

# A Cascaded Framework for Statistical Machine Translation System Combination

**Jinhua Du**

Institute of Automation  
Chinese Academy of Sciences  
Beijing, China, 100080  
jhdu@hitic.ia.ac.cn

**Wei Wei, Zhendong Yang and Bo Xu**

Institute of Automation  
Chinese Academy of Sciences  
Beijing, China, 100080  
{weiwei, zdyang, xubo}  
@hitic.ia.ac.cn

## Abstract

This paper investigates an extensive evaluation of combination techniques and presents a cascaded framework for combining multiple Machine Translation (MT) system outputs. A Word Transition Network (WTN) is constructed from an  $N$ -best list by aligning the hypotheses against an alignment reference, where the alignment is based on minimising an modified Translation Edit Rate (TER) with word or phrase reordering. The Minimum Bayes Risk (MBR) decoding technique is investigated for the selection of an appropriate alignment reference. Pairwise word alignment is created by an enhanced statistical alignment algorithm that explicitly models word reordering. Experimental results are presented based on three MT systems of Chinese-English translation outputs. It is shown that worthwhile improvements in translation performance can be obtained using the proposed framework.

## 1 Introduction

Recently, multiple system combination is an active research topic in statistical machine translation (SMT). In the past decade, machine translation (MT) systems based on statistical theory have achieved considerable progress such as phrase-based system (Koehn et al., 2003; Och and Ney, 2004; Chiang, 2005), syntax-based systems (Yamada and Knight, 2001; Liu et al., 2006) and so on.

And phrase-based system only reflects the consistency of N-gram string while syntax-based system focuses more on syntactic structure information. So how to combine all the different information to obtain a better performance is becoming more and more significant.

Many approaches of post-processing, derived from automatic speech recognition (ASR) such as ROVER (Fiscus, 1997), Minimum Bayes-Risk decoding (MBR) (Kumar and Byrne, 2004) and so on, have been successfully applied in ASR. And in machine translation, these techniques also demonstrated powerful significance. Recently, some new improved combination strategies based on the above approaches have been proposed. For instance, (Hewavitharana et al., 2005) described the ROVER by combining N-gram language model and some features. (Matusov et al., 2006) proposed a new strategy that performed word reordering before building a word transition network. (Sim et al., 2007) presented a new alignment scheme based on minimum Translation Error Rate (TER) to construct a word-level consensus network. They all achieved a better translation quality in their experiments.

In this paper, a cascaded framework for statistical machine translation system combination is proposed. The framework integrates the MBR and word alignments techniques to process the outputs of multiple machine translation systems. It is easy for the proposed framework to combine different systems like phrase-based system, syntax-based systems and so on. Our framework includes the following three steps:

- (1) Use MBR decoder to select a best hypothesis  $T1$  as the alignment reference from the outputs of multiple translation systems.

(2) Use the alignment reference  $T_1$  to perform the word alignment with the other hypotheses by GIZA++ toolkit (Och and Ney, 2003), and then build the word transition network based on our modified TER scheme.

(3) Use beam search decoding to search the best translation from the constructed WTN.

The remainder of this paper is organized as follows: Section 2 presents a review of related work. Two typical system combination approaches that are often applied in post-processing are discussed in Section 3. Section 4 gives a detailed description of the proposed framework. In this section, we also describe our modified word alignment strategy. Experimental results on three translation systems are presented in Sections 5 using Chinese-English translation tasks. And Section 6 gives our conclusions and future work.

## 2 Related Work

Recently, more and more researches on multiple system combination or post-processing have been performed.

The most straightforward approach is that some other knowledge sources, different from those in decoding, are used to rescore for  $n$ -best outputs. For example, (Chen et al., 2005) used nine feature functions to rescore for the 1000-best translation hypotheses of each source sentence. In their experiments, the N-gram ( $n > 3$ ) language model contributes more to the system performance.

(Hewavitharana et al., 2005) used ROVER to build a WTN to get a consensus translation. They also introduced a modified ROVER that combined with a language model (Schwenk and Gauvain, 2000). In these experiments, single ROVER without other knowledge sources does not demonstrate an excellent capability of tuning the system performance well, and moreover, it weakens the performance to some extent. In fact, ROVER is initially designed for reducing word error rate for automatic speech recognition without word reordering.

