Using TERp to Augment the System Combination for SMT

Jinhua Du and **Andy Way**
CNGL, School of Computing
Dublin City University, Dublin, Ireland
{jdu, away}@computing.dcu.ie

Abstract

TER-Plus (TERp) is an extended TER evaluation metric incorporating morphology, synonymy and paraphrases. There are three new edit operations in TERp: Stem Matches, Synonym Matches and Phrase Substitutions (Paraphrases). In this paper, we propose a TERp-based augmented system combination in terms of the backbone selection and consensus decoding network. Combining the new properties of the TERp, we also propose a two-pass decoding strategy for the lattice-based phrase-level confusion network (CN) to generate the final result.The experiments conducted on the NIST2008 Chinese-to-English test set show that our TERp-based augmented system combination framework achieves significant improvements in terms of BLEU and TERp scores compared to the state-of-the-art word-level system combination framework and a TER-based combination strategy.

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 have become the state-of-the-art methodology to implement the combination strategy (Bangalore et al., 2001; Matusov et al., 2006; Sim et al., 2007; Rosti et al., 2007a; Rosti et al., 2007b; He et al., 2008). A CN is built by aligning a set of translation hypotheses against a reference or "backbone" which is usually generated by a minimum Bayes-risk decoder (MBR) (Kumar and Byrne, 2004). Generally, as with 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 in system combination for SMT is constructed on the word level constrained by the inherent property of the CN. Basically, there are two critical parts to build a word-level CN, namely hypothesis alignment and the structure of the CN.

Hypothesis alignment involves aligning a set of hypotheses against the "backbone" under a specific alignment metric, such as TER (Snover et al., 2006), HMM (Matusov et al., 2006), IHMM (He et al., 2008), TERp (Snover et al., 2009) etc. Synonym matching is the most challenging issue for the hypothesis alignment metric because it has an important impact on alignment accuracy and the final consensus decoding. As a consequence, many hypothesis alignment metrics integrate rich linguistic features to increase the capability of synonym matching. IHMM uses a similarity function to perform synonym matching and it significantly outperforms the TER method. Ayan et al. (2008) modified TER to consider substitutions of synonyms using WordNet (Fellbaum, 1980). Snover et al. (2009) extended TER to TERp in a similar idea

that incorporates the stems and synonym matching (Banerjee and Lavie, 2005) and paraphrase matching (Kauchak and Barzilay, 2006; Zhou et al., 2006) to increase the alignment accuracy.

Regarding the TERp metric, Rosti et al. (2009) firstly used it to increase the hypothesis alignment in the WMT2009 system combination shared task and achieved the best performance in their experiments. Barrault (2010) developed an open source MT system combination using TERp which is still based on a word-level CN. In this paper, we make good use of the synonyms and paraphrases recognised by the TERp metric to upgrade our word-level combination framework to the phrase level. Additionally, we develop a weighted MBR using TERp as the loss function to train system weights for our proposed framework.

As to the structure of the CN, the state-of-the-art form is a word-level network. A CN is essentially a directed acyclic graph which includes weighted arcs and nodes. Each arc between two nodes in the CN denotes a word or token, possibly a *null* item, with an associated posterior probability. Feng et al. (2009) proposed a lattice-based network which allows several words to connect with other several words, i.e., many-to-many mappings. Phrase pair alignment can reduce the risk of producing ungrammatical phrases because of the coherence between the words in a phrase.

In this paper, we propose a TERp-based augmented system combination network in which firstly, TERp is used as a loss function in a weighted MBR (wMBR) to select a backbone; secondly, TERp is employed as the hypothesis alignment to carry out the word alignment between the backbone and the set of hypotheses; and then to build a lattice-based phrase-level network by extending the TERp-based alignment points. During the network decoding process, we present a two-pass decoding strategy to leverage the selection preference to obtain better results.

The remainder of this paper is organised as follows. In section 2, we introduce the mechanisms of TER and TERp metrics as well as the word-level CN and phrase-level CN. In section 3, we describe a weighted MBR using TERp as the loss function to select the backbone. Section 4 proposes our TERp-based augmented combination framework which is built on the phrase level. Furthermore, we present a two-pass decoding strategy to generate the final consensus. The experiments conducted on the NIST Chinese-to-English test set are reported in Sections 5 and 6. Section 7 concludes and gives avenues for future work.

2 Background

2.1 TER-Plus

TERp is closely related to TER, so in order to fully understand TERp, we first introduce TER.

