The CASIA Statistical Machine Translation System for IWSLT 2008

Yanqing He, Jiajun Zhang, Maoxi Li, Licheng Fang,

Yufeng Chen, Yu Zhou and Chengqing Zong

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, 100190, China
{yqhe,jjzhang,mxli,lcfang,chenyf,yzhou,cqzong}@nlpr.ia.ac.cn

Abstract

This paper describes our statistical machine translation system (CASIA) used in the evaluation campaign of the International Workshop on Spoken Language Translation (IWSLT) 2008. In this year's evaluation, we participated in challenge task for Chinese-English and English-Chinese, BTEC task for Chinese-English. Here, we mainly introduce the overview of our system, the primary modules, the key techniques, and the evaluation results.

1. Introduction

This paper describes the statistical machine translation system of CASIA (Institute of Automation, Chinese Academy of Sciences), which is used for the evaluation campaign of the International Workshop on Spoken Language Translation (IWSLT) 2008. We participated in challenge task for Chinese-English and English-Chinese, BTEC task for Chinese-English.

Our system combines the output results of multiple machine translation systems. These systems are listed as follows:

- Three phrase-based statistical machine translation (SMT) models: Moses decoder (MOSES) [1], an in-home phrase-based decoder (PB) [2] and a sentence type-based reordering decoder (Bandore) [3];
- Two formal syntax-based translation models: a hierarchical phrase-based model (HPB) [4] and a maximum entropy-based reordering model (MEBTG)[5];
- A linguistically syntax-based translation model: a syntax-augmented machine translation (SAMT) decoder [6].

Then by using some global features we rescore the combination results to get our final translation outputs.

This paper is structured as follows: Section 2 presents the overview of CASIA system. In Section 3, the experimental results of our system are reported and the details on analyses of the results are given. Section 4 gives the conclusions.

2. System Overview

Figure 1 depicts our system architecture. After the test data are preprocessed, they are passed into multiple translation systems respectively to produce an N-Best translation list, and then all the N-Best translations in the list are combined to obtain a new *N*-Best in the combination module and the final 1-Best translation is selected by rescoring the new *N*-Best. We post-process the best translation to get the final translation results. We will detail each module as follows:

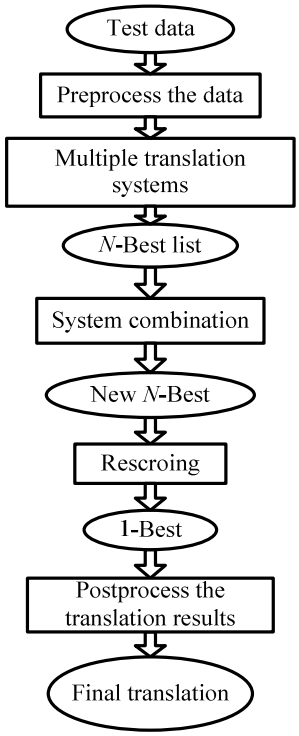

Figure 1: Our system architecture.

2.1. Preprocessing

For the Chinese part of the training data, development data and test data, two types of preprocessing are performed:

- Segmenting the Chinese characters into Chinese words using the free software toolkit ICTCLAS3.0[1];
- Transforming the SBC case into DBC case;

For the English part of the training data and development data and test data, also two types of preprocessing are performed:

- Tokenization of the English words: which separates the punctuations with the English words;
- Transforming the uppercase into lowercase.

[1] http://www.nlp.org.cn

2.2. Multiple translation systems

2.2.1. *Three phrase-based SMT systems*

Phrase-based translation systems are usually modeled through a log-linear model [7]. In the log-linear model, given the sentence f (source language), the translating process is searching the translation e (target language) with the highest probability. The translation probability and the decision rule are given as Formula (1).

$$e^* = \arg\max_{e} \sum_{m=1}^{M} \lambda_m h_m(e, f) \quad (1)$$

Where $h_m(e,f)$ is a feature function and λ_m is the weight of the feature. The entire λ_m are obtained by the minimum error rate training [8].

