Source-Side Context-Informed Hypothesis Alignment for Combining Outputs from Machine Translation Systems

Jinhua Du, Yanjun Ma and Andy Way
Centre for Next Generation Localisation,
School of Computing,
Dublin City University
Dublin 9, Ireland
{jdu, yma, away}@computing.dcu.ie

Abstract

This paper presents a new hypothesis alignment method for combining outputs of multiple machine translation (MT) systems. Traditional hypothesis alignment algorithms such as TER, HMM and IHMM do not directly utilise the context information of the source side but rather address the alignment issues via the output data itself. In this paper, a source-side context-informed (SSCI) hypothesis alignment method is proposed to carry out the word alignment and word reordering issues. First of all, the source–target word alignment links are produced as the hidden variables by exporting source phrase spans during the translation decoding process. Secondly, a mapping strategy and normalisation model are employed to acquire the 1-to-1 alignment links and build the confusion network (CN). The source-side context-based method outperforms the state-of-the-art TER-based alignment model in our experiments on the WMT09 English-to-French and NIST Chinese-to-English data sets respectively. Experimental results demonstrate that our proposed approach scores consistently among the best results across different data and language pair conditions.

1 Introduction

In the past several years, multiple system combination has been shown to be helpful in improving translation quality. Recently, confusion network-based networks, either single (Bangalore et al., 2001; Matusov et al., 2006; Rosti et al., 2007a; Sim et al., 2007; He et al., 2008) or multiple networks (Rosti et al., 2007b; Rosti et al., 2008), have become the state-of-the-art methodology to implement the combination strategy. A CN is essentially a directed acyclic graph which is built by aligning a set of translation hypotheses against a reference or "backbone". Each arc between two nodes in the CN denotes a word or token, possibly a *null* item, with an associated posterior probability. Generally, like the translation decoding, the CN decoding process also uses a log-linear model, which combines a set of different features, to search for the best path or an N-best list by dynamic programming algorithms.

Typically, the dominant CN is constructed on the word level by a state-of-the-art framework. Firstly, a minimum Bayes-risk (MBR) decoder (Kumar and Byrne, 2004) is utilised to choose the backbone from a merged set of hypotheses, and then the remaining hypotheses are aligned against the backbone by a specific alignment approach. Currently, most research on system combination has focussed on hypothesis alignment due to its significant role in combination. Since the TER-based (Snover et al., 2006) system combination strategy was introduced to system combination in (Sim et al., 2007) and was shown to outperform the Word Error Rate (WER) alignment metric, many hypothesis alignment metrics have been proposed and successfully applied in system combination, such as ITER (Rosti et al., 2008), ITG (Karakos et al., 2008) and IHMM (He et al., 2008). In all these papers, the proposed alignment method outperformed the TER-based baseline system.

In system combination, source—target-related

knowledge has been shown to significantly improve translation quality (Huang and Papineni, 2007; Rosti et al., 2007a; He et al., 2008). At present, although mainstream statistical machine translation (SMT) systems are implemented based on various paradigms—phrasal, hierarchical and syntax-based—all of them still use the word alignment between the source and target side as the cornerstone. Such systems have demonstrated that word alignment accuracy plays a crucial role when it comes to translation quality. Intuitively, such bilingual word alignment contextual information could be useful in the post-processing stage, especially for the system combination phase.

This paper proposes a source-side context-informed hypothesis alignment for system combination. We employ the source-side word alignment links and source-side phrase span information to heuristically carry out the hypothesis alignment. Firstly, in the translation decoding stage, the spans of translated source-side phrases are kept as the hidden word alignment information. Secondly, we retrieve the phrase table to acquire the word alignment links between the source and target phrases. Finally, by mapping the word alignment links between the backbone and the hypothesis based on the same span of a source phrase, associated with a normalisation model, we can perform the hypothesis alignment and CN efficiently. Our approach does not need any complicated estimation algorithm, nor does it require additional training data or any other resources.

