I^{2}R's Machine Translation System for IWSLT 2009

Xiangyu Duan, Deyi Xiong, Hui Zhang, Min Zhang and Haizhou Li

Institute for Infocomm Research, Singapore
{xduan, dyxiong, mzhang, hli}@i2r.a-star.edu.sg zhangh1982@gmail.com

Abstract

In this paper, we describe the system and approach used by the Institute for Infocomm Research (I^{2}R) for the IWSLT 2009 spoken language translation evaluation campaign. Two kinds of machine translation systems are applied, namely, phrase-based machine translation system and syntax-based machine translation system. To test syntax-based machine translation system on spoken language translation, variational systems are explored. On top of both phrase-based and syntax-based single systems, we further use rescoring method to improve the individual system performance and use system combination method to combine the strengths of the different individual systems. Rescoring is applied on each single system output, and system combination is applied on all rescoring outputs. Finally, our system combination framework shows better performance in Chinese-English BTEC task.

1. Introduction

This paper describes the machine translation (MT) system and approach explored by the Institute for Infocomm Research (I^{2}R) for the International Workshop on Spoken Language Translation (IWSLT) 2009.

Basically, our MT system is a system combination framework. System combination [1, 2, 3] has demonstrated its advantage in the recent machine translation evaluation campaign [4, 5]. In our system combination framework, we adopt mainly two kinds of statistical machine translation (SMT) methods: phrase-based SMT and syntax-based SMT. For syntax-based system, we developed three variations. Totally, we applied four SMT systems. Based on outputs of four single systems, we applied rescoring method [4] to incorporate rich global features. Finally, we adopt two kinds of system combination methods, namely, n-gram expansion [3] and weighted voting, on all rescoring outputs.

The rest of paper is organized as follows. Section 2 presents each individual SMT system used in our framework. Section 3 details the rescoring method. Section 4 describes two system combination strategies: n-gram expansion and weighted voting. Section 5 reports the experimental setups and results while Section 6 concludes the paper.

2. The SMT Models

To integrate the advantages of the state-of-the-art translation methods, we use two different SMT systems, phrase-based, and BTG-based systems. The two systems share some common features: word alignment of training data obtained from Berkeley alignmentor [18], Language model(s) (LM) trained using SRILM toolkit [7] with modified Kneser-Ney smoothing method [8].

2.1. Lavender: Phrasal Translation System

Lavender [19] is our newly-developed in-house SMT translation platform, including a phrase-based decoder and most of the current linguistically motivated syntax-based system. Its phrase-based component, which functions very similar to Moses [12], is used as the phrase-based decoder for this campaign. Phrase-based SMT usually adopt a log-linear framework [9]. By introducing the hidden word alignment variable *a* [10], the optimal translation can be searched for based on the following criterion:

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

where $\tilde{e}$ is a string of phrases in the target language, $\tilde{f}$ is the source language string of phrases, $h_m(\tilde{e}, \tilde{f}, a)$ are feature functions, weights λ_m are typically optimized to maximize the scoring function [11]. IBM word reordering constraints [13] are applied during decoding to reduce the computational complexity. The other models and feature functions employed by Lavender are:

- Translation model(s) (TM), direct and inverse phrase/word based translation model
- Distortion model, which assigns a cost linear to the reordering distance, the cost is based on the number of source words which are skipped when translating a new source phrase
- Lexicalized word reordering model [14] (RM)
- Word and phrase penalties, which count the numbers of words and phrases in the target string

The translation model, reordering model and feature weights are trained and optimized using Moses training and tuning toolkits [12].

2.2. Tranyu: Syntax-based Translation System

Tranyu is our another in-house translation platform. It is a formally syntax-based SMT system, which adapts the bracketing transduction grammars (BTG) for phrase translation and reordering. The BTG lexical rules (A --> x/y) are used to translate source phrase x into target phrase y while the BTG merging rules (A --> [A, A]|<A, A>) are used to combine two neighboring phrases with a straight or inverted order. All these rules are weighted with various features in a log-linear form. For lexical rules, phrase/lexical translation probabilities in both directions, word/phrase penalties, as well as the language model are used as features. For merging rules, we incorporate maximum entropy (MaxEnt) based reordering models to predict orders between two neighboring phrases. We train all the model scaling factors on the development set to maximize the

BLEU score. A CKY-style decoder is developed to generate the best BTG binary tree for each input sentence, which yields the best translation.

We develop three variations of Tranyu. Each variation is tuned independently on the development set. All variations share the same phrase table, language model and boundary word based reordering model. We give brief introductions of these variations as follows.