(Jayaraman and Lavie, 2005) presented a versatile word alignment algorithm to produce a word-to-word alignment. In conjunction with confidence estimates and a language model, these alignments were used to select a new hypothesis. Their alignment algorithm was motivated by (Bangalore et al., 2001) where the edit distance was used to produce

monotone alignments. (Matusov et al., 2006) used a statistical alignment which can carry on word reordering, where an algorithm was proposed to train the word alignments and proceed the word reordering. In their algorithm, because every hypothesis was used to play the role of the alignment reference to get a word alignment, the complication of training is increased and more noises by some worse hypotheses are probably introduced.

(Kumar and Byrne, 2004) used Minimum Bayes-Risk decoding to choose a best translation from the  $n$ -best list under some evaluation metric. And they presented three types of translation loss functions. In their experiments, the best performance was achieved when the same loss function was used in both the error rate and the decoder design, that is, when the loss function is BLEU, the performance under the BLEU metric is best.

(Sim et al., 2007) proved that MBR decoding was useful for alignment reference selection. They also considered a different scheme (TER) for aligning hypotheses and presented convictive comparison experimental results.

In the following sections, we will discuss two critical techniques applied in MT in detail.

## 3 Two Typical Combination Methods

In this section, we will discuss two critical techniques used in system combination from a theoretical point of view. We will introduce how to use ROVER to build a WTN, and furthermore, investigate how the MBR chooses a minimal cost hypothesis.

### 3.1 ROVER

ROVER was developed by J. Fiscus of NIST. It seeks to reintegrate word sequences to decrease word error rates by exploiting the difference of the word errors made by multiple speech recognizers in ASR. ROVER mainly includes the following two steps:

- (1) An output of one of the multiple systems is designated as the base, and the other outputs are aligned to this base to build a single word transition network. Figure 1 shows an example of a word transition network. The word sequence begins with  $\langle s \rangle$  and ends with  $\langle /s \rangle$ . The symbol of "@" which means an insertion or deletion in one position is no-cost word.

- (2) A best word sequence with the highest score (with the highest number of votes) is selected from this WTN.

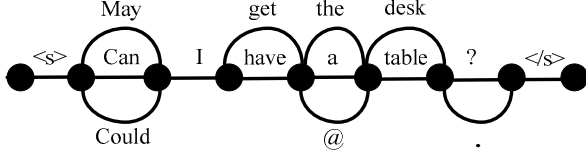


Figure 1 Word Transition Network

Given a WTN, the best scored translation  $\hat{i}$  is selected according to Equation (1) (Fiscus, 1997):

$$\begin{aligned} \hat{i} &= \arg \max \sum_{n=1}^L P_r(w_n) \\ &= \arg \max \sum_{n=1}^L N(w, n) / N_s \end{aligned} \quad (1)$$

where  $L$  is the length of the word sequence,  $p_r(w)$  is a relative probability of  $w$ ,  $N(w, n)$  is the number of occurrence of  $w$  in column  $n$ , and  $N_s$  is the number of combined systems.

(Schwenk and Gauvain, 2000) proposed an improved ROVER. They used language model information to provide contextual information to decrease the perplexity of word sequence. (Hewavitharana et al., 2005) also presented an improved ROVER as Equation (2):

$$\hat{i} = \arg \max \prod_{n=1}^L P_r(w_n) P(w_n / w_{n-1} \dots w_1)^\lambda \quad (2)$$

where  $p(w_n / w_{n-1} \dots w_1)$  is a N-gram language model.  $\lambda$  is the scale factor.

In our framework, we use the principle of ROVER to build a Word Transition Network according to a selected alignment reference based on an improved alignment scheme. And beam search decoding is adopted to get the consensus translation.

### 3.2 Minimum Bayes-Risk Decoding

(Kumar and Byrne, 2004) used Minimum Bayes-Risk decoding to minimize expected loss of translation errors to get a best hypothesis. And they introduced three lexical loss functions like BLEU, WER and PER (Och, 2002). The BLEU loss function is:

$$L_{BLEU}(E, E') = 1 - BLEU(E, E') \quad (3)$$

where  $E$  is the reference translation, and  $E'$  is the hypothesis translation.