The remainder of this paper is organised as follows. In section 2, we study the influence of the word order of the backbone and different hypothesis alignment metrics on the confusion network. Section 3 describes the working mechanism of our proposed source-side context-informed hypothesis alignment approach. The experiments conducted on different language pairs are reported in Sections 4 and 5. Section 6 concludes and gives avenues for future work.

2 The Impact of Hypothesis Alignment on the Confusion Network

The methodology of hypothesis alignment is similar to the word alignment between source and target language. The distinct differences are firstly that the source and target sides are the same language; and secondly, the word alignment types are limited to 1-to-1 and 1-to-$null$. Currently, the CN has two crucial characteristics: 1) it is a word-level graph; 2) a monotone decoding process is selected. Therefore, hypothesis alignment plays a vital role in the CN because the backbone sentence decides the skeleton and word order of the consensus output.

Figure 1 shows the main steps of how to align the hypotheses and how to carry out the word reordering as well as construct the CN. In Fig. 1(a), hypotheses from different MT systems are merged to form a new N-best list, from which the backbone is selected using the MBR decoder. The most frequently used loss functions in MBR are TER (Snover et al., 2006) and BLEU (Papineni et al., 2002). Then as illustrated in Fig. 1(b), E_1 is assumed as the backbone, with the rest of the hypotheses aligned against it. The symbol @ denotes a $null$ word. Note that there are only three types of word alignment in system combination, namely, 1-to-1, 1-to-$null$ and $null$-to-1 in terms of bidirectional alignment. According to the word alignment, word reordering is carried out and a CN is constructed based on the reordered hypotheses as Fig. 1 (c) shows. Finally, a set of global and local features are integrated into a log-linear model to decode the CN.

The most challenging problem for CN decoding is the phenomenon of "non-grammatical" phrases, which are mainly caused by the arbitrary word reordering and decoding strategy inside the CN. There might be several arcs between any two adjacent nodes. Each arc indicates an alternative word or $null$. The search process is to produce a sequence with the best overall score, while at each position, the selected word is mainly decided by methods such as voting. Thus in some sense, there is no direct grammatical relationship between any adjacent words in the voting decision. Although nowadays most MT research introduces some syntax-like features into the CN (such as a language model, for instance), it still cannot avoid producing "non-grammatical" phrases. However, a high-quality hypothesis alignment can reduce this kind of influence to some extent, since the more accurately the words are aligned, the less noise is produced.

When we examine the impact of hypothesis align-

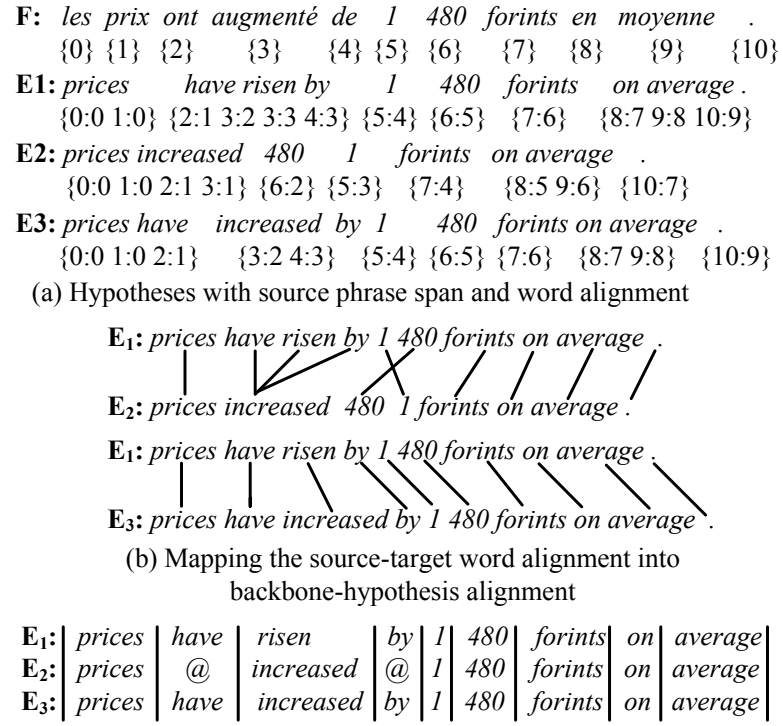

Figure 1: Mapping Source–Target Word Alignment to Hypothesis Alignment

ment on the CN, two key issues should be studied. The first one is that of *word order*: how does the word order impact on the skeleton of the consensus output? The second one is the *hypothesis alignment accuracy*: how does the hypothesis alignment influence the word sequence of the consensus output?