We use three phrase-based machine translation systems, Moses system (MOSES) [1], an in-home phrase-based system (PB) [2] and a sentence type-based reordering model (Bandore) [3].

The Moses decoder provided in the open source Moses package[1] is run by the default parameters. We only train 3-gram language model and extract phrase pairs no more than 10 words.

Our in-home PB system's word alignment is based on the training results of the GIZA++[2] toolkit under the default parameters. We obtain word alignment by the method of grow-diag-final on the bi-directional word alignments of GIZA++. PB's phrase extraction is same with Moses with the maximum length 10. We use the following features in a monotone decoding process:

- Phrase translation probability $p(\tilde{e} \mid \tilde{c})$;
- Lexical phrase translation probability $lex(\tilde{e} \mid \tilde{c})$;
- Inversed phrase translation probability $p(\tilde{c} \mid \tilde{e})$;
- Inversed lexical phrase translation probability $lex(\tilde{c} \mid \tilde{e})$;
- English language model based on 3-gram $lm(e_1^I)$;
- English sentence length penalty I;
- Chinese phrase count penalty N.

Bandore is a sentence type-based reordering model, which divides the Chinese sentences into three types and employs different reordering model for each sentence type. Bandore serves as a preprocessing module for SMT system. Firstly, SVM is used to classify Chinese sentences into three types: special interrogative sentences, other interrogative sentences and non-question sentences. We directly use all the words occurring in the sentence as features. Secondly, corresponding reordering model is developed for specific sentence types. Phrase-ahead model is employed for special interrogative sentences and phrase-back model is employed for other sentence types. The framework of Bandore is illustrated in Figure 2, where *C1* means the special interrogative sentences, *C2* is other interrogative sentences and *C3* is non-question sentences.

[1] http://www.statmt.org./moses/

[2] http://www.fjoch.com/GIZA++.html

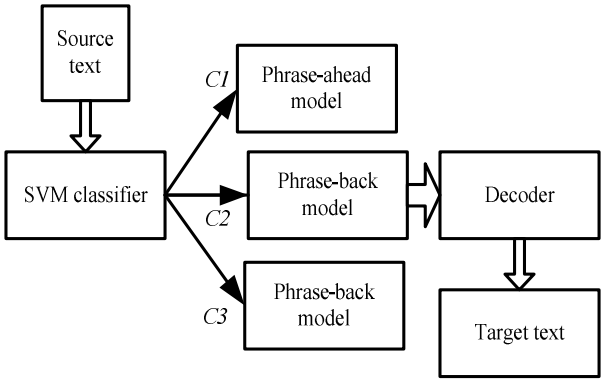

Figure 2: Architecture of Bandore.

For the Chinese special interrogative sentence, there is a fixed phrase that usually occurs at the end of Chinese sentence but appears at the beginning part of its English translation. We define such special question phrase (*SQP*) as the syntactic component containing the key word in special interrogative sentences. The key words, listed in Table 1, are found from corpus by mutual information. Let *S* be a Chinese special interrogative sentence, we utilize a CRF toolkit named FlexCrfs [9] to train, test and predict the *SQP*s chunking. If we have known the *SQP*, *S* becomes S^0 *SQP* S^1 where S^0 is the left part of the sentence before *SQP*, and S^1 is the right part of the sentence after *SQP*. We note that there are only three positions where the *SQP* will be moved to: (1) the beginning of the sentence; (2) just after the rightmost punctuation3 before the *SQP*; (3) or after a regular phrase such as "请问 (May I ask)" and "告诉 我 (Please tell me)". Therefore, we have learned the reordering templates from bilingual corpus to find the right position in S^0 where *SQP* will be moved to.

Table 1: The special key words set.