Given the loss function, the Minimum Bayes-Risk decoder described in (Kumar and Byrne, 2004) is applied as

$$\hat{i} = \arg \min_{i \in \{1, 2, \dots, N\}} \sum_{j=1}^N L(E_j, E_i) P_r(E_j / F) \quad (4)$$

where  $\hat{i}$  is the translation with minimal cost,  $N$  is the size of  $n$ -best hypotheses.  $P_r(E_j / F)$  is the posterior probability, and is approximated using an  $n$ -best list as

$$P(E_j / F) = \frac{P(E_j, F)}{\sum_{i=1, \dots, N} P(E_i, F)} \quad (5)$$

where  $P_r(E_j, F)$ , is the joint probability of the source and target sentence, could be assigned by the translation model. (Sim et al., 2007) assumed the posterior probability distribution was uniform.

Equation (4) shows a rescoring procedure whose goal is to search the best translation from a  $n$ -best set.

## 4 Proposed Cascaded Framework

According to the above analysis, we present a cascaded framework for system combination. The framework is initially motivated by the work of (Matusov et al., 2006) and (Sim et al., 2007). The former used GIZA++ to train word alignment and then performed the word reordering before build a WTN; the latter used MBR decoding to select an alignment reference and then performed word alignment based on TER metric. In our proposed framework, we propose an improved alignment strategy called GIZA-TER that uses GIZA++ to train word alignment and then use TER metric to minimize the number of edit distance. In our scheme, the shift operation in TER is different from that in (Sim et al., 2007). We perform phrase shift operation based on the alignment point generated by GIZA++.

The flowchart of the framework is shown in Figure 2. The number 1, 2 and 3 represent each step in the whole framework.

In the following subsection, we will describe each level of our framework.

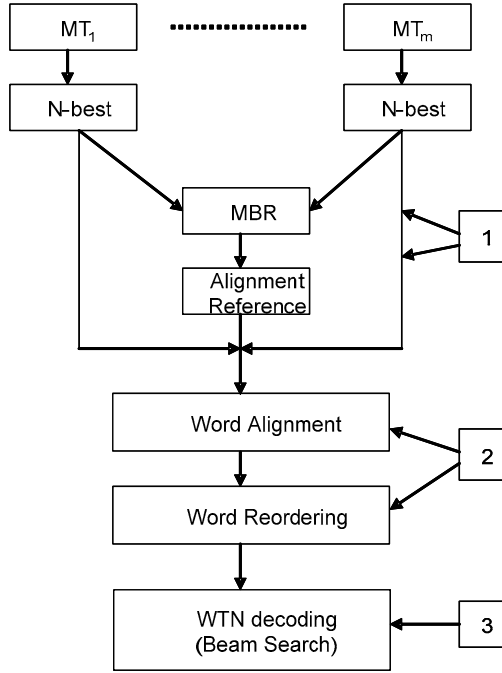


Figure 2. Cascaded Framework of Multiple System Combination

#### 4.1 Select the Alignment Reference

This section is illustrated as the part of the number 1 in Figure 2. Given  $M$  MT engines of which each engine produces a  $n$ -best list, we need to combine all the translation hypotheses and use MBR decoding to select one as the alignment reference.

The alignment reference is achieved with the following steps:

- Combine all these  $n$ -best lists. Use BLEU as the loss function, then the MBR decoding is performed using Equation (4) to generate one translation as the alignment reference with the minimal risk.
- Use the alignment reference and the other hypotheses to train word alignment by the GIZA++ toolkit.

#### 4.2 Word Alignment for Word Reordering

GIZA++ toolkit is used to train the word alignment. Similar to the training process in statistical machine translation, the hypotheses pair is regarded as the source and target sentence. Accord-

ing to the algorithm of IBM model, the alignment between the two hypotheses is expressed as:

$$P_r(E_b, A / E_m) = P_r(A / E_b) P_r(E_b / A, E_m) \quad (6)$$

where  $E_b$  is the alignment reference selected by MBR decoder,  $E_m (m=1,2,\dots,n)$  is a hypothesis in the above hypotheses set  $H$ ,  $A$  is the set of the alignment between these two sentences,  $P_r(E_b / A, E_m)$  is the lexicon probability, and  $P_r(A / E_b)$  is the alignment probability.

