shot	killer
		a					b	

Figure 2: Example of two directions TER-based word alignments.

For CLA word alignment, if we use the same association score, direct and inverse CLA word alignments should be exactly the same. Therefore, we use different functions to compute the surface similarities, such as using maximum common subsequence (*MCS*) to compute inverse word alignment, and using longest matched prefix (*LMP*) for computing direct word alignment, as in Equation (4).

$$S(e'_j, e_i) = \frac{2 \times len(LMP(e'_j, e_i))}{len(e'_j) + len(e_i)} \quad (4)$$

2) When two word alignments are ready, we start from the intersection of the two word alignments, and then continuously add new links between backbone and hypothesis if and only if both of the two words of the new link are unaligned and this link exists in the union of two word alignments. If there are more than two links share a same hypothesis or backbone word and also satisfy the constraints, we choose the link that with the highest similarity score. For example, in Figure 2, since *MCS*-based similarity scores $S(shot, shoot) > S(shot, the)$, we choose alignment (a).

4 Experiments and results

4.1 Tasks and single systems

Experiments are carried out in two domains. One is in spoken language domain while the other is on newswire corpus. Both experiments are on Chinese-to-English translation.

Experiments on spoken language domain were carried out on the *Basic Traveling Expression Corpus* (BTEC) (Takezawa et al., 2002) Chinese- to-English data augmented with *HIT-corpus*[1]. BTEC is a multilingual speech corpus which contains sentences spoken by tourists. 40K sentence-pairs are used in our experiment. *HIT-corpus* is a balanced corpus and has 500K sentence-pairs in total. We selected 360K sentence-pairs that are more similar to BTEC data according to its sub-topic. Additionally, the English sentences of *Tanaka* corpus[2] were also used to train our language model. We ran experiments on an IWSLT *challenge task* which uses IWSLT-2006[3] DEV clean text set as development set and IWSLT-2006 TEST clean text as test set.

[1] http://mitlab.hit.edu.cn/
[2] http://www.csse.monash.edu.au/~jwb/tanakacorpus.html
[3] http:// www.slc.atr.jp/IWSLT2006/

Experiments on newswire domain were carried out on the FBIS[4] corpus. We used NIST[5] 2002 MT evaluation test set as our development set, and the NIST 2005 test set as our test set.

Table 1 summarizes the statistics of the training, dev and test data for IWSLT and NIST tasks.

task	data		Ch	En
IWSLT	Train	Sent.	406K	
		Words	4.4M	4.6M
	Dev	Sent.	489	489×7
		Words	5,896	45,449
	Test	Sent.	500	500×7
		Words	6,296	51,227
	Add.	Words	-	1.7M
NIST	Train	Sent.	238K	
		Words	7.0M	8.9M
	Dev 2002	Sent.	878	878×4
		Words	23,248	108,616
	Test 2005	Sent.	1,082	1,082×4
		Words	30,544	141,915
	Add.	Words	-	61.5M

Table 1: Statistics of training, dev and test data for IWSLT and NIST tasks.

In both experiments, we used four systems, as listed in Table 2, they are phrase-based system Moses (Koehn et al., 2007), hierarchical phrase-based system (Chiang, 2007), BTG-based lexicalized reordering phrase-based system (Xiong et al., 2006) and a tree sequence alignment-based tree-to-tree translation system (Zhang et al., 2008). Each system for the same task is trained on the same data set.

4.2 Experiments setting

For each system, we used the top 10 scored hypotheses to build the confusion network. Similar to (Rosti et al., 2007a), each word in the hypothesis is assigned with a rank-based score of $1/(1+r)$, where r is the rank of the hypothesis. And we assign the same weights to each system.

For selecting the backbone, only the top hypothesis from each system is considered as a candidate for the backbone.

Concerning the four alignment methods, we use the default setting for GIZA++; and use toolkit TERCOM (Snover et al., 2006) to compute the TER-based word alignment, and also use the default setting. For fair comparison reason, we decide to do not use any additional resource, such as target language synonym list, IBM model lexicon; therefore, only surface similarity is applied in IHMM-based and CLA-based methods. We compute the distortion model by following (He et al., 2008) for IHMM and CLA-based methods. The weights for each model are optimized on held-out data.