什么	What
哪 (哪里 / 哪儿…)	Where
多 (多大 / 多长…)	How much/many/old…
怎么 (怎么办 / 怎么样…)	How
怎样	What about
谁 (谁的 / 是谁…)	Who/whose/whom
几 (几点 / 几个…)	How many/old When…
为什么	Why
何(何时 / 何地…)	When/where

For Chinese other interrogative sentences and non-question sentences, we only consider the *VP* (verb phrase) modifiers *PP* (prepositional phrase), *TP* (time phrase) and SP (spatial phrase) as triggers, and the first *VP* occurring after triggers will be the candidate position where the triggers may be moved to. To deal with the case that there is no *VP* in a sentence due to recognition error, we define a fake verb phrase (*FVP*): the phrase after *PP* (*TP* or *SP*) until the punctuation (",", ";" or "."). Here, *FVP* is given the same function with *VP*, thus it makes our model suitable for more

[3] The punctuation is ",", ";" or ":" in Chinese.

situations. We develop a probabilistic reordering model to alleviate the impact of the errors caused by the parser when recognizing *PP*s, *TP*s, *SP*s and *VP*s. The form of phrase-back reordering rules:

$$A: A_1 X A_2 \Rightarrow \begin{cases} A_1 X A_2 & straight \\ X A_2 A_1 & inverted \end{cases}$$

where $A_1 \in \{PP, TP, SP\}$, $A_2 \in \{VP, FVP\}$ and $X \in$ $\{phrases\ between\ A_1\ and\ A_2\}$. We use Maximum Entropy Model [10] which is trained from bilingual spoken language corpus to determine whether A_1 should be moved after A_2. The features that we investigate include the leftmost, rightmost, and their POSs of A_1 and A_2. It leads to the following formula:

$$P(O \mid A) = \frac{\exp(\sum_i \lambda_i h_i(O, A))}{\sum_O \exp(\sum_i \lambda_i h_i(O, A))}$$

where $O \in \{straignt, inverted\}$, $h_i(O, A)$ is a feature, and λ_i is the weight of the feature.

After reordering the Chinese sentences of training set and test set, we pass the reordered sentences into a phrase-based decoder such as Moses or PB decoder to get the final translation results. In our experiments Bandore uses Moses as its decoder.

2.2.2. Two formal syntax-based translation models

Here we use two formal syntax-based translation models, a maximum entropy-based reordering model (MEBTG) [5] and a hierarchical phrase-based translation model (HPB) [4].

The system of MEBTG is realized in home according to [5] and [11]. In this model the prediction of relative orders of any two adjacent blocks is considered as a problem of classification. We extract reordering examples from the word-aligned training corpus and extract the following features from every two consecutive phrase pairs:

- Lexical features: the last word of two source phrases or target phrases;
- Collocation features: the combination of lexical features.

With these features we train a MaxEnt classifier[1]. We extract phrase pairs using Och's algorithm [12]. The maximum length of source phrase is limited in 10 words. We use a CKY style decoder which limits the phrase table within 40 and the partial hypotheses is within 200.

HPB translation engine is a re-implementation of David Chiang's hierarchical phrase translation model. Based on the union of the bidirectional alignment results of GIZA++, initial rules consistent with the alignment are extracted using Och's algorithm [12] and then rule subtraction is performed to obtain rules with no more than two non-terminals. Null-aligned words are allowed at the boundary of phrases. We set a limitation that initial rules are of no more than 10 words and other rules should have no more than 5 terminals and non-terminals. The decoder is CYK-style chart parser that maximizes the derivation probability. A 3-gram language model generated by SRILM is used in the cube-pruning process. The search space is pruned with a chart cell size limit of 30. To limit the number of rules applicable to a single sentence, a maximum of 50 is set for rules with the same source side. Threshold pruning is also used to prune the translation hypotheses which are worse than the current best hypothesis in the cell by a factor of 10. Minimum error rate training [8] is used to tune the BLEU score on a development set.

2.2.3. A linguistic syntax-based translation model

For the Chinese-English task, we used the latest version of the syntax-augmented machine translation (SAMT) source under the GNU General Public License[2]. We extract phrases no more than 10 words and run the decoder with default parameters.