The TER (translation edit rate) metric measures the ratio of the number of edit operations between the hypothesis E' and the reference E_b to the total number of words in the E_b. Here the backbone E_b is assumed as the reference. The allowable edits include insertions (Ins), deletions (Del), substitutions (Sub) and phrase shifts (Shft). The TER of E' compared to E_b is computed as in (1):

$$\text{TER}(E', E_b) = \frac{Ins + Del + Sub + Shft}{N_b} \times 100\% \quad (1)$$

where N_b is the total number of words in E_b. The difference between TER and classical Edit Distance (or WER) is the sequence shift operation, which allows phrasal shifts in the output to be captured.

The $Shft$ edit is carried out by a greedy algorithm and restricted by three constraints: 1) the shifted words must exactly match the reference words in the destination position; 2) the word sequence of the hypothesis in the original position and the corresponding reference words must not match exactly; 3) the word sequence of the reference that corresponds to the destination position must be misaligned before the shift (Snover et al., 2006).

TER-Plus (TERp) is an extension of TER that aligns words in the hypothesis and reference not only when they are exact matches but also when the words share a stem or are synonyms. In addition, it uses probabilistic phrasal substitutions

to align phrases in the hypothesis and reference. These phrases are generated by considering possible paraphrases of the reference words. Different from the constant edit cost for all operations such as shifts, insertion, deleting or substituting in TER, all edit costs in TERp are optimized to maximize correlation with human judgments.

TERp uses all the edit operations of TER as well as three new edit operations: Stem Matches, Synonym Matches and Phrase Substitutions. TERp employs the Porter stemming algorithm (Porter, 1980) and WordNet (Fellbaum, 1980) to perform the "stem match" and "synonym match" respectively. Sequences of words in the reference are considered to be paraphrases of a sequence of words in the hypothesis if that phrase pair occurs in the TERp phrase table (Snover et al., 2009).

2.2 Lattice-based CN

A lattice-based CN is an extension of the word-level CN. The word-level CN limits the word alignment between the backbone and the hypothesis to 1-to-1, 1-to-$null$ and $null$-to-1, while the lattice-based network allows many-to-many mappings. Therefore, each arc in the word-level CN indicates "one" word or $null$. However, in the lattice-based network, the arc might represent a phrase which includes several words. As a result, we define the lattice-based network as the phrase-level CN. Figure 1 is a simple example which respectively uses TER and TERp to carry out the alignment and the CN construction to illustrate the differences between the word-level and the phrase-level CN.

Figure 1 (a) shows the backbone and a translation hypothesis. Figure 1 (b) and Figure 1 (c) are the TER-based word alignment and the TER-based word-level CN respectively. "@" represents the $null$ arc. Figure 1 (d) and Figure 1 (e) indicate the TERp-based word alignment and the TERp-based phrase-level CN.

It can be found in Figure 1 that the TERp aligns "early next week" against "at the beginning of next week" as paraphrases, as well as the paraphrases "take place " and "start". In this case, if the phrase pairs are kept as a whole as shown in Figure 1 (e) rather than broken them into individual words as shown in Figure 1 (c), then the ungrammatical risk during the decoding process would be decreased. This example also shows the obvious advantage of TERp: using WordNet and paraphrases to recognise and align the synonymous words and phrases to increase the alignment accuracy.

3 TERp-based Weighted Minimum Bayes-Risk Decoding

In state-of-the-art MT system combination, MBR decoding plays an important role to select the backbone for the CN. The backbone decides the word orders of the CN and the consensus output. In our framework, we employ TERp as the Loss Function in MBR to select the backbone as in (2):

$$E_b = \arg\min_{E \in E_i} \sum_{j=1}^{N_s} \texttt{TERp}(E_j, E_i) \qquad (2)$$

where N_s is the number of systems.

