The MSRA Machine Translation System for IWSLT 2010

Chi-Ho Li, Nan Duan[1], Yinggong Zhao[2], Shujie Liu[3], Lei Cui[3]

Natural Language Computing, Microsoft Research Asia
49 Zhichun Road, Beijing 100190, China

Mei-yuh Hwang, Amittai Axelrod[4], Jianfeng Gao, Yaodong Zhang[5], Li Deng

Natural Language Processing, Microsoft Research
One Microsoft Way, Redmond WA98052, United States

chl@microsoft.com

Abstract

This paper describes the systems of, and the experiments by, Microsoft Research Asia (MSRA), with the support of Microsoft Research (MSR), in the IWSLT 2010 evaluation campaign. We participated in all tracks of the DIALOG task (Chinese/English). While we follow the general training and decoding routine of statistical machine translation (SMT) and that of MT output combination, it is our first time to try our ideas in post-processing output of automatic speech recognition (ASR) before feeding it to SMT decoders. Our findings are: (1) it does not help to use the complete N-best ASR output; rather, the best translation performance is achieved by taking the top one candidate after Minimum Bayes Risk re-ranking of the N-best ASR output; (2) as to punctuation recovery, the best performance is achieved by splitting the problem into two steps, viz. the prediction of punctuation position and the prediction of punctuation given a position.

1. Introduction

This paper is a description of all system modules and the associated experiments used by MSRA for its very first participation in the IWSLT evaluation exercise. In the 2010 campaign we took part in the following tracks of the DIALOG task:

a) Chinese-to-English / CRR
b) Chinese-to-English / ASR
c) English-to-Chinese / CRR
d) English-to-Chinese / ASR.

Here CRR refers to correct speech recognition output while ASR refers to automatic speech recognition output. This distinction is about the nature of the input to the MT system.

The structure of the paper is as follows. Section 2 summarizes the various modules in the MSRA SMT system, including the translation decoders and the MT output combination module. Section 3 explains a few important techniques for the best IWSLT performance. Section 4 elaborates how we tackle two special problems when taking ASR output as SMT input, viz. lack of punctuation, and output in the form of N-best list. Finally, Section 5 reports the experiments done in preparation of the 2010 evaluation.

2. SMT System Architecture

Figure 1 illustrates the basic framework of the MSRA SMT system. Essentially it can be divided into two phases. The first phase is translation by individual MT decoders, and the second phrase is MT output combination. Between the two phases is re-ranking of translation hypotheses produced by the individual decoders. For the ASR tracks, there is an additional process of selecting the best candidate from the N-best list of ASR output.

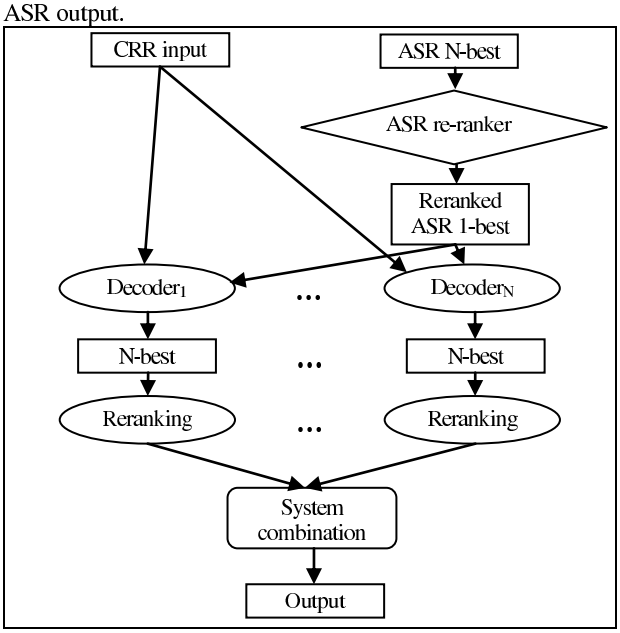

Figure 1: The MSRA SMT System Architecture.

[1] This author is an intern from Tianjin University.
[2] This author was an intern from Nanjing University.
[3] These authors are interns from Harbin Institute of Technology.
[4] This author was an intern from University of Washington.
[5] This author was an intern from Massachusetts Institute of Technology.