2.3. System combination

We implement system combination on our *N*-Best list from multiple translation systems. The overall framework of system combination is shown in Figure 3.

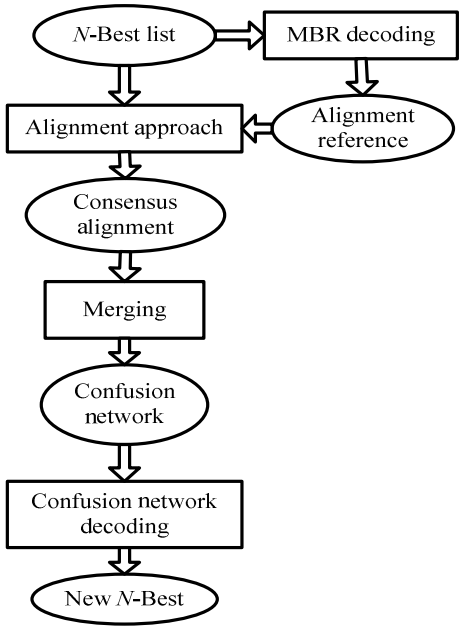

Figure 3: System combination architecture.

We collect the *N*-Best list translation hypotheses from each translation system in Section 2.2, and find a hypothesis as the alignment reference with the minimum Bayes risk [13]. We exploit word reordering alignment approaches to align all the hypotheses against the alignment reference and form a consensus alignment. Given *N* (*N*=3) translation hypotheses:

please show me on this map .
please on the map for me .
show me on the map , please .

when the first translation hypothesis is chosen as the alignment reference, the result of consensus alignment may look something like Figure 4, where "null" strings are used to accommodate insertions and deletions.

[1] http://maxent.sourceforge.net/

[2] http://www.cs.cmu.edu/~zollmann/samt

null	please	show	me	on	this	map	.
null	please	for	me	on	the	map	.
,	please	show	me	on	the	map	.

Figure 4: An example of consensus alignment.

After obtaining consensus alignment, by merging similar words being aligned together at the same position and assigning each word an alignment score based on a simple voting scheme, it forms a confusion network. In Figure 5, an example of confusion network is provided.

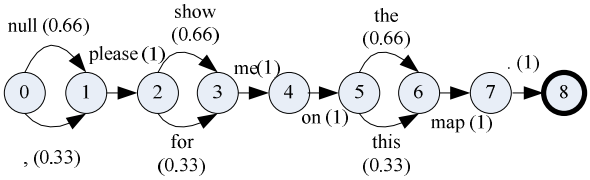

Figure 5: An example of confusion network.

With the language model feature and word penalty introduced, we will get the final translation by confusion network decoding. The decoding process may be written as

$$e^* = \arg\max_e(\lambda_{AL}\log P_{AL} + \lambda_{LM}\log P_{LM} + \lambda_{WP}\log P_{WP})$$

where λ_{AL} , λ_{LM} , λ_{WP} are the weights of the alignment feature, language model feature, and words penalty and they are constrained to sum to one. The probability P_{AL} , P_{LM} , P_{WP} represent alignment score, language model, and word penalty. The argmax operation denotes the search problem, that is, the generation of the new *N*-Best after combining the input *N*-Best list hypotheses from multiple translation systems.

2.4. Rescoring

Because we have employed several different SMT systems and combination technology, the local feature functions of each translation hypothesis cannot be used in the rescoring module. Therefore, we should use the global feature functions to score the new *N*-Best generated by system combination. In our experiments we set *N*=200. The 9 global functions we apply are listed as follows and most of them are referred to [14].

- Direct and inverse IBM model 1 and model 3 [14].
- 2, 4, 5-gram target language model.
- 3, 4, 5-gram target pos language model.
- Bi-word language model [15].
- Length ratio between source and target sentence.
- Question feature [14].
- Frequency of its *n*-gram (*n*=1, 2, 3, 4) within the *N*-Best translations.
- N-gram posterior probabilities within the *N*-Best translations [16].
- Sentence length posterior probabilities [16].