- Tranyu(Bound). In this variation, we use a boundary word based reordering (BWR) model [20] to predict phrase orders for merging rules. We define boundary words as words at the begining/ending positions of source/target sides of two neighboring phrases.Supposing the left phrase pair is "于 7 月 15 日|on July 15", the right phrase pair is "举行 总统 与 国会 选举| held its presidential and parliament elections", source words {“ 于”, “15 日”, “举行”, “选举” } and target words {“on”, “15”, “held”, “elections”} are boundary words. Training a BWR model proceeds through 3 steps. First, we extract reordering examples from word-aligned bilingual data, then generate features using boundary words of these examples and finally estimate feature weights.

- Tranyu(LAR). In order to employ more linguistic knowledge in the ITG reordering, we extend boundary word based reordering further by linguistically annotating each node involved in reordering according to the source-side parse tree. We call this linguistically annotated reordering (LAR). In LAR, we annotate each BTG node with three annotation elements: (1) head word, (2) the part-of-speech (POS) tag of head word and (3) syntactic category. We use these three elements, together with boundary words described above, as our reordering features. The weights of these features are tuned using a MaxEnt trainer. For more details, please refer to [21].

- Tranyu(UniBrack). Syntactic analysis influences the way in which the source sentence is translated. In this variation, except for the reordering model BWR, we incorporate a syntax-driven bracketing model (UniBrack) which predicts whether a phrase (a sequence of contiguous words) is bracketable or not using rich syntactic constraints. If a source phrase remains contiguous after translation, we refer this type of phrase {\bf bracketable}, otherwise {\bf unbracketable}. We parse the source language sentences in the word-aligned training corpus. According to the word alignments, we define bracketable and unbracketable instances. For each of these instances, we automatically extract relevant syntactic features from the source parse tree as bracketing evidences. Then we tune the weights of these features using a maximum entropy trainer. For more details, please refer to [22].

To further improve reordering between two neighboring phrases, we introduce two hard constraints. The first one is the swapping window, which only allows reordering within a predefined window (we set the window size to 15 words on the source side). The second one is the punctuation restriction, which prohibits any inverted orders if two neighboring phrases include any of the punctuation marks {， 、 ： ； 「 」 《 》 （ ） “ ”}. For more details, please refer to [23]. The two constraints are implemented in three Tranyu variations described above.

3. Rescoring Models

Rescoring operation plays a very important role in our system. A rich global feature functions set benefits our system greatly. The rescoring models are the same ones which were used in our SMT system for IWSLT 2007 [4]. We apply the following feature functions. Weights of feature functions are optimized by the MERT tool in Moses package.

- direct and inverse IBM model 1 and 3
- association scores, i.e. hyper-geometric distribution probabilities and mutual information
- lexicalized reordering rule [15]
- 6-gram target language model and 8-gram target word-class based LM, word-classes are clustered by GIZA++
- length ratio between source and target sentence
- question feature
- Linear sum of n-grams (n=1,2,3,4) relative frequencies within all translations, which favors the hypotheses containing popular n-grams of higher order [16]
- n-gram posterior probabilities within the N-best translations [17]
- sentence length posterior probabilities [17]

4. System Combination

After rescoring, we perform system combination. In our system combination framework, two different system combination strategies are used in a two-stage procedure to find the final translation. They are n-gram expansion in the first stage, and weighted voting in the second stage.

4.1. N-gram Expansion

N-gram expansion [3] combines the sub-strings occurred in the rescored N-best translations to generate new hypotheses. Firstly, all n-grams from the rescored N-best translations are collected. Then the partial hypotheses are continuously expanded by appending a word through the n-grams collected in the first step.

During the new hypotheses generation step, the translation outputs are computed through a beam-search algorithm with a log-linear combination of the feature functions. In addition to n-gram frequency and n-gram posterior probability used in [3], we follow the suggestion of [16] and also use language model, direct/inverse IBM model 1, and word penalty in this work.

4.2. Weighted Voting

Given all 1-best hypotheses generated from different systems, the final 1-best translation is selected by weighted voting. In our weighted voting, a binary feature function is used to indicate the system in which hypothesis is generated from. Note that we have four individual systems and one combination strategy of n-gram expansion. Totally, we have five systems to vote. The feature weight of each system is tuned over the development set.

5. Experiments

We participate Chinese-to-English BTEC task (BT) of IWSLT 2009.

5.1. Data

Experiments are carried out on the Basic Traveling Expression Corpus (BTEC) Chinese-English data provided by the IWSLT 2009 organizer. Given the IWSLT 2009 provided train set and several development set, we re-divide them in the following way. We use official train set (20k), devset2_IWSLT04, devset3_IWSLT05, devset6_IWSLT07 all together as our train set, devset1_CSTAR03 as our development set, and devset7_IWSLT08 as our test set. Language model is trained on English side of training data. All the experiments hereafter are based on the above setting unless specified otherwise.