2.1. MT Decoders

A wide variety of MT decoders are used to maximize the performance gain by MT output combination, including:

1) Moses [1].
2) BTG. This is an in-house phrase-based decoder with a maximum entropy based, lexicalized reordering model proposed by [2].
3) Hiero. This is an in-house re-implementation of hierarchical phrase-based SMT [3].
4) DepHiero. This is similar to Hiero but augmented with a dependency tree language model as in [4]. Such language model is learned from the dependency trees of the target side of the bilingual training dataset as parsed by the Berkeley Parser.
5) DepBTG. This is similar to BTG but, again, augmented with a dependency tree language model.
6) Syntax. This is an in-house syntax-based decoder based on [5] and [6]. The minimal GHKM and SPMT rules are extracted from bilingual training dataset whereas the composed rules are generated by combining two or three minimal rules. The target side of the training dataset is parsed by the Berkeley Parser and the parse trees are binarized by the method in [7].
7) Treelet [8]. Note that this decoder is used for English-to-Chinese translation only.

2.2. MT Output Combination

The translation hypotheses from the individual decoders in Section 2.1 are fed to an MT output combination module for producing the final translation. The hypotheses are first aligned and then converted to a confusion network using incremental HMM alignment [9]. The confusion network is then decoded in the conventional way as in [10]. The incremental HMM alignment requires probabilistic bilingual dictionaries, which are obtained from HMM alignment over the given bilingual training dataset.

Note that all text is lower-cased at the very beginning of the entire MT routine. The final translation from the MT output combination module will be passed to a case restoration module, which uses a simple language model based method [11].

3. Useful Techniques for IWSLT

There have been many techniques which were reported to improve performance in IWSLT exercises in previous years. We tried a number of them and found that the following three techniques are of particular importance.

3.1. Translation of Numeral/Temporal Expressions

Numeral and temporal expressions are too many to be covered by training data yet their variations can be well handled by a handful of rules. Therefore, these expressions are first identified by manually written rules. During training phase, all tokens of numerals are replaced by a special symbol, and similarly for all tokens of temporal expressions. Such treatment greatly improves the quality of word alignment and phrase/rule extraction. During decoding, the numeral and temporal tokens are translated by manual rules as well.

3.2. Combination of Word Alignment

Different word aligners commit different kinds of mistakes, and therefore we may lessen the impact of alignment errors by considering the alignments produced by several word aligners. We simply apply each word aligner to the bilingual training data and then merge the phrases/rules extracted from the alignment matrices by all these aligners. Note that this is itself a kind of weighing/voting mechanism, since the alignment links agreed by more aligners will be counted for more times. The word aligners used include GIZA++ [12], MSRA ITG Aligner [13], and MSRA Discriminative Aligner (which is similar to the model in [14] but the parameter training is MERT [15]).

3.3. Re-ranking of Translation Hypotheses

Minimum Bayes Risk (MBR) decoding has received more and more attention in recent years. In order to keep our decoders efficient, we do not apply the MBR technique to the decoding process; rather, we do MBR re-ranking of translation hypotheses produced by each decoder. The re-ranking is a log-linear model with the features:

1) N-gram posterior probabilities [16].
2) Sentence length posterior probabilities [17].
3) N-gram language model probabilities.
4) Length ratio between source sentence and translation hypothesis.

The feature weights are tuned by MERT, on certain development set with reference translations.

4. Taking ASR Output as SMT Input

The input to the SMT system in the ASR tracks has two characteristics. First, it is subject to ASR error. Thus we tried to lessen the impact of ASR error by taking the N-best ASR output in consideration. Secondly, the ASR output does not contain punctuation at all.[1] Hence the need of punctuation recovery.

4.1. Re-ranking ASR Output

A naive idea to make use of the N-best ASR output is to produce M translation hypotheses for each ASR output, and then select the optimal translation out of these MN hypotheses. It is found that this simple method gives even worse result than translating the 1-best ASR output only. Analysis shows that, while the ASR errors in the 1-best ASR output may be fixed by some other hypothesis in the N-best list, it is indeed much more likely that the other hypotheses in the N-best list commit more ASR errors than the 1-best.

Therefore, we do not feed to the SMT decoders with all N-best ASR output. Instead we select the best candidate in the N-best list for translation. The technique of MBR re-ranking is applied again. The model is the same as in the one in Section 3.3, but the features for ASR output are:

1. the three scores coming with the N-best list;
2. language model probabilities;
3. number of words/characters.

As the re-ranking model is discriminative, it needs a training dataset in which the references (correct answers) of the ASR task are known. For this purpose we used the CRR data of our development sets as references.