The weights of the feature functions are optimized by downhill simplex algorithm which is implemented by us. After the rescoring on the new *N*-Best we obtain the 1-Best translation for each input source sentence.

2.5. Post-processing

The post-processing for the output results mainly includes:

- Case restoration in English words;
- Recombination the separated punctuations with its left closest English words;
- Segmenting the Chinese outputs into characters.

3. Experiments

Experiments were carried out on each track task. We will describe each step in detail and give our analysis on the experimental results.

3.1. Corpus

Besides the training data provided by IWSLT 2008, we collected all the data from the website[1]. All the linguistical sources used in our experiments are listed in Table 2.

Table 2: Training data list

Names	Sentence pairs
IWSLT2008	**19,972**
LDC2005T06	**10,317**
LDC2005T10	**282,176**
LDC2002T01	**993**
LDC2003T17	**878**
LDC2004T07	**935**
LDC2006T04	**919**
LDC2004T08	**1,767,609**
LDC2002L27	**82,099**
LDC2005T34	**2,345,276**
HIT-corpus	**132,514**
CLDC-LAC-2003-004	**304,502**
CLDC-LAC-2003-006	**200,082**
Chinese LDC (2004-863-008)	**52,227**
Total	**5,200,499**

Table 3: The detailed statistics of our corpus for development set

Track	Data		Sen.	Running words	Voc.
CT CE CRR	**Train set**	**Chi**	**324,626**	**2.4M**	**11,214**
		Eng	**324,626**	**2.57M**	**9,488**
	Dev set	**Chi**	**534**	**3,163**	**649**
		Eng	**3,204**	**22,861**	**1,132**
CT EC CRR	**Train set**	**Chi**	**311,438**	**2.28M**	**11,113**
		Eng	**311,438**	**2.42M**	**9,370**
	Dev set	**Chi**	**2275**	**15,266**	**797**
		Eng	**325**	**2,061**	**404**
BTEC CE CRR	**Train set**	**Chi**	**321,770**	**2.38M**	**11,202**
		Eng	**321,770**	**2.51M**	**9,493**
	Dev set	**Chi**	**764**	**4,899**	**910**
		Eng	**4,584**	**34,310**	**1,536**

We extract the bilingual data which are highly correlative with the training data of each track. Given a track task, if all the words in a sentence pair are all falling into the word

[1] http://www.slc.atr.jp/IWSLT2008/archives/2008/10/resources.html#LinguisticResources

vocabulary of the training data of the track, we add such sentence pair into the training data of the track. After the filtering step, we obtain the training data of each track. Then based on the test data of each track, we also filter some development sentences and their reference sentences from all the released development data of the track as our development data according to the similarity calculation. Because the development sets provided by IWSLT08 have different numbers of multiple reference translations, we only choose the minimum number of reference translations, such as 6. We use the English side or the Chinese side of the filtered training set of each track task to train language model by SRILM. In our experiment, we only use 3-gram language model.

For the CRR translation of each track, we first obtain our model parameters of each translation system by the minimum error rate training on the development data filtered according to the above principle. The detailed statistics of our development data are shown in Table 3. Here "CT" means challenge task, "BTEC" means BTEC task and "CE" or "EC" respectively denote the translation direction from Chinese to English or from English to Chinese. "Sen." denotes sentence pair and "Voc." denotes the vocabulary of words. After the model parameters are obtained on development set, we add the development set of each track into the training set to form the final training set. The detailed statistics of the corpus for test set are given in Table 4.