Equation (2) indicates an MBR decoder with uniform posterior probabilities. In fact, the uniform posterior distribution only performs robustly when the individual systems have a similar quality and are less correlated (Macherey and Och, 2007). Generally, there are two ways to leverage the robustness of the MBR decoder. One way is the empirical way that filters out the worse or closely relevant individual systems based on some specific metric scores and keeps the better systems with similar quality (Macherey and Och, 2007); the other way is the discriminative way that trains system weights through the discriminative training algorithm. Sim et al. (2007) and Rosti et al. (2007) employed a TER-based weighted MBR to achieve better results than the uniform distributed MBR. In our TERp-based method, we also use the second way – system weights estimation – to optimise system weights on the development set (devset) and then apply them to the test set.The weighted MBR is written as in (3):

$$E_b = \arg\min_{E \in E_i} \sum_{j=1}^{N_s} \omega_j \cdot \texttt{TERp}(E_j, E_i) \qquad (3)$$

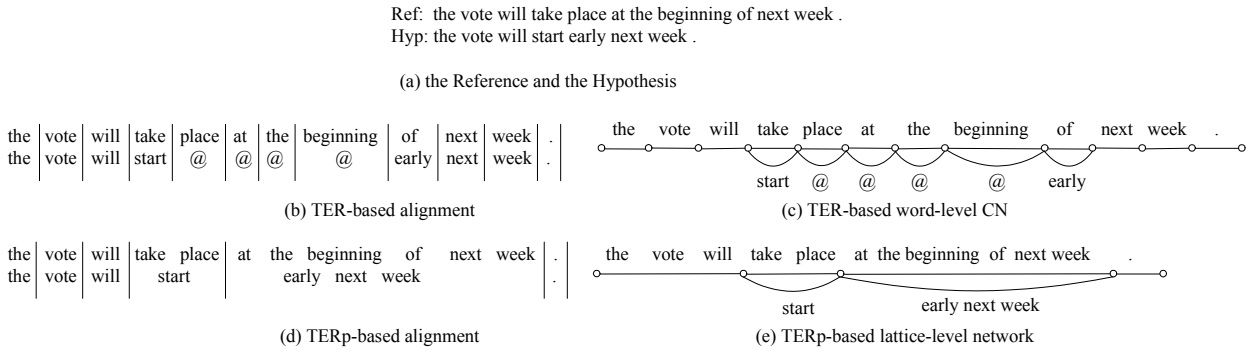

Figure 1: Comparison between a word-level CN and a lattice-based CN

Different from the use of wMBR in (Sim et al., 2007) and (Rosti et al., 2007a), who only use the weights at the stage of the MBR decoding, we use the weights trained by the MBR as the system confidence for each individual system in the lattice-based CN construction and decoding.

4 TERp-based Phrase-level Combination Framework

In this section, we propose a TERp-based phrase-level combination framework and a two-pass decoding strategy. Specifically, we extend the word-level "Stem", "Synonyms" and "Paraphrases" to a phrasal alignment under some restricted conditions. Furthermore, we design a two-pass decoding strategy to leverage the arc confidence to search for the best path.

4.1 Motivation

Typically, in MT system combination, the CN is restricted only to 1-to-1, $null$-to-1 and 1-to-$null$ alignments. The advantage of this method is that if the word alignment performs very well (i.e., the synonyms are accurately aligned), then the candidates on each arc in the CN are not only diverse but also context-correlated, which can reduce the number of ungrammatical errors. However, the potential disadvantage of this method is that if the hypothesis alignment performs badly, the word-level CN maybe increase the risk of producing ungrammatical phrases or fragments. Therefore, the combination quality is heavily reliant on the performance of the hypothesis alignment metric. Feng et al. (2009) proposed a lattice-based CN which not only allows 1-to-1 mappings but also many-to-many mappings, which allows a phrase or several words to be an arc. The advantage of this method is that it can keep the coherence and consistency of a phrase and its context. They used the IHMM method to carry out the bidirectional word alignment and extract the phrasal alignment. IHMM has a limited capability of synonym matching by using a similarity function.

Motivated by the fact that TERp integrates WordNet, Porter Stemming and a big Paraphrase Table to increase the capability of synonym matching, we propose using TERp to build a lattice-based phrase-level combination framework. Consequently, we design a two-pass decoding strategy to search the best path.

4.2 Lattice-based Phrase-level Confusion Network

In TERp, the "stem match", "synonym match" and the paraphrases are respectively notated as "T", "Y" and "P". Additionally, the "exact match", "substitution", "insertion" and "deletion" are defined as "E", "S", "I" and "D" respectively. Since the alignment of synonyms has a significant influence on the consensus quality, we are considering in terms of the consensus decoding that

- the substitution match is more like noise to break the consecutive phrase into words;
- the best path should prefer the candidates between the synonymous words;

- the synonyms, stems and paraphrases should have a higher confidence.