4.2. End-to-End Re-ranking

Section 4.1 considers the re-ranking of ASR output as an isolated task. We also attempted an end-to-end framework in which the re-ranking of both ASR output and translation candidates of the ASR output are jointly trained in a way similar to the minimum classification error method, using a Bayesian decision function that integrates ASR scores, language model scores, and translation scores. (Please refer to [18] for details.) Unfortunately the framework failed to improve translation performance in the experiments in next section.

4.3. Punctuation Recovery

We tried two approaches to punctuation recovery, viz. tagging, and implicit recovery through translation model. For implicit recovery we drop, in the training dataset, all punctuations on the source side but still keep those on the target side. Then the translation process itself will also produce punctuations on the target side.

The tagging approach decides for each word in an input sentence whether the word is followed by a punctuation. This question can be further analyzed into two questions: 1) is the word followed by some punctuation (no matter what it is)? And 2) if yes to question 1, then exactly what is the punctuation? The first question is the prediction of punctuation position and the second the prediction of punctuation given specific position. There are therefore two versions of the tagging approach. The first version makes the two kinds of prediction separately with two taggers. The second version makes the two kinds of prediction at the same time with only one tagger. In our implementation the taggers are based on CRF modeling, and the features are about current word/POS, neighboring words/POSs, and language model probabilities.

The experiments in the next section show that the tagging approach in two stages achieves the best performance.

[1] Note that in IWSLT 2010 even the CRR datasets have all punctuations stripped.

5. Experiments

In this section the contributions of various techniques are shown by experiments. Except the fourth one, the experiments are all about Chinese-to-English DIALOG task. The test set is devset9; the development set for MERT comprises both devset8 and the Chinese DIALOG set; all other devsets are merged with BTEC and SLDB into a training set. The English side of the training set is also taken for language model training. Case-insensitive IBM-BLEU is used as evaluation metric.

Table 1 shows the effect of combining different alignments on improving the performance of individual MT decoders. Here the experiment uses CRR as translation input. The baseline is to use GIZA++ (G) only, and the test cases are to combine output from GIZA++ with that of either MSRA discriminative aligner (D) or MSRA ITG aligner (I). It can be seen that in all cases the combination of alignments does improve translation quality, and in many cases the BLEU gain is larger than 1 point.

Setting	BTG	Hiero	Dep-Hiero	Dep-BTG	Syntax
G	45.69	45.45	48.47	44.15	44.45
G+D	46.54	46.40	48.59	46.13	45.62
G+I	46.99	46.96	48.86	45.03	45.79

Table 1. Effect of alignment output combination

Table 2 shows the effect of MBR re-ranking of translation hypotheses. Here the experiment uses CRR as translation input. The first row is about the Bleu scores by the 1-best translations from various decoders while the second row is about the scores of the top one translations after MBR re-ranking. It is observed that in general MBR re-ranking improves translation quality significantly (more than 1 point).

Setting	BTG	Hiero	Dep-Hiero	Dep-BTG	Syntax
original 1-best	45.69	45.45	48.47	44.15	44.45
reranked 1-best	46.07	48.42	47.80	45.74	45.65

Table 2. Effect of MBR reranking of translation candidates

Table 3 shows the effect of MBR re-ranking of ASR output. The first row is about the Bleu scores achieved by using the original 1-best ASR output, and the second row is about the scores for the new top one candidates after MBR re-ranking of the N-best ASR output. It is observed that reranking

of ASR output slightly reduces ASR error and thus improves translation quality.

Setting	BTG	Hiero	Dep-Hiero	Dep-BTG	Syntax
original 1-best	41.05	40.43	42.80	39.16	39.28
reranked 1-best	41.98	41. 24	43.57	39.93	39.99

Table 3. Effect of MBR reranking of ASR output

Table 4 compares different methods in punctuation recovery. This experiment uses a different setting than all other experiments. It is about English-to-Chinese DIALOG task. The test set is devset11; the development set for MERT comprises both devset10 and the English DIALOG set; all other devsets are merged with BTEC and SLDB into a training set. It can be seen that the tagging approaches perform much better than implicit recovery through translation model, and that it is slightly better to separate the prediction of punctuation position and that of punctuation given specific position.