Table 4: The detailed statistics of the corpus for test set

Track	Data		Sen.	Running words	Voc.
CT CE CRR	Train set	Chi	349,297	2.55M	11,358
		Eng	349,297	2.74M	9,713
	Test set	Chi	504	3,098	377
CT EC CRR	Train set	Chi	314,185	2.36M	11,269
		Eng	314,185	2.44M	9,448
	Test set	Eng	498	3,529	310
BTEC CE CRR	Train set	Chi	347,554	2.54M	11,356
		Eng	347,554	2.73M	9,707
	Test set	Chi	507	3,531	870

3.2. ASR translation

For the ASR translation of each track, we first translate the ASR *N*-Best list. For our experiments the value *N*=5 is used and we translate the 5-Best of ASR output to get 1-Best translation of each translation system. In each translation system we just translate ASR output by using the parameters and data trained on CRR translation for the same track. To use the features of acoustic model and source language model, we pass these translation results into our combination module and rescore all the translation hypotheses with the feature functions of translation hypotheses plus the features of ASR to get the final 1-Best result of ASR of the track.

3.3. Dealing with the named entities

The test data includes some named entities such as person names, location names, organization names, numbers and dates. If we ignore such named entities, much useful information will be lost. It will result in worse translation result. Aiming at such named entities, we first identify and extract them from the test data [17] and then deal with them individually with their different characters.

- For the person names and location names, we translate them only by looking up its translations in the common phrase pair table which is obtained from the training data on word alignments;
- For the organization names, we translate them using the model based on a synchronous CFG grammar [18];
- For the numbers and dates, we adopt the method based on the man-written rules to translate.

For SAMT and MOSES decoder we add all the named entity translation pairs into their training data. For other decoders we add all the named entity translation pairs in the phrase pair table or rule table with all the probabilities as 1.0.

3.4. Experimental results

For each track we participant in, we give the experimental results on development set shown from Table 5 to Table 7. Here "PB" represents our in-home phrase-based translation system used in IWSLT 2007. MOSES system and SAMT are free toolkits from website. MEBTG and HPB are realized in home according to [5] [11] [4]. Bandore is newly-developed machine translation model by our lab. "COM" means system combination and "Rescore" represents the rescoring module.

For each translation system we extract 10-Best translations for each source input, with duplicates found in each *N*-Best. We do punctuation insertion before feeding sentences into decoders by using *hidden-ngram* command in SRILM toolkit. All the scores on development set are computed based on case non-insensitive and without punctuation.

Table 5: Results of development set for CT_CE track

	CRR		ASR	
	BLEU	NIST	BLEU	NIST
PB	0.4505	7.4649	0.4732	7.4777
MOSES	0.5048	7.9175	0.4980	7.7488
Bandore	0.5033	8.0267	0.4651	7.4983
MEBTG	0.4571	7.6887	0.4969	7.8267
HPB	0.4412	6.8600	0.4536	7.4474
COM	0.5109	8.1780	0.5093	8.0045
Rescore	0.5741	8.3162	0.5787	8.7570

Table 6: Results of development set for BTEC_CE track

	CRR		ASR	
	BLEU	NIST	BLEU	NIST
PB	0.4659	7.9333	0.4831	7.8623
MOSES	0.5100	8.0298	0.4870	7.4720
Bandore	0.5127	8.3513	0.4856	7.7699
MEBTG	0.4717	7.8045	0.4915	7.7357
HPB	0.4764	6.5603	0.4445	5.9105
COM	0.5308	8.5689	0.5087	8.0778
Rescore	0.6100	8.7823	0.5235	8.2364

Table 7: Results of development set for CT_EC track

	CRR		ASR	
	BLEU	NIST	BLEU	NIST
PB	0.4385	7.0469	0.4350	7.3629
MEBTG	0.4399	7.5303	0.4569	7.5691
MOSES	0.4522	7.3626	0.4676	7.5165
HPB	0.4298	7.0914	0.4544	7.5165
COM	0.4555	7.6200	0.4578	7.5600
Rescore	0.5242	7.7361	0.5011	7.9627

Table 8 shows the systems combined on development set for each track. In the experiments on development set SAMT is not used because it needs longer time on larger training data. So we only run SAMT on the test data based on the released training data in each track by IWSLT 2008. Bandore and SAMT can not be applied to the tracks from English to Chinese.