5.2. Preprocessing

Preprocessing includes Chinese word segmentation, English tokenization and lower-casing. The word segmentation tool we used is developed by NUS [24]. Punctuation insertion is not performed because gold standard punctuations are not available in training data.

5.3. Post-processing

The evaluation of IWSLT'09 is case sensitive. To reduce data sparseness, we lowercase the target language in the preprocessing step. Thus, a case restoration post-processing step is required to recover the correct case information. It is done on the final MT output by using disambig tool from SRILM toolkit. Besides, we also remove OOV before case restoration.

5.4. Experimental Results

Our evaluation metric is BLEU, which is to perform n-grams matching up to n=4. Please note that all in-house evaluation BLEU scores are computed with closest, case-insensitive option, and with punctuation.

5.4.1. Baseline Performance and Effectiveness of Adding Development Set into Training Set

We use Lavender as our baseline system to tune the basic settings. We find that Berkeley alignmentor [18] performs much better than Giza++ in the BTEC data used in the IWSLT 2009. Thus, we adopt Berkeley alignmentor in our experiment. After tuning on the development set, we add development set into training set to re-train the translation model and reordering model. Table 1 reports the baseline performance and demonstrates the effectiveness of adding development set into training set.

Table 1: Baseline performance using training set only and adding dev set to training set

Alignmentor		***Baseline***	***+dev***
Berkeley	***Dev***	0.4630	-
	Tst	0.4526	**0.4629**

5.4.2. Effectiveness of Combination of Phrase Tables

In phrase-based system training, there are several widely-used heuristics in building bi-directional word alignments for phrase generation. For Berkeley alignment, we investigate combination of their only two provided heuristics of "grow" and "grow-diag". Table 2 shows that the combination of the two heuristics is helpful for performance improvement.

Table 2: Performance of combination of Berkeley alignments

	grow	***grow-diag***	***grow +grow-diag***
Dev	0.4630	0.4609	0.4635
Tst	0.4526	0.4472	0.4598
Tst (+dev)	0.4629	0.4523	**0.4732**

Then we further investigate the combination of Giza++ and Berkeley alignments. Various combinations between them are explored, and we find that combining Giza++'s grow-diag-final-and and Berkeley's grow+grow-diag achieves the best performance (as shown in Table 3) among all this kinds of combinations. But it is still lower than the best performance in Table 2. Therefore, we use Berkeley's grow+grow-diag as our word alignments in our experiments, on which previously mentioned four SMT systems are trained.

Table 3: Performance of combination between Giza++ and Berkeley's alignments

	Giza++_grow-diag-final-and + Berkeley_grow+grow-diag
Dev	0.4760
Tst	0.4636
Tst (+dev)	0.4710

5.4.3. Effectiveness of Rescoring

Given the models, we translate the test set using all our four MT systems. Then we rescore these outputs with additional features. The performance is shown in Table 4. Row "before" reports performance before rescoring while row "after" reports performance after rescoring.

Table 4: Performance of rescoring

		Lavender	***Tranyu: Bound***	***Tranyu: UniBrack***	***Tranyu: LAR***
Dev	***before***	0.4635	0.4719	0.4478	0.4597
	after	0.4858	0.4951	0.4882	0.5008
Tst	***before***	0.4598	0.4521	0.4471	0.4594
	after	0.4618	0.4715	0.4743	0.4760
Tst (+dev)	***before***	0.4732	0.4604	0.4572	0.4589
	after	0.4799	0.4755	**0.4816**	0.4790

Rescoring on the development set improves performance dramatically, but rescoring on test set behaves differently between phrase-based system and syntax-based systems. On test set, rescoring on Lavender's outputs improves performance marginally while rescoring on syntax-based systems improves performance dramatically. UniBrack achieves the best performance after rescoring.

5.4.4. Effectiveness of System Combination

We adopt two kinds of system combination: n-gram expansion and weighted voting. Besides the four single systems' outputs, we add n-gram expansion's outputs into weighted voting frame. Their performances are shown in Table 5, where UniBrack is used as baseline since it is the best single system after rescoring.

Table 5: Performance of system combination

	Dev	*Tst*	*Tst (+dev)*
UniBrack	0.4882	0.4743	0.4816
n-gram expan	0.5106	0.4841	0.4880
weighted vote	0.5185	0.4897	**0.4944**

5.4.5. Effectiveness of Post-processing

Table 6 reports the performance in BLEU score in our test set after post-processing. Note that we add dev set into training set in this experiment. We can see that post-processing improves the performance further from 0.4944 (See Table 5) to 0.5033 in Bleu score.