To study the first issue, considering that the word order of CN is decided by the backbone, we performed a set of experiments to compare the influence on consensus output of selecting different backbones for our CN. Table 1 shows the comparison results. We use the WMT09 English-to-French system combination shared task as the evaluation data set, including 2525 sentences and 16 1-best systems. TER is used as the default hypothesis alignment metric. The results are reported in TER, case-sensitive BLEU, NIST (Doddington, 2002) and Meteor (MTR) (Banerjee and Lavie, 2005).

The Worst-CN, Best-CN and MBR-CN are the outputs of the CNs using the worst single, best single and MBR result as the backbone respectively. We can see that 1) MBR is better than the best single system; 2) the MBR-CN obtains the best performance in terms of the four automatic evaluation metrics. The better the word order of the backbone is, the better the performance.

Backbone	TER	BLEU	NIST	MTR
Oracle	52.58	33.84	8.04	23.95
Worst Single	69.19	14.73	5.57	12.40
Best Single	59.21	25.43	6.99	18.97
MBR	58.05	26.54	7.12	19.81
Worst-CN	59.16	23.53	7.04	17.63
Best-CN	57.03	26.73	7.29	19.84
MBR-CN	**56.84**	**27.56**	**7.33**	**20.33**

Table 1: The influence of backbone on CN

By using the same backbone but different hypothesis alignment methods, we compare the results to address the second issue. Correctly aligning synonyms to each other is a challenging issue. For instance, in Fig. 1 (b), "risen" in $E1$ and "increased" in $E2$ and $E3$ express the same meaning with different morphologies. Of course, a simple 'exact match' algorithm is incapable of dealing with this issue. In our experiments, three dominant types of

hypothesis alignment metrics are used, namely TER, HMM (Matusov et al., 2006) and IHMM. The data set we used is still the WMT09 English-to-French system combination shared task. TER aligns the words based on the exact match principle; HMM uses the same principle as the word alignment model of (Vogel et al., 1996), while IHMM uses two similarity models and one distortion model to perform the alignment. Table 2 shows the results for these three metrics.

Alignment	TER	BLEU	NIST	MTR
TER	56.84	27.56	7.33	20.33
IHMM	56.83	27.27	7.24	20.27
HMM	**56.56**	**27.64**	**7.38**	**20.52**

Table 2: The influence of alignment metrics on the CN

In this experiment, the three CNs are built on the MBR-based backbone, and decoded using the same features and weights. We can see that in this task, the HMM approach outperforms the other two methods. When we manually examine the alignment result, the HMM method has a higher word alignment accuracy and produces a lower non-grammatical error rate.

3 Source-side Context-Informed Hypothesis Alignment

3.1 Motivation

The major difficulties in system combination are to capture the internal structures of the various results and normalize them. However, for an MT system, it is feasible to provide more powerful source-side information for system combination, as opposed to just a 1-best or an N-best list. In the past, bilingual information has been demonstrated to be helpful for improving translation quality. (Rosti et al., 2007a) incorporated target-to-source information into the phrase-level combination, and (Huang and Papineni, 2007) proposed to incorporate bilingual information from source and target sentences in word-level system combination. However, they did not directly use the source-target word alignment links to provide the information for the hypothesis alignment. Considering this point, we propose a source–side context-informed hypothesis alignment which can carry out the hypothesis alignment by employing the word alignment links derived from the GIZA++ (Och and Ney, 2003) training.