Based on the considerations above, we come up with an idea that extends word-level synonyms or paraphrases to phrasal synonymous alignments if and only if the following conditions are satisfied:

- the word alignment link is marked as "T", "Y" or "P";
- the words in front of or behind the synonym word are marked as "E", "T", "Y" or "P";
- these words can be connected into a consecutive sequence of words;
- the maximum length of the consecutive sequence of words are limited to 7 words.

Accordingly, we define an extra operation as "arc combination" to adjust the arcs and the arc confidence, which will be discussed later. The detailed realisation is shown in Figure 2 and Figure 3.[1]

E_b : e_1 e_2 e_3 e_4 [e_5 e_6 e_7]
E' : e'_1 e'_2 e'_3 * [e'_4 e'_5]
A : E T S D P
(a) alignment between E_b and E' produced by TERp

E_b : e_1 e_2 e_3 e_4 e_5 e_6 e_7
E'' : e''_3 e''_2 e''_1 e''_5 e''_6 e''_7
A : E S E D E Y S
(b) alignment between E_b and E'' produced by TERp

Figure 2: Examples of TERp alignment

In Figure 2, E_b is the backbone selected by the TERp-based wMBR decoder. E' and E'' are two hypotheses aligned against E_b. We can see that in Figure 2 (a), e_1, e'_1 are the "exact match", e_2 and e'_2 are the "stem match", e_3 and e'_3 are the "substitution match", e_4 is aligned to $null$ which is a "deletion match", $e_5e_6e_7$ and $e'_4e'_5$ are paraphrases which comprise the phrasal alignment. In Figure 2 (b), e_6 and e''_6 are the "synonym match" as well.

[1]In order to easily explain our method and contain as many synonym matching phenomena, here we use sequences of pseudo words e_i.

Figure 3 shows an example of the detailed construction process of a lattice-based phrase-level CN using TERp-based alignment links. As mentioned before, one important operation we defined in the network construction is the "arc combination" which combines the synonym and stem matching arcs into one arc. In Figure 3 (a), the base lattice is firstly constructed by the alignment links between E_b and E'. Secondly, since the words e_2 and e'_2 are marked as the "stem match", then we combine them as one arc using the "arc combination" operation. The deleted arc during the combination operation is shown by the dashed lines. Meanwhile, the data information and the position of the deleted arc are stored in the combination node which will be used when tracing back during our two-pass decoding. Similarly, the arcs of $e_5e_6e_7$ and $e'_4e'_5$ are combined into one arc as well, as shown in Figure 3 (b). After the "arc combination" operation, we then add the second alignment pair of E_b and E'' into the base lattice network which is shown in Figure 3 (c). Finally, we carry out the "arc combination" operation again to extend the stem/synonym match to the phrasal alignment and combine the "same" arcs and store the deleted arcs. Figure 3 (d) shows the final lattice-based phrase-level CN.

4.3 Two-pass Decoding Algorithm

In CN decoding, a log-linear model with several features, such as the word posterior, the langauge model etc., is employed to search for a best path by traversing all the nodes from left-to-right. Similarly, in the lattice-based CN, the searching algorithm also needs to travel all the nodes to calculate the best path. The log-linear model and features we used is similar to that in (Feng et al., 2009).

Since we combined the "synonym" phrasal arcs and paraphrase arcs during the lattice construction and kept the "deleted" arcs in the corresponding nodes, we need to restore and evaluate these arcs during the decoding process. Therefore, a two-pass decoding strategy is proposed for our TERp-augmented lattice-based CN.

Typically, each arc is assigned a confidence score based on the different system confidence

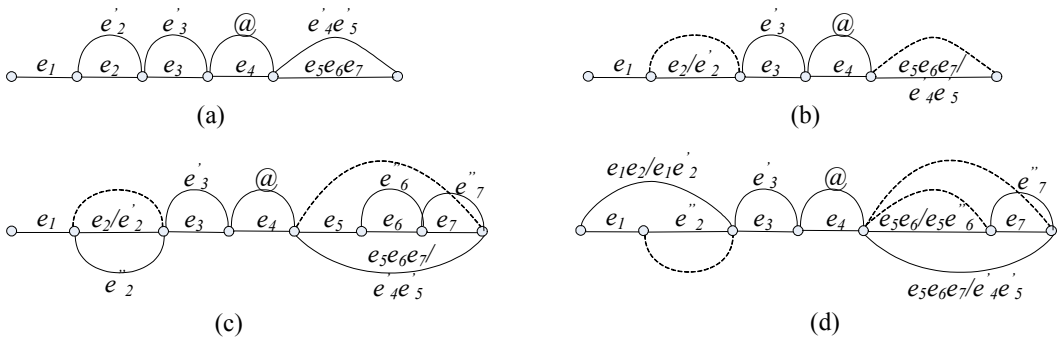

Figure 3: TERp-based Phrase-level CN using paraphrases and extended stems, synonyms