In the process of training, we use the IBM model 4 to estimate the alignment model. And we perform the training with the two directions  $E_b \rightarrow E_m$  and  $E_m \rightarrow E_b$ . Two alignment relations are finally obtained:

$$A_1 = \{(i, a_i) | a: i \rightarrow j\}$$

$$A_2 = \{(b_j, i) | b: j \rightarrow i\}$$

$a: i \rightarrow j$  is a alignment function from  $E_b \rightarrow E_m$  and  $b: j \rightarrow i$  is an alignment function from  $E_m \rightarrow E_b$ .

To get a set of refined alignment, the intersection of the two alignment relations  $A_1$  and  $A_2$  are computed as

$$P = A_1 \cap A_2$$

The intersection relation  $P$  represents a high-precision alignment

We use a large English-English lexicon in the training to improve the accuracy of word alignment.

#### 4.3 Building WTN and Retrieving the Consensus Translation

In Section 4.2, alignment matrix between two sentences is obtained. We use this alignment matrix to perform word or phrase reordering by minimizing the Word Error Rate between a hypothesis and the alignment reference, which we call GIZA-TER. This approach is similar to the method described in (Sim et al., 2007), but the difference is that we use GIZA++ to obtain the word alignment and then perform the phrase shift (reordering) while they used greedy search to perform phrase shift.

Here, a simple example of creating a WTN by GIZA-TER scheme is provided for illustration in Figure 3.

Original hypotheses	1. Economic Indicators Will Show That the European Economy Remained Weak Euro Area this Week 2. Euroland Economic Indicator Will Show That Europe Economy Remained Weak this Week
Alignment Pair by GIZA++	{1:2 2:3 3:4 4:5 5:6 7:7 8:8 9:9 10:10 13:11 14:12}
Original WER:	1. ***** Economic Indicators Will Show That the European Economy Remained Weak Euro Area this Week 2. Euroland Economic Indicator Will Show That ***** Europe Economy Remained Weak ***** this Week Edit Distance: Subs-2; Ins-1; Del-3      WER:42.86%
Phrase Shift according to Alignment Pair	1. Economic Indicators Will Show That the European Economy Remained Weak Euro Area this Week 2. Economic Indicator Will Show That Euroland Europe Economy Remained Weak this Week
WER after Phrase Shift	1. Economic Indicators Will Show That the European Economy Remained Weak Euro Area this Week 2. Economic Indicator Will Show That Euroland Europe Economy Remained Weak ***** this Week Edit Distance: Subs-3; Ins-0; Del-2      WER:35.71%
WTN	Economic Indicators Will Show That the European Economy Remained Weak Euro Area this Week Economic Indicator Will Show That Euroland Europe Economy Remained Weak @ @ this Week

Figure 3 Example and flowchart of creating a WTN by GIZA-TER scheme. Sentence 1 is the alignment reference. The symbol of @ denotes a null alignment.

In Figure 3, although word “Indicators “ in Sentence 1 is different from the word “Indicator” in Sentence 2, the phrase shift can be performed according to alignment pair from 1:2 to 5:6; while in standard TER scheme, this shift can not be performed. So GIZA-TER has stronger capability for phrase reordering than standard TER.

## 5 Experiments

In this section, we will perform a set of experiments and verify how the proposed framework works. Experiments results are presented on two domains. One is on the Chinese-to-English Basic Travel Expression Corpus (BTEC) task, and the other is on the NIST newswire task.

In the experiments of system combination, 3 statistical machine translation systems are used, including a typical phrase-based system, a string-to-tree system and a hierarchical phrase-based system.

### 5.1 Corpus Statistics

The two domains are spoken language translation domain and newswire domain respectively. We select Chinese-to-English translation tasks: NIST MT-05 (newswire domain) and IWSLT-05 (spoken language domain).