3.2 Description of Algorithm

As the source–target word alignment task, the aim of hypothesis alignment is to obtain the best word alignment links between the hypothesis and the backbone. Intuitively, this task has been performed in the process of training GIZA++ (Och and Ney, 2003), extracting the phrases and decoding. However, this kind of alignment information is subsequently abandoned during the translation decoding phase. Our goal is to keep the source-side word alignment information and utilise it in the hypothesis alignment phase.

Whether the system is syntax-based, hierarchical phrase-based or just plain phrase-based, each requires an initial phrase table containing word alignments. The best search path is definitely to expand a sequence of source spans without any overlap. Then when generating the target phrases, we also set the system to export the source span information. By integrating these source–target word alignments, we can align the hypotheses via a mapping algorithm and normalisation model.

3.2.1 Mapping Source–Target Word Alignment to Hypothesis Alignment

Let us use Figure 1 as an example to describe the mapping algorithm. In Fig. 1(a), all the hypotheses are collected with the span of each source phrase. Employing this span information, we can retrieve the specific word alignment links from the initial phrase table as shown in (a). In this step, the alignment links consist of 1-to-1, 1-to-$null$, 1-to-N and N-to-1. types. However, only 1-to-1 and 1-to-$null$ word alignment are needed for the CN, so the normalisation model is designed to process the 1-to-N links. This process is described further in section 3.2.2.

Now we study how to map source–target word alignments to the backbone-to-hypothesis alignments. Assuming E_1 is the selected backbone E_b and E' is the hypothesis, we use $\boldsymbol{F} = \{f_1, \ldots, f_k\}$ as a source-side word (or minimum span), $\Lambda_b = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k\}$ as the set of word alignments between $\boldsymbol{F}$ and the counterpart of E_b, and $\Lambda' = \{\boldsymbol{A}'_1, \ldots, \boldsymbol{A}'_k\}$ as the set of word alignments between $\boldsymbol{F}$ and the cor-

responding part of E'. $\boldsymbol{A}_i$ and $\boldsymbol{A}'_i$ are represented as a set of alignment pairs $\langle f_i, \{e_l, \ldots, e_m\} \| (m \geq l \geq 0) \rangle$ and $\langle f_i, \{e'_p, \ldots, e'_q\} \| (q \geq p \geq 0) \rangle$ respectively, which indicates that each source-side word f_i could be aligned to multiple target words or a $null$ word. Mapping Λ and Λ' to the word alignment between E_b and E' can be achieved as in (1) and (2):

$$\Lambda_b \cap \Lambda' = \{\boldsymbol{A}_1 \cap \boldsymbol{A}'_1, \ldots, \boldsymbol{A}_k \cap \boldsymbol{A}'_k\} \quad (1)$$

$$\boldsymbol{A}_i \cap \boldsymbol{A}'_i = \langle \tilde{E}_i, \tilde{E}'_i \rangle \quad (2)$$

where $\tilde{E}_i$ is a set of words in E_b, and $\tilde{E}'_i$ is the set of words in E', both of which are aligned to f_i. Fig. 1 (a) and (b) demonstrate the hypotheses alignment results of our mapping strategy.

3.2.2 Normalisation Model for Hypothesis Alignment

The general normalisation algorithm for processing the 1-to-N or N-to-1 word alignments is to keep the link which gives the highest translation probability (Matusov et al., 2006; He et al., 2008). One problem for this algorithm is that the sparseness of limited training data for hypothesis alignment would compromise the accuracy of the word alignments and lexical probabilities.

Considering that the backbone and the hypothesis are the same language, it is easy to integrate some morphological and syntactic features into a normalisation model to reduce the alignment error rate, so we applied a modified similarity model. This model resembles the similarity model in IHMM alignment proposed by (He et al., 2008). The essential difference is that we use this similarity model at the word level to select a 1-to-1 link in a set of 1-to-N or N-to-1 alignment links, whereas they apply it as the emission model to search for a best alignment sequence at the sentence level.