score. In our framework, we employ a weighted MBR to train the system weights. Directly, the system weights are used as the system confidence in our lattice-based phrase-level network. When two synonym arcs or paraphrase arcs are combined, then the confidence of the combined arc is the sum of the confidence scores of these two arcs weighted by the two system weights. This step ensure that the synonymous/paraphrasal arcs have a higher confidence for the purpose of selecting a more consistent phrase. The two-pass decoding algorithm is described as:

- First pass: traverse all the nodes in the lattice and find a path with the maximum probability as the candidate path;
- Second pass: trace back along the candidate path and check whether it has any combined arcs. If so, then restore all the combined arcs to a new lattice and decode all the nodes again to generate the final consensus. This step is similar to the N-best generation process in SMT decoding (Koehn, 2004).

The purpose of the first pass is to provide a selection preference of synonymous phrases or paraphrases for the decoder which can guarantee the coherence and consistency of the phrases and the context, while the second pass carries out a fair competition between the different synonyms and paraphrases which can guarantee a best fluency of the translation.

5 Experimental Settings

The experiments are conducted and reported on the NIST 2008 test data. The NIST 2006 test set includes 1,664 sentences and is used as the devset, while the NIST 2008 is used as the test set which contains 1,357 sentences. Each source sentence has 4 references in the two sets. The training data includes 2.5 million pairs of Chinese and English parallel sentences.

There are three SMT systems used in our experiments, namely, 1) baseline: Moses (Koehn et al., 2007); 2) R-HPB: our own re-implemented hierarchical phrase-based (R-HPB) system (Chiang, 2005); 3) Moses-chart: a re-implemented HPB in Moses.

In order to increase the diversity of MT systems, we also reorder the Chinese sentences using the DE classifier (Chang et al., 2009). Therefore, in our experiments, there are 6 individual systems in all which are trained on the non-reordered and reordered data. The alignment is carried out by GIZA++ (Och and Ney, 2003) and then we symmetrize the word alignment using the grow-diag-final heuristic. Parameter tuning is performed using Minimum Error Rate Training (MERT) (Och, 2003). The results of the 6 SMT systems on the NIST 2008 test set are reported in terms of BLEU (Papineni et al., 2002) and TERp scores and shown in Table 1.

In Table 1, "Baseline", "R-HPB" and "Moses-chart" indicate that the systems are trained and tested on non-reordered training data and test set. "+DE" indicates the SMT systems are built and run on a DE-reordered data set. We can see that

SYS	BLEU	TERp
Baseline	22.42	63.10
Baseline+DE	23.47	62.89
R-HPB	20.53	64.39
R-HPB+DE	22.36	63.15
Moses-chart	24.36	62.58
Moses-chart+DE	**24.75**	**62.19**

Table 1: Individual system results on the reordered and non-reordered data.

the "Moses-chart+DE" is the best individual system.

6 Experimental Results and Analysis

In this section, in order to compare the performance between 1) the weighted MBR and the uniform distributed MBR; 2) TER and TERp; 3) the word-level CN and the phrase-level CN, we perform a series of comparison experiments, as shown in Table 2.

system	**TERp**	**BLEU**
Worst Single	64.39	20.55
Best Single	62.19	24.75
TER-MBR-U	63.01	24.14
TER-MBR-W	**62.55**	**24.98**
TERp-MBR-U	62.87	24.33
TERp-MBR-W	**61.46**	**25.36**
TER Word-level CN	61.22	25.88
TERp Word-level CN	**60.71**	**26.71**
TERp-Two-pass CN	**60.24**	**27.15**

Table 2: Comparison on word-level and the phrase-level combination frameworks

In Table 2, the "TER-MBR-U" and "TER-MBR-W" indicate the TER-based MBR decoding with a uniform distribution and with a weighted distribution respectively, while "TERp-MBR-U" and "TERp-MBR-W" represent the TERp-based MBR decoder with the uniform weights and the trained system weights respectively. The "TER Word-level CN" represents the weighted word-level CN built on the TER-based alignment, and "TERp Word-level CN" is the weighted word-level CN built on the TERp alignment. In addition, the "TERp-Two-pass CN" stands for our proposed TERp-augmented phrase-level framework.