#### NIST MT-05.

In this task, the bilingual training data comes from the Hong Kong News and Hansards corpus as well as Xinhua Parallel News. In the process of training the language model, besides the monolingual part of the above corpus, we also add the Xinhua News portion of Gigaword.

#### IWSLT-05.

In this domain, the translation task is on small data track. All provided bilingual corpra are used to train the translation model, and the monolingual portion to train the language model. Table 1 shows the statistics in these two tasks.

Corpus	#Sent.	#Eng. Voc	#Chn. Voc
NIST-05	2M	93K	117K

IWSLT-05	20K	7308	9070
----------	-----	------	------

Table 1. Statistics in NIST-05 and IWSLT-05 training corpora.

## 5.2 Multiple Machine Translation Systems

3 statistical machine translation systems that are used in our experiments are listed in Table 2.

Sys	Type	Feature	Expansion
P-B	Phrase-based	8	Non-Monotone
S-T	String-to-Tree	6	CKY
Hiero	Hierarchical	6	CKY

Table 2. Multiple Machine Translation Systems

A detailed illustration about individual system is as follows:

- System P-B is a phrase-based system with variable-based alignment template (Wei et al., 2006). The 8 features are the bidirectional phrase translation probabilities  $p(e/f)$  and  $p(f/e)$ , the bidirectional lexical probabilities  $l(e/f)$  and  $l(f/e)$ , the language model, the distortion model, word penalty and phrase penalty.
- System S-T is a string-to-tree system. In extracting alignment templates, two variables of  $X1$  and  $X2$  are introduced. The decoder is CKY decoder. The 6 features are the same to System P-B except for word and phrase penalty.
- System Hiero is a hierarchical phrase-based system. Like in System S-T, two variables of  $X1$  and  $X2$  are used to extract phrases. The 6 features are the same to System S-T.

System P-B uses a 4-gram language model, and the other two systems use a 3-gram language model trained by SRILM toolkit (Stolcke, 2002).

## 5.3 Experimental Setup

We perform a set of experiments to implement the following three tasks:

- To prove MBR decoding is useful for selecting a better hypothesis from the outputs of multiple systems.

- To prove that selecting an alignment reference as the base to build a WTN is very important for the final performance.
- To prove the effectiveness of our proposed alignment method for WTN. In this experiment, we will compare our alignment method with WER and TER.

## 5.4 Task 1: MBR Decoding for Combination

SYS <sup>1</sup>	IWSLT05	NIST05
P-B	0.4770	0.2308
S-T	0.4305	0.2137
Hiero	0.4691	0.2188
MBR-BLEU	0.4817	0.2402
+weights	0.4880	0.2473

Table 3 BLEU scores for individual systems and MBR decoding for combination

Table 3 shows the BLEU performance of individual systems and MBR decoding of the 1-best translation from each system. MBR-BLEU denotes MBR decoding using BLEU loss function. For MBR decoding, there are only three 1-best hypotheses to select from and the posterior distribution is assumed to be uniform. +weights denotes that the posterior distribution is tuned as system weights that sum to one. When tuned to maximize the BLEU scores, 0.63% and 0.71% absolute improvements in BLEU were obtained. The improvements prove that the MBR decoding is useful for combination.

## 5.5 Task 2: Comparative Experiments on Selecting an Alignment Reference

In this task and the next Task 3, each system provides 3-best hypotheses, and MBR decoder is used to select an alignment for word transition network. In Table 4, the effect of alignment reference on WTN was examined. Using P-B 1-best hypothesis as alignment reference yields better BLEU performance compared to using the S-T and Hiero 1-best hypothesis. When using the output from MBR-BLEU (with tuning) as the alignment reference, 1.72% and 1.4% improvements on BLEU were obtained over using the P-B 1-best output. These results suggest that MBR decoding

<sup>1</sup> Our S-T system and Hiero system were just developed, so they didn't demonstrate better performance in these experiments. We now go on improving their performance.

is useful for alignment reference selection. In this task, TER alignment method was used.