We take the backbone-to-hypothesis direction as an example to illustrate our algorithm. The hypothesis-to-backbone direction can be done likewise. Given a backbone e_1^I consisting of I words $e_1, \ldots, e_I$ and a hypothesis $e_1'^J$ consisting of J words $e'_1, \ldots, e'_J$, $A_{E \to E'}$ denotes the backbone-to-hypothesis word alignment $a_1^I = (a_i, \ldots, a_I)$ between e_1^I and $e_1'^J$. Since the similarity model primarily normalises the 1-to-N alignments, $A_{E \to E'}$ can be represented as a set of pairs $a'_j = \langle E_j, e'_j \rangle$ denoting a link between one single hypothesis word e'_j and several backbone words $E_j = \{a_i = j \| i = m, \ldots, n; I \geq n \geq m \geq 1\}$. If the word e'_j is aligned to a $null$ word, the set E_j is empty. Given this notation, the modified equation can be written as in (3):

$$p(e'_j|e_i) = \alpha \cdot p_{lex}(e'_j|e_i) + (1 - \alpha) \cdot p_{sim}(e'_j|e_i)$$

$$\hat{a} = \max_{i=m,\ldots,n} \{p(e'_j|e_i)\} \quad (3)$$

where p_{lex} and p_{sim} denote the lexical alignment probability and the similarity between the backbone word e_i and hypothesis word e'_j respectively. α is the interpolation factor. $\hat{a}$ is the best 1-to-1 link in the set of 1-to-N alignments. Similarly, the 1-to-N links in the hypothesis-to-backbone alignment direction can be computed by the above algorithm.

Since the word alignment links between the backbone and the hypothesis are derived from source-side-informed word alignments, p_{lex} can be calculated by summing the bidirectional lexical translation probabilities between source and target aligned words (He et al., 2008). In order to compute the similarity of two linked words, we can utilise the longest matched prefix (LMP), longest common subsequence (LCS) (He et al., 2008) or cosine similarity algorithm to compute p_{sim}.

After the bidirectional normalisation, we employ the intersection rule to acquire the 1-to-1, 1-to-$null$ and $null$-to-1 links.

Out-of-vocabulary words (OOVs) are one important issue for our proposed hypothesis alignment. If the systems adopt different training data, then the outputs from these systems could include different OOVs. If we directly remove the OOVs from the hypothesis, when performing the mapping step, it would cause problems. To solve this issue, we should keep the completeness of the source-side span, and regard the source OOVs as the corresponding target phrases, and then mark them as the 1-to-1 word alignments.

3.2.3 Word reordering

The word alignments between the backbone and the hypothesis are used to carry out word reordering on the hypothesis side. Firstly, we move a word

or phrase with a 1-to-1 alignment link so that it is aligned to the corresponding position in the backbone. We then perform the edit distance operation on both the backbone and hypothesis sides. For instance, if a word in the backbone has no alignment link with the word in the hypothesis, we insert a particular $null$ mark in this position on the hypothesis side which denotes "deletion"; similarly, the "insertion" operation can be performed. Fig. 1 (c) shows the results after performing word reordering and the edit operation.

4 Experimental Settings

In this section, we introduce the experimental settings for evaluating our source-side context-informed hypothesis alignment method.

4.1 Training Data

To verify the effectiveness of our method, we performed experiments on Chinese-to-English (C2E) and English-to-French (E2F) data sets.

Diversity has a significant influence on the performance of system combination (Macherey and Och, 2007). Although we have different types of MT systems to generate a set of translations, the training data for these systems are basically the same. This would cause a high correlation between the hypotheses and would potentially decrease the system combination performance. In order to increase the diversity, we sample the training data to train a number of translation models. Furthermore, we can adjust parameters such as the distortion limit or use different development sets to reduce any such correlation.

Chinese-to-English Task
5 sub-training data sets are randomly sampled from a large-scale database, each of which includes 400K sentence pairs, including the HK parallel corpus, ISI parallel data, UN data and other news data.