	System	Dev	Test
IWSLT	Sys1	30.75	27.58
	Sys2	30.74	**28.54**
	Sys3	29.99	26.91
	Sys4	**31.32**	27.48
NIST	Sys1	25.64	**23.59**
	Sys2	24.70	23.57
	Sys3	25.89	22.02
	Sys4	**26.11**	21.62

Table 2: Results (BLEU% score) of single systems involved to system combination.

4.3 Experiments results

Our evaluation metric is BLEU (Papineni et al., 2002), which are to perform case-insensitive matching of *n*-grams up to $n = 4$.

Performance comparison of four methods: the results based on direct word alignments are reported in Table 3, row Best is the best single systems' scores; row MBR is the scores of backbone; GIZA++, TER, CLA, IHMM stand for scores of systems for four word alignment methods.

- MBR decoding slightly improves the performance over the best single system for both tasks. This suggests that the simple voting strategy to select backbone is workable.
- For both tasks, all methods improve the performance over the backbone. For IWSLT test set, the improvements are from 2.06 (CLA, 30.88-28.82) to 2.52 BLEU-score (IHMM, 31.34-28.82). For NIST test set, the improvements are from 0.63 (TER, 24.31-23.68) to 1.40 BLEU-score (IHMM, 25.08-23.68). This verifies that the confusion network decoding is effective in combining outputs from multiple MT systems and the four word-alignment methods are also workable for hypothesis-to-backbone alignment.
- For IWSLT task where source sentences are shorter (12-13 words per sentence in average), the four word alignment methods achieve similar performance on both dev and test set. The biggest difference is only 0.46 BLEU score (30.88 for CLA, *vs*. 31.34 for IHMM). For NIST task

[4] LDC2003E14
[5] http://www.nist.gov/speech/tests/mt/

where source sentences are longer (26-28 words per sentence in average), the difference is more significant. Here IHMM method achieves the best performance, followed by GIZA++, CLA and TER. IHMM is significantly better than TER by 0.77 BLEU-score (from 24.31 to 25.08, p<0.05). This is mainly because IHMM exploits more knowledge source and Viterbi decoding allows more thorough search for the best alignment while other methods use less optimal greedy search. Another reason is that TER uses hard matching in computing edit distance.

	method	Dev	Test
IWSLT	Best	31.32	28.54
	MBR	31.40	28.82
	GIZA++	34.16	31.06
	TER	33.92	30.96
	CLA	33.85	30.88
	IHMM	**34.35**	**31.34**
NIST	Best	26.11	23.59
	MBR	26.36	23.68
	GIZA++	27.58	24.88
	TER	27.15	24.31
	CLA	27.44	24.51
	IHMM	**27.76**	**25.08**

Table 3: Results (BLEU% score) of combined systems based on direct word alignments.

Performance improvement by intersection word alignment: Table 4 reports the performance of the system combinations based on intersection word alignments. It shows that:

- Comparing Tables 3 and 4, we can see that the intersection word alignment-based expansion method improves the performance in all the dev and test sets for both tasks by 0.2-0.57 BLEU-score and the improvements are consistent under all conditions. This suggests that the intersection word alignment-based expansion method is more effective than the commonly used direct word-alignment-based hypothesis alignment method in confusion network-based MT system combination. This is because intersection word alignments are more reliable compared with direct word alignments, and so for heuristic-based expansion which is based on the aligned words with higher scores.
- TER-based method achieves the biggest performance improvement by 0.4 BLEU-score in IWSLT and 0.57 in NIST. Our statistics shows that the TER-based word alignment generates more inconsistent links between the two-directional word alignments than other methods. This may give the intersection with heuristic-based expansion method more room to improve performance.
- On the contrast, CLA-based method obtains relatively small improvement of 0.26 BLEU-score in IWSLT and 0.21 in NIST. The reason could be that the similarity functions used in the two directions are more similar. Therefore, there are not so many inconsistent links between the two directions.
- Table 5 shows the number of links modified by intersection operation and the BLEU-score improvement. We can see that the more the modified links, the bigger the improvement.

	method	Dev	Test
IWSLT	MBR	31.40	28.82
	GIZA++	34.38	31.40
	TER	34.17	31.36
	CLA	34.03	31.14
	IHMM	**34.59**	**31.74**
NIST	MBR	26.36	23.68
	GIZA++	27.80	25.11
	TER	27.58	24.88
	CLA	27.64	24.72
	IHMM	**27.96**	**25.37**

Table 4: Results (BLEU% score) of combined systems based on intersection word alignments.