We can see that the "TER-MBR-U" and the "TERp-MBR-U" are 0.42 and 0.61 absolute BLEU points (1.7 and 2.46 relative percent) lower than the best individual system. We argue that this is caused by the distinctly different quality of the individual systems. That is, the "R-HPB" is far lower than the "Moses-chart+DE" system. However, the "TER-MBR-W" and "TERp-MBR-W" achieved 0.23 and 0.61 absolute BLEU points (0.93 and 2.46 relative percent) improvements compared to the best individual system. From these results, we analyse that if the performance between the individual systems is quite different, the discriminative MBR performs more robustly and can achieve better results than the uniform distributed MBR.

Regarding the "TER Word-level CN" and the "TERp Word-level CN", we can see that the latter obtains 0.83 absolute BLEU points (3.21 relative percent) improvement than the former. The comparison shows that the TERp metric performs better in the hypothesis alignment which attributes to the powerful capacity of synonym matching.

From Table 2, we can also find that the "TERp-Two-pass CN" outperformed any of the "Word-level CN" in terms of the BLEU and TERp scores. Our proposed framework obtained an absolute improvement by 2.4 BLEU points (9.7 relative percent) over the best single system and 0.44 absolute BLEU points (1.65 relative percent) over the "TERp Word-level CN". The results indicate that the phrase-level network can perform better than the word-level CN because it can keep the consistency of the paraphrases, and so reduce the ungrammatical errors.

7 Conclusions and Future Work

In this paper, we proposed a lattice-based phrase-level system combination framework using the TERp alignment metric. The properties of "stem match", "synonym match" and "paraphrase match" of TERp are fully used to build a phrase-level lattice-based CN. We also proposed a two-pass consensus decoding process to generate the final output. We performed a series of experiments to compare the uniform dis-

tributed MBR and a weighted MBR, TER and TERp, the word-level CN and the phrase-level CN. The experimental results conducted on the NIST 2008 set show that our proposed method significantly outperformed the word-level combination framework, and using TERp can significantly improve the combination performance.

As for future work, we need to carry out more experiments and deep analysis on the experimental results to fully explore the advantages of the TERp metric. In addition, we need to investigate one of the potential problems in lattice-based CN, which is the normalisation of arc confidence scores.

Acknowledgment

Many thanks to the reviewers for their insightful and valuable comments and suggestions. This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University.

References

Necip Fazil Ayan, Jing Zheng and Wen Wang. 2008. Improving alignments for better confusion networks for combining machine translation systems. In *Proceedings of the Coling'08*, pages 33–40.

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.

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.

Loic Barrault. 2010. MANY: Open Source Machine Translation System Combination. In *The Prague Bulletion of Mathematical Linguistics No. 93, 2010*, pages 147–155.

Pi-Chuan Chang, Dan Jurafsky and Christopher D. Manning. 2009 Disambiguating "DE" for Chinese-English machine translation. In *Proceedings of the Fourth Workshop on SMT*, pages 215–223.

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

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.

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. In *MIT Press. http://www.cogsci.princeton.edu/ wn*, [2000, September 7].

Yang Feng, Yang Liu, Haitao Mi, Qun Liu and Yajuan L. 2009. Lattice-based system combination for statistical machine translation. In *Proceedings of the EMNLP'09*, pages 1105–1113.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of the HLT-NAACL'06*, pages 455–462.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA'04*, pages 115–124.

Philipp Koehn, Hieu Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, Wade Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL'07*, pages 177–180.

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 EMNLP'07*, 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. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL'03*, pages 160–167.

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.

Martin F. Porter. 1980. An algorithm for suffic stripping. In *Program, 14(3)*, pages 130–137.

Antti-Veikko I.Rosti, Bing Zhang, Spyros Matsoukas and Richard Schwartz. 2009. Incremental hypothesis alignment with flexible matching for building confusion networks: BBN system description for WMT09 system combination task. In *Proceedings of the WMT'09*, pages 61–65.

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 AMTA'06*, pages 223–231.

Matthew Snover, Nitin Madnani, Bonnie J.Dorr and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the WMT'09*, pages 259–268.

Liang Zhou, Chon-Yew Lin and Eduard Hovy. 2006. Re-evaluating Machine Translation Results with Paraphrase Support. In *Proceedings of the EMNLP'06*, pages 77–84.