English-to-French Task
We also build 5 sub-training data sets, each of which includes 500k sentence pairs and is sampled randomly from the total parallel corpus, which comprises Europarl data and Giga News data.

4.2 Development and Test Data

Chinese-to-English
The devset used for translation system parameter training is the NIST MT05 test set which contains 1082 sentences; the devset used for system combination parameter tuning (including MBR decoding tuning, CN tuning) is the NIST MT06 test set which contains 1664 sentences. The test set is the NIST MT08 "current" test set which has 1357 sentences from two different domains, namely newswire and web-data genres. All the dev and test sets have 4 references per source sentence.

English-to-French
The devset used for translation system parameter training is the WMT2009-dev-a set which contains 1025 sentences; the devset used for system combination parameter tuning is the WMT2009-dev-b set which consists of 1026 sentences. The test set is the WMT2009 shared task test set which includes 3027 sentences. All the sets are from the news domain and only have one reference per source sentence.

All the results are reported using BLEU, TER, NIST and Meteor scores. The parameters and weights are optimized on the BLEU score.

4.3 Baseline Combination System

Experiments are conducted to compare the proposed source-side context-informed hypothesis alignment approach with the dominant TER-based method by a standard combination framework: the backbone selection via MBR decoder, the CN decoding and a log-linear re-score module with global features (Du et al., 2009).

5 Experimental Results

To save computation effort, we use only the 1-best hypothesis from the individual systems as input to the combination phase. Five individual systems are built according to the sampled sub-training data.

5.1 Chinese-to-English Translation

Table 3 shows the performance of the best and the worst of the single systems as well as the Oracle result in terms of the BLEU score. The results for the consensus translation demonstrate a significant improvement compared to the best single system. For the SSCI-based approach, TER is reduced from 66.17% to 64.80%, and the BLEU, NIST, METEOR are improved by 6.75%, 6.20% and 0.75% relative

points respectively. We also compared the SSCI-based combination result with the TER-based combination result. It can be observed that the SSCI-based consensus translation is superior to the TER-based combination translation in terms of all evaluation measures.

System	TER	BLEU	NIST	MTR
Worst Single	68.86	17.33	6.59	39.82
Best Single	66.17	21.64	6.94	42.95
Oracle	62.88	26.67	7.93	44.95
TER-based	65.70	22.47	7.36	43.11
SSCI-based	**64.80**	**23.10**	**7.37**	**43.27**

Table 3: Results on Chinese-to-English Test Set

5.2 English-to-French Translation

In this task, the results for the SSCI-based consensus translation show a significant improvement in translation quality compared to the best single system as illustrated in Table 4. The TER is reduced from 64.97% to 63.12%, and the BLEU, NIST, METEOR are improved by 6.19%, 3.67% and 5.26% relative points respectively. In this language pair, the SSCI-based framework consistently outperforms the TER-based combination system in terms of all evaluation measures. The results in Table 3 and Table 4 are verified by significance test (Zhang and Vogle, 2004). We found the SSCI-based outputs are significantly better than those from TER-based combination.

System	TER	BLEU	NIST	MTR
Worst Single	71.72	15.18	5.33	11.88
Best Single	64.97	20.04	6.27	15.22
Oracle	60.43	24.84	6.91	17.77
TER-based	63.32	21.09	6.47	15.95
SSCI-based	**63.12**	**21.28**	**6.50**	**16.02**

Table 4: Results on English-to-French Test Set

5.3 Intrinsic Comparison with Other Post-Aligning Methods

We define our proposed method as the pre-aligning approach which actually performs the alignment before MT decoding, while the traditional alignment methods such as TER, HMM and IHMM are defined as the post-aligning methods, which carry out the alignment after MT decoding. The essential difference is that we utilise the source–target word alignment information to acquire the hypothesis alignment. This approach has two advantages: 1) saving computation effort and reducing the complexity—we just need to map the source–target alignment links to the backbone-to-hypothesis alignment and select the best 1-to-1 links using a normalisation model; and 2) we do not need to perform global word reordering in order to reduce the risk of non-grammatical fragments. The post-aligning methods need to search for the best alignment and then perform the word reordering at sentence level. In most cases, 'word reordering' involves moving words rather than phrases. Therefore, some function words like prepositions would be isolated and are forced to align to $null$. This kind of global word reordering will break the original meaning of the hypothesis and increase non-grammatical fragments. However, our approach is to align phrases according to the source-side span and primarily perform the word reordering locally that can decrease the risk of "non-grammatical" fragments.