Because the development sets of each track released for CRR translation and ASR translation may be different, in our experiment we use a different development set for ASR from CRR. So the ASR score may be higher than CRR in a same translation system for the same track.

Table 8: systems for combination on development set.

	CT_CE		CT_EC		BTEC_CE	
	CRR	ASR	CRR	ASR	CRR	ASR
PB			√	√		√
MOSES	√	√	√	√	√	√
Bandore	√	√			√	
MEBTG	√	√	√	√	√	√
HPB					√	√

Table 9 gives the experimental results on test set for each track we participated in. Here "Con1" denotes our system combination and "Con2" represents the rescoring module. For the reason that we cannot judge clearly which one is better than the other one, we RE-rescore "Con1" and "Con2" to choose the better one as our primary results by using the feature of the prior probability of the length-ratio of source sentence to target sentence in training corpus.

Table 9: Results of test set for each track

Track	System	CRR		ASR	
		BLEU	NIST	BLEU	NIST
CT CE	Primary	0.4844	7.5859	0.4066	6.6384
	Con1	0.4803	7.4277	0.3750	6.3134
	Con2	0.4767	7.4237	0.4067	6.5887
CT EC	Primary	0.5122	7.3513	0.4312	6.6867
	Con1	0.4968	7.1525	0.4172	6.4864
	Con2	0.4817	6.7254	0.4162	6.4713
BTEC CE	Primary	0.5077	8.5389	0.4339	7.7247
	Con1	0.4842	8.4094	0.4303	7.6550
	Con2	0.5162	8.2884	0.4318	7.6203

Table 10: systems for combination on test set.

	CT_CE		CT_EC		BTEC_CE	
	CRR	ASR	CRR	ASR	CRR	ASR
PB			√	√		
MOSES	√	√	√	√	√	√
Bandore	√	√			√	√
MEBTG	√	√	√	√	√	√
HPB	√	√			√	√
SAMT	√	√			√	√

Table 10 shows the systems combined on test set for each track. We use as many as systems to combination module. Here we only give the performance on test set which score are computed based on case insensitive and with punctuation.

3.5. Experimental analyses

From the translation results on development set, we find that the translation systems newly added almost outperform the PB system used for IWSLT'07. Table 11 gives the best performance relatively compared with PB decoder among the scores on development set. Bandore has an outstanding performance among the three systems. One reason is that it uses Moses as its decoder. In all the translation system MOSES has a performance with considerable robust. Another reason is that the reordering model of Bandore aims at the speech language. So Bandore has an effective ability to the domain of IWSLT 2008. In Table 5 and Table 6 Bandore failed to achieve significant improvements over Moses because Moses itself has lexicalized reordering features.

Table 11: systems comparison

System	Compared with PB
Bandore	11.72%
MEBTG	5.03%
HPB	4.45%

For comparison of systems combination and rescoring module, we give Table 12 to illustrate their performance. We compare the two modules respectively with PB, MOSES and the best system in each track and compute the best improvement. From Table 12 we can see system combination and rescoring module are effective in our experiments. Especially the rescoring module has a relative improvement of 30.93% on PB and 19.6% on MOSES. From Table 5, 6, 7 and 9 the performance of rescoring module is surprising on the development set while it has a modest performance on test set

Table 12: the performance of system combination and rescoring

Module	PB	MOSES	best translation system	Com
Com	13.91%	4.45%	3.5%	-
Rescore	30.93%	19.6%	18.98%	15.18%

because the rescoring module has a development set to tune its weights and so the Bleu score on same development set is certainly higher.

4. Conclusions

In summary, this paper presents our statistical machine translation system in IWSLT 2008 evaluation campaign. Our system combines the output results of multiple machine translation systems, such as 1) three phrase-based SMT model: Moses decoder, an in-home phrase based decoder and a sentence type-based reordering decoder; 2) two formal syntax-based translation models: a hierarchical phrase-based model and a maximum entropy-based reordering model; 3) A linguistic syntax-based translation model: a syntax-augmented machine translation decoder. Then by using some global features we rescore the combination results to get our final translation outputs.