Table 6: Performance of post-processing

Case-insensitive	*Case-sensitive*
0.5033	0.4900

5.4.6. Official Automatic Evaluation Result

Finally we add all the dev sets and our test set into our training set and train a new model as the final model for the IWSLT 2009 evaluation. We denote it as "train+d+t". We use weighted vote output trained on "train+d+t" as our primary run result, #weighted vote output trained on "train+d" as our contrastive1 run, n-gram expansion trained on "train+d+t" as our contrastive2 run, and UniBrack trained on "train+d+t" as our contrastive3 run. Table 7 shows BLEU score (case-sensitive+punc) of official automatic evaluation on our submissions.

Results show that contrastive1 is slightly better than primary. However, the difference is not significant. N-gram expansion (contrastive2) performs much worse than UniBrack (contrastive3). We will do more experiments to study the reason.

Table 7: Official automatic evaluation result

Primary	*Contrastive1*	*Contrastive2*	*Contrastive3*
0.4595	0.4599	0.4441	0.4527

6. Conclusions

This paper describes I^2R's SMT system that is used in the IWSLT 2009 MT campaign. We use a system combination framework that incorporates mainly two kinds of our in-house SMT systems: phrase-based and syntax-based systems in the IWSLT 2009. We explain the details of our experiments and report how we achieve the final performance from single systems to the combined systems step by step.

7. References

[1] E. Matusov, Nicola Ueffing, and Hermann Ney. 2006. "Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment". In Proceeding of EACL-2006, Trento, Italy.

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

[3] B. Chen, M. Federico and M. Cettolo, "Better N-best Translation through Generative n-gram Language Model", Proceeding of MT Summit XI, Copenhagen, Denmark, September, 2007.

[4] B. Chen, Jun Sun, Hongfei Jiang, Min Zhang and Aiti Aw. "I2R Chinese-English Translation System for IWSLT-2007", Proceeding of IWSLT 2007. pp. 55-60. Oct. Trento, Italy.

[5] The MSR-NRC-SRI MT System for NIST Open Machine Translation 2008 Evaluation. Available at http://research.microsoft.com/~xiaohe/publication/NIST_MT08_sys_desc_MSR-NRC-SRI_Chinese.pdf

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

[7] A. Stolcke, "SRILM -- an extensible language modeling toolkit", *Proceeding of International Conference on Spoken Language Processing*, 2002.

[8] S. F. Chen and J. T. Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. *Technical Report TR-10-98*, Computer Science Group, Harvard University.

[9] F. J. Och, and H. Ney. "Discriminative Training and Maximum Entropy Models for Statistical Machine Translation." In *Proceeding of ACL-2002*. 2002.

[10] P. F. Brown, V. J. Della Pietra, S. A. Della Pietra & R. L. Mercer. "The Mathematics of Statistical Machine Translation: Parameter Estimation." Computational Linguistics, 19(2) 263-312. 1993.

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

[12] 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." In *Proceedings of ACL-2007*. pp. 177-180, Prague, Czech Republic. 2007.

[13] A. L. Berger, P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, A. S. Kehler, and R. L. Mercer. "Language translation apparatus and methods using context-based translation models". US Patent 5,510,981. 1996.

[14] P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne and D. Talbot. "Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation." In *Proceeding of IWSLT-2005*.

[15] Boxing Chen, Mauro Cettolo and Marcello Federico, "Reordering Rules for Phrase-based Statistical Machine Translation", *Proceeding of International Workshop on Spoken Language Translation*, pp. 182-189, Kyoto, Japan, November, 2006.

[16] Boxing Chen, R. Cattoni, N. Bertoldi, M. Cettolo and M. Federico, "The ITC-irst SMT System for IWSLT-2005",

Proceeding of International Workshop on Spoken Language Translation, pp.98-104, Pittsburgh, USA, October, 2005.

[17] R. Zens and H. Ney, "N-gram Posterior Probabilities for Statistical Machine Translation", *Proceedings of the HLT-NAACL Workshop on Statistical Machine Translation*, pp. 72-77, New York City, NY, June 2006.

[18] P. Liang, B. Taskar and D. Klein, "Alignment by Agreement", Proceedings of North American Association for Computational Linguistics (NAACL), 2006.

[19] Min Zhang et al., "Lavender: I2R Statistical Machine Translation Platform", Technical-report-2009-008, Institute for Infocomm Research, 2009.

[20] Deyi Xiong, Qun Liu, and Shouxun Lin. "Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation". *In Proceedings of COLING-ACL 2006*, Sydney, Australia.

[21] Deyi Xiong, Min Zhang, Ai Ti Aw, and Haizhou Li. "A Linguistically Annotated Reordering Model for BTG-based Statistical Machine Translation". *In Proceedings of ACL 2008*.

[22] Deyi Xiong, Min Zhang, Aiti Aw and Haizhou Li. "A Syntax-Driven Bracketing Model for Phrase-Based Translation". *In Proceedings of ACL-IJCNLP 2009*.

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

[24] Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo , "A Maximum Entropy Approach to Chinese Word Segmentation". *In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*