6 Conclusions and Future Work

In this paper, we presented a source-side context-informed hypothesis alignment for system combination. The motivation was that the source-side contextual knowledge has been shown to be helpful in SMT. In the proposed approach, we employed the source-side word alignment links and source-side phrase span information to heuristically carry out the hypothesis alignment. Firstly, the span of the translated source-side phrase was kept during the MT decoding stage. Secondly, the source–target word alignment links were retrieved from the phrase table. Finally, the SSCI-based CN was constructed by our mapping algorithm followed by a normalisation process. On two different data sets, and for two different language pairs, we demonstrated that our model improves translation quality according to four different evaluation metrics. Furthermore, our approach does not need to estimate any complicated alignment model, and it is easy to integrate rich features to the normalisation model.

As for future work, firstly we plan to automatically evaluate the alignment performance of differ-

ent approaches. Secondly, we plan to examine how source–target word alignment quality influences the accuracy of the hypothesis alignment. We also intend to integrate richer features such as part-of-speech tags into our normalisation model to improve the 1-to-1 links.

Acknowledgments

This work is supported by Science Foundation Ireland (Grant No. 07/CE/I1142). Thanks also to the reviewers for their insightful comments and suggestions.

References

Srinivas Bangalore, German Bordel and Giuseppe Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *Proceedings of 2001 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 351–354.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, ACL-2005*, pages 65–72.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of the Second international conference on Human Language Technology Research*, pages 138–145.

Jinhua Du, Yifan He, Sergio Penkale and Andy Way. 2009. MaTrEx: The DCU MT System for WMT2009. In *Proceedings of the Third Workshop on Statistical Machine Translation, EACL 2009*, pages 95–99.

Fei Huang and Kishore Papineni. 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*, pages 277–286.

Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen and Robert Moore. 2008. Indirect HMM-based hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 98–107.

Damianos Karakos, Jason Eisner, Sanjeev Khudanpur and Markus Dreyer. 2008. Machine translation system combination using ITG-based alignments. In *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 81–84.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-Risk Decoding for Statistical Machine Translation. In *Proceedings of HLT-NAACL*, pages 169–176.

Wolfgang Macherey and Franz J. Och. 2007. An empirical study on computing consensus translations from multiple machine translation systems. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 986–995.

Evgeny Matusov, Nicola Ueffing and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proceedings of EACL'06*, pages 33–40.

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

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL-02*, pages 311–318.

Antti-Veikko I. Rosti, Bing Xiang, Spyros Matsoukas, Richard Schwartz, Necip F. Ayan and Bonnie J. Dorr. 2007a. Combining outputs from multiple machine translation systems. In *Proceedings of HLT-NAACL*, pages 228–235.

Antti-Veikko I. Rosti, Bing Zhang, Spyros Matsoukas and Richard Schwartz. 2008. Incremental Hypothesis Alignment for Building Confusion Networks with Application to Machine Translation System Combination. In *Proceedings of ACL/WMT 2008*, pages 183–186.

Antti-Veikko I. Rosti, Spyros Matsoukas and Richard Schwartz. 2007b. Improved Word-Level System Combination for Machine Translation. In *Proceedings of ACL-07*, pages 312–319.

Khe Chai Sim, William J. Byrne, Mark J.F. Gales, Hichem Sahbi, and Phil C. Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 105–108.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231.

Stephan Vogel, Hermann Ney and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 836–841.

Ying Zhang and Stephan Vogel. 2004. Measuring Confidence Intervals for the Machine Translation Evaluation Metrics. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 85–94.