The translation result proves that the combination module and rescoring module are effective in the SMT system. But there are much more space for us to ameliorate. In our experiment we only translate the 5-Best ASR output and much information of ASR such as word lattice need to be mined for ASR translation. Even for CRR translation we will add the semantic information into our model to obtain better translations.

5. Acknowledgments

The research work described in this paper has been funded by the Natural Science Foundation of China under Grant No. 60575043 and 60736014, the National Key Technology R&D Program under Grant No. 2006BAH03B02, the Hi-Tech R&D Program ("863" Program) of China under Grant No. 2006AA01Z194 and 2006AA010108-4, and Nokia Research Center, Beijing as well.

6. References

[1] 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, " Moses: Open Source Toolkit for Statistical Machine Translation", *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Poster Session, pp. 177-180,* Prague, Czech Republic, June 2007.

[2] Yu Zhou, Yanqing He, and Chengqing Zong, The CASIA Phrase-Based Statistical Machine Translation System for IWSLT 2007. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, October 15-16, 2007, Trento, Italy.

[3] Jiajun Zhang, Chengqing Zong, Shoushan Li. 2008. Sentence Type Based Reordering Model for Statistical Machine Translation. To appear in the *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, August 18-22, 2008. Manchester, UK.

[4] David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2): 201-228.

[5] Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum Entropy Based phrase reordering model for statistical machine translation. In *proceedings of COLING-ACL*, Sydney, Australia.

[6] A. Zollmann and A. Venugopal, "Syntax augmented machine translation via chart parsing," in *Proc. of the Workshop on Statistical Machine Translation, HLT/NAACL*, June 2006.

[7] F. J. Och, and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models", *Computationa Linguistics, volume 29, number 1, pp. 19-51* March 2003.

[8] Ashish Venugopal, Stephan Vogel. Considerations in Maximum Mutual Information and Minimum Classification Error training for Statistical Machine Translation. In *the Proceedings of the Tenth Conference of the European Association for Machine Translation (EAMT-05)*, Budapest, Hungary May 30-31, 2005.

[9] Xuan-Hieu Phan, Le-Minh Nguyen, and Cam-TuNguyen. 2005. "FlexCRFs: Flexible Conditional Random Field Toolkit". http://flexCRF.sourceforge.net

[10] Le Zhang. 2004. Maximum entropy modeling toolkit for python and c++. http://homepages.inf.ed.ac.uk /s0450736 / maxent_toolkit.

[11] Deyi Xiong, Min Zhang, Aiti Aw, Haitao Mi, Qun Liu and Shouxun Lin, Refinements in BTG-based Statistical Machine Translation, In *Proceedings of IJCNLP 2008.*

[12] Franz J. Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417-449

[13] S. Kumar and W. Byrne, "Minimum Bayes-risk decoding for statistical machine translation," in *Proc. of HLT*, 2004.

[14] Boxing Chen, Jun Sun, Hongfei Jiang, Min Zhang, Ai Ti Aw, I^2R Chinese-English Translation System for IWSLT 2007, In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, October 15-16, 2007, Trento, Italy.

[15] Pi-Chuan Chang and Kristina Toutanova. 2007. A Discriminative Syntactic Word Order Model for Machine Translation. *In Proceedings of 45th Meeting of the Association for Computational Linguistics* .

[16] Richard Zens and Hermenn Ney. 2006. N-gram Posterior Probabilities for Statistical Machine Translation. In *Proceedings of the HLT-NAACL Workshop on Statistical Machine Translation.*

[17] Youzheng Wu, Jun Zhao, Bo Xu, Chinese Named Entity Recognition Model Based on Multiple Features. In *Proceedings of HLT/EMNLP 2005*, pages: 427~434, October 6-8, Vancouver, B.C., Canada.

[18] Yufeng Chen, and Chengqing Zong. A Structure-based Model for Chinese Organization Name Translation. *ACM Transactions on Asian Language Information Processing*, 7(1): 1-30, February 2008