	IWSLT		NIST	
system	Inc.	Imp.	Inc.	Imp.
CLA	1.2K	0.26	9.2K	0.21
GIZA++	3.2K	0.36	25.5K	0.23
IHMM	3.7K	0.40	21.7K	0.29
TER	4.3K	0.40	40.2K	0.57
#total links	284K		1,390K	

Table 5: Number of modified links and absolute BLEU(%) score improvement on test sets.

Effect of fuzzy matching in TER: the previous work on TER-based word alignment uses hard match in counting edits distance. Therefore, it is not able to handle cognate words match, such as in Figure 2, original TER script count the edit cost of (shoot, shot) equals to word pair (shot, the). Following (Leusch et al., 2006), we modified the TER script to allow fuzzy matching: change the substitution cost from 1 for any word pair to

$$COST_{sub}(e'_j, e_i) = 1 - S(e'_j, e_i) \quad (5)$$

which $S(e'_j, e_i)$ is the similarity score based on the length of longest matched prefix (*LMP*) computed as in Equation (4). As a result, the fuzzy matching reports $SubCost(shoot, shot) = 1-(2\times3)/(5+4) = 1/3$ and $SubCost(shoot, the) = 1-(2\times0)/(5+3) = 1$ while in original TER, both of the two scores are equal to 1. Since cost of word pair (shoot, shot) is smaller than that of word pair (shot, the), word "shot" has higher chance to be aligned to "shoot" (Figure 2 (a)) instead of "the" (Figure 2 (b)). This fuzzy matching mechanism is very useful to such kind of monolingual alignment task as in hypothesis-to-backbone word alignment since it can well model word variances and morphological changes.

Table 6 summaries the results of TER-based systems with or without fuzzy matching. We can see that the fuzzy matching improves the performance for all cases. This verifies the effect of fuzzy matching for TER in monolingual word alignment. In addition, the improvement in NIST test set (0.36 BLEU-score for direct alignment and 0.21 BLEU-score for intersection one) are more than that in IWSLT test set (0.15 BLEU-score for direct alignment and 0.11 BLEU-score for intersection one). This is because the sentences of IWSLT test set are much shorter than that of NIST test set.

TER-based systems	IWSLT		NIST	
	Dev	Test	Dev	Test
Direct align	33.92	30.96	27.15	24.31
+fuzzy match	34.14	31.11	27.53	24.67
Intersect align	34.17	31.36	27.58	24.88
+fuzzy match	34.40	31.47	27.79	25.09

Table 6: Results (BLEU% score) of TER-based combined systems with or without fuzzy match.

5 Conclusion

Confusion-network-based system combination shows better performance than other methods in combining multiple MT systems' outputs, and hypothesis alignment is a key step. In this paper, we first compare four word alignment methods for hypothesis alignment under the confusion network framework. We verify that the confusion network framework is very effective in MT system combination and IHMM achieves the best performance. Moreover, we propose an intersection word alignment-based expansion method for hypothesis alignment, which is more reliable as it leverages on both direct and inverse word alignment. Experimental results on Chinese-to-English spoken and newswire domains show that the intersection word alignment-based method yields consistent improvements across all four word alignment methods. Finally, we evaluate the effect of fuzzy matching for TER.

Theoretically, confusion network decoding is still a word-level voting algorithm although it is more complicated than other sentence-level voting algorithms. It changes lexical selection by considering the posterior probabilities of words in hypothesis lists. Therefore, like other voting algorithms, its performance strongly depends on the quality of the *n*-best hypotheses of each single system. In some extreme cases, it may not be able to improve BLEU-score (Mauser et al., 2006; Sim et al., 2007).

References

N. F. Ayan. J. Zheng and W. Wang. 2008. Improving Alignments for Better Confusion Networks for Combining Machine Translation Systems. In *Proceedings of COLING 2008*, pp. 33–40. Manchester, Aug.

S. Bangalore, G. Bordel, and G. Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *Proceeding of IEEE workshop on Automatic Speech Recognition and Understanding*, pp. 351–354. Madonna di Campiglio, Italy.