Alignment Reference	IWSLT-05	NIST-05
	BLEUr16n4	BLEUr4n4c
P-B	0.4720	0.2377
S-T	0.4278	0.2223
Hiero	0.4569	0.2273
MBR-BLEU	0.4892	0.2517

Table 4. Comparison of alignment references for WTN using TER alignment

### 5.6 Task 3: Comparison of Alignment Methods for WTN

Another important factor which greatly influences the performance of the system combination is the alignment method used to construct the network. Here, the WER, TER and GIZA-TER alignment methods were compared in Table 5. It was found that using GIZA-TER alignment yields better BLEU (0.18-1.05%) and (0.15-0.45%) than using WER and TER respectively.

Alignment Method	IWSLT-05	NIST-05
	BLEUr16n4	BLEUr4n4c
WER	0.4805	0.2487
TER	0.4892	0.2517
GIZA-TER	0.4910	0.2532

Table 5. Comparison of alignment methods for WTN using MBR-BLEU (+weights) output as alignment reference

In Table 5, the result of GIZA-TER was obtained under our proposed cascaded framework, that is, we use MBR-BLEU output as the alignment reference to train the GIZA++ alignment matrix and then use GIZA-TER method to build a WTN, and finally search a consensus translation.

## 6 Conclusion and Future Work

In this paper, we investigate deeply the methods of post-processing in machine translation. And from various comparison experiments, we learn more how to employ a combination algorithm in different situations. Simultaneously, on the basis of related work, we propose a modified alignment

method of GIZA-TER and a complete framework of system combination. Through a set of detailed experiments, our framework is proved to be effective and significant.

In future, we will introduce word lattice to perform the system combination. Word lattice includes more useful information to rescore and search a consensus translation.

## References

- S. Bangalore, G. Bordel, G. Riccardi. 2001. Computing Consensus Translation from Multiple Machine Translation Systems. *IEEE Workshop on Automatic Speech Recognition and Understanding*.
- David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. *In Proceedings of ACL 2005*, pp. 263–270.
- B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo and M. Federico. 2005. The ITC-irst SMT System for IWSLT-2005. *In Proceedings of IWSLT 2005*.
- George Doddington. 2002. Automatic evaluation of machine translation quality using N-gram co-occurrence statistics, *In Proc. ARPA Workshop on Human Language Technology*.
- J. G. Fiscus. 1997. A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER). *IEEE Workshop on Automatic Speech Recognition and Understanding*.
- Sanjika Hewavitharana, Bing Zhao, Almut Silja Hildebrand, Matthias Eck, Chiori Hori, Stephan Vogel and Alex Waibel. 2005. The CMU Statistical Machine Translation System for IWSLT2005. *In the proceedings of International Workshop on Spoken Language Translation (IWSLT 2005)*.
- S. Jayaraman and A. Lavie. 2005. Multi-Engine Machine Translation Guided by Explicit Word Matching. *10th Conference of the European Association for Machine Translation*, pp. 143-152.
- Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. *In Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas*, pp. 115–124.
- Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. *In Proceedings of HLT-NAACL 2003*, pp. 127-133.

- S. Kumar and W. Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of HLT-NAACL*.
- Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of 39<sup>th</sup> Annual Meeting of the ACL*, pp. 523-530
- Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-String Alignment Template for Statistical Machine Translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, pages 609-616.
- E. Matusov, N. Ueffing and H. Ney. 2006. Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. In *Proceedings of EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*, pp. 33-40.
- Franz J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417-449.
- F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19-51.
- F. J. Och. 2002. Statistical Machine Translation: From Single Word Models to Alignment Templates. Ph.D. thesis, RWTH Aachen, Germany.
- Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 901-904.
- K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division.
- H. Schwenk and J. -L. Gauvain. 2000. Improved ROVER using language model information. In *Proceedings of the ISCA ITRW Workshop Automatic Speech Recognition: Challenges for the new Millennium*, pp. 47-52.
- K.C. Sim, W. Byrne, M. Gales, H. Sahbi and P. Woodland. Consensus Network Decoding For Statistical Machine Translation System. In *the 32nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hawaii, USA, 2007*
- Wei Wei, Wei Pang, Zhendong Yang, Zhenbiao Chen, Chengqing Zong, Bo Xu. 2006. CASIA SMT System For TC-STAR Evaluation Campaign 2006. In *Proceedings of TC-STAR Workshop on Speech-to-Speech Translation*, pp. 69-74.