Setting	BTG	Hiero
Implicit recovery	47.87	40.98
One stage tagging	48.63	41.51
Two stage tagging	48.96	41.78

Table 4. Comparison of Punctuation Recovery Strategies

Finally, Table 5 shows the effect of MT output combination. It is obvious that MT output combination helps both the CRR and ASR tracks.

Setting	Best single decoder	MT output combination
CRR	50.02	51.43
ASR	45.87	46.53

Table 5. Effect of MT output combination

6. Conclusion

This report briefly describes the architecture and the individual modules of the MSRA SMT system for the IWSLT 2010 evaluation exercise. While our experience in evaluations like NIST'2008 shows that the MSRA SMT system achieves state-of-the-art quality in written text translation, it is our first opportunity to look into issues related to speech translation. Our experiments show that ASR errors can be lessened by MBR re-ranking of ASR output, and the recovery of punctuation should be done with the two-stage tagging approach. Moreover, it is also found that MBR re-ranking of translation hypotheses and combination of alignments are simple but useful techniques in boosting up translation performance.

In future, our focus will be parameter tuning using more than one evaluation metric. According to the unofficial results released by the time of writing, the MSRA SMT system achieves very high Bleu, which emphasizes precision and fluency. On the other hand, the MSRA system achieves not very satisfactory results with respect to recall/fidelity oriented metrics like Meteor and unigram recall. That is a natural consequence from our use of Bleu as the only criterion in training. Therefore we need to work on the pursuit of a suitable balance between recall and precision in translation.

7. References

[1] Koehn, P. et. al. "Moses: Open Source Toolkit for Statistical Machine Translation", 45th ACL demo session, 2007.

[2] Xiong, D., Liu, Q., and Lin, S. "Maximum Entropy based Phrase Reordering Model for Statistical Machine Translation" Proc. 44th ACL, 2006.

[3] Chiang, D. "Hierarchical Phrase based Translation" Computational Linguistics, 33(2), 2007, pp. 201-228.

[4] Shen, L., Xu, J., and Weischedel, R. "A New String-to-Dependency Machine Translation Algorithm with a Target Dependency Language Model" Proc. 46th ACL, 2008.

[5] Galley, M. et. al. "Scalable Inference and Training of Context-Rich Syntactic Translation Models" Proc. 44th ACL, 2006.

[6] Marcu, D., Wang, W., Echihabi, A., and Knight, K. "SPMT: Statistical Machine Translation with Syntactified Target Language Phrases" Proc. EMNLP, 2006.

[7] Wang, W., Knight, K., and Marcu, D. "Binarizing Syntax Trees to Improve Syntax-based Machine Translation Accuracy" Proc. EMNLP, 2007.

[8] Quirk, C., Menezes, A., and Cherry, C. "Dependency Treelet Translation: Syntactically Informed Phrasal SMT" Proc. 43th ACL, 2005.

[9] Li, C.-H., He, X., Liu, Y., and Xi, N. "Incremental HMM Alignment for MT System Combination", Proc. 47th ACL, 2009.

[10] Rosti, A.-V., Matsoukas, S., and Schwartz, R. "Improved Word-level System Combination for Machine Translation", Proc. 45th ACL, 2007.

[11] Lita, L.V., Ittycheriah, A., Roukos, S., and Kambhatla, N. "tRuEcasIng", Proc. 41st ACL, 2003.

[12] Och, F. J. and Ney, H. "Improved Statistical Alignment Models" Proc. 39th ACL, 2000.

[13] Liu, S., Li, C.-H., and Zhou, M. "Discriminative Pruning for Discriminative ITG Alignment" Proc. 48th ACL, 2010.

[14] Moore, R. "A Discriminative Framework for Bilingual Word Alignment" Proc. EMNLP, 2005.

[15] Och, F. J. "Minimum Error Rate Training in Statistical Machine Translation" Proc. 41st ACL, 2003.

[16] Kumar, S., Macherey, W., Dyer, C. and Och, F. J. "Efficient Minimum Error Rate Training and Minimum Bayes-Risk Decoding for Translation Hypergraphs and Lattices" Proc. 47th ACL, 2009.

[17] Zens, R. and Ney, H. "N-gram Posterior Probabilities for Statistical Machine Translation" Proc. HLT-NAACL, 2006.

[18] Zhang, Y., Deng, L., He, X., and Acero, A. "Integrative Scoring and End-to-End Discriminative Training for Spoken Utterance Translation" submitted to IEEE Trans. Audio, Speech and Language Processing.