B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo and M. Federico. 2005. The ITC-irst SMT System for IWSLT-2005. In *Proceeding of IWSLT-2005*, pp.98-104, Pittsburgh, USA, October.

B. Chen, M. Zhang, A. Aw and H. Li. 2008. Regenerating Hypotheses for Statistical Machine Translation. In: *Proceeding of COLING 2008*. pp105-112. Manchester, UK. Aug.

D. Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

C. Fellbaum. editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

X. He, M. Yang, J. Gao, P. Nguyen, R. Moore, 2008. Indirect-HMM-based Hypothesis Alignment for Combining Outputs from Machine Translation Systems. In *Proceeding of EMNLP*. Hawaii, US, Oct.

F. Huang and K. Papinent. 2007. Hierarchical System Combination for Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and*

Computational Natural Language Learning (EMNLP-CoNLL'2007), pp. 277 − 286, Prague, Czech Republic, June.

S. Jayaraman and A. Lavie. 2005. Multi-engine machine translation guided by explicit word matching. In *Proceeding of EAMT*. pp.143–152.

D. Karakos, J. Eisner, S. Khudanpur, and M. Dreyer. 2008. Machine Translation System Combination using ITG-based Alignments. In *Proceeding of ACL-HLT 2008*, pp. 81–84.

O. Kraif, B. Chen. 2004. Combining clues for lexical level aligning using the Null hypothesis approach. In: *Proceedings of COLING 2004*, Geneva, August, pp. 1261-1264.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL-2007*. pp. 177-180, Prague, Czech Republic.

S. Kumar and W. Byrne. 2004. Minimum Bayes Risk Decoding for Statistical Machine Translation. In *Proceedings of HLT-NAACL 2004*, May 2004, Boston, MA, USA.

G. Leusch, N. Ueffing and H. Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. In *Proceedings of EACL*. pp. 241-248. Trento Italy.

E. Matusov, N. Ueffing, and H. Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proceeding of EACL*, pp. 33-40, Trento, Italy, April.

E. Matusov, G. Leusch, R. E. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y. Lee, J. B. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney. System Combination for Machine Translation of Spoken and Written Language. *IEEE Transactions on Audio, Speech and Language Processing*, volume 16, number 7, pp. 1222-1237, September.

A. Mauser, R. Zens, E. Matusov, S. Hasan, and H. Ney. 2006. The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation. In *Proceeding of IWSLT 2006*, pp. 103-110, Kyoto, Japan, November.

I. D. Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2), pp. 221-249.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL-2003*. Sapporo, Japan.

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

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceeding of ACL-2002*, pp. 311-318.

A. I. Rosti, N. F. Ayan, B. Xiang, S. Matsoukas, R. Schwartz and B. Dorr. 2007a. Combining Outputs from Multiple Machine Translation Systems. In *Proceeding of NAACL-HLT-2007*, pp. 228-235. Rochester, NY.

A. I. Rosti, S. Matsoukas and R. Schwartz. 2007b. Improved Word-Level System Combination for Ma-chine Translation. In *Proceeding of ACL-2007*, Prague.

A. I. Rosti, B. Zhang, S. Matsoukas, and R. Schwartz. 2008. Incremental Hypothesis Alignment for Building Confusion Networks with Application to Machine Translation System Combination, In *Proceeding of the Third ACL Workshop on Statistical Machine Translation*, pp. 183-186.

K. C. Sim, W. J. Byrne, M. J.F. Gales, H. Sahbi, and P. C. Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *Proceeding of ICASSP-2007*.

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

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceeding of LREC-2002*, Las Palmas de Gran Canaria, Spain.

D. Xiong, Q. Liu and S. Lin. 2006. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. In *Proceeding of ACL-2006*. pp.521-528.

R. Zens and H. Ney. 2006. N-gram Posterior Probabilities for Statistical Machine Translation. In *Proceeding of HLT-NAACL Workshop on SMT*, pp. 72-77, NY.

M. Zhang, H. Jiang, A. Aw, H. Li, C. L. Tan, and S. Li. 2008. A Tree Sequence Alignment-based Tree-to-Tree Translation Model. In *Proceeding of ACL-2008*. Columbus, US. June.

Y. Zhang, S. Vogel, and A. Waibel 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of LREC 2004*, pp. 2051-2054.

---

* The first author has moved to National Research Council, Canada. His current email address is: Boxing.Chen@nrc.ca.