The ICT,CAS MT Systems for the IWSLT09 Evaluation

Hailao Mi, Yang Liu, Tian Xia, Yang Feng, Xinyan Xiao, Jun Xie, Zhaopeng Tu, Hao Xiong, Daqi Zheng, Yajuan Lü and Qun Liu
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
{htmi,yl,xiatian,fengyang,xiaoxinyan,lvajuan,liuqun}@ict.ac.cn

1. Overview
ICT,CAS participated in three tasks:
1. BTEC task, Chinese-English direction;
2. Challenge task, Chinese-English direction;

For each task, we finally submitted a single system which achieved a maximum BLEU score on development set among four different systems.

2. Single Systems
2.1 Silenus
Silenus (Mi et al., 2008; Mi and Huang, 2008) is a forest-based tree-to-string SMT system. A packed parse forest is a compact representation of all derivations (i.e., parse trees) for a given sentence under a context-free grammar. A tree-to-string rule describes the correspondence between a source parse tree and a target string.

The first two rules are used to merge two neighboring blocks into a larger block either in a monotonic or an inverted order. A block is a pair of source and target contiguous sequences of words. The last rule translates a source phrase into a target phrase and generate a block. Figure 2 gives some blocks. The first block and the second block is connected in a monotonic order. The third and the fourth block is connected in an inverted order.

The ordering problem is a typical two-class classification. So we build a maximum entropy model to predict the merging order of two phrases.

2.3 Chiero
Chiero is a re-implementation of the state-of-the-art hierarchical string-to-string translation system (Chiang, 2007). The model can formalized as a synchronous context-free grammar.

2.4 Moses
Moses is a phrase-based model. It is an open source system and uses beam-search to reduce the searching space. We use the default settings for this model.

3. Data Preparation
We only provide the data collected by the organizer for each task. We first used the Chinese lexical analysis system ICTCLAS for splitting Chinese characters into words and a rule-based tokenizer for tokenizing English words. Then, we convert all alphabetic characters to a 2-byte representation. Finally, we use GIZA++ and used the “grow-diag-final” heuristic to get many-to-many word alignments.

We use the Chiero baseline system as the reference point. The model is a synchronous context-free grammar.

4 Development Set Selection
Our development set for each task is selected automatically from all the development sentences according to the n-gram similarity, which is calculated against the current test set sentence.

5. Additional Experiments
We also conducted several experiments of system combination after the Evaluation Campaign. Finally, we applied two kinds of word level combination systems, which are based on the techniques of BM25 [Mi et al., 2008] and TIE [Dorr et al., 2006] respectively. But all systems are failed due to the poor hypothesis alignments, on which the Oracle BLEU score is only 6 points higher than the score of single best system. Finally, we remarked the merged n-best lists of all single systems with one sentence level combination system, which is global linear model with a series of sentence level combination system, and obtain significant improvements of +4 BLEU on BTEC, CT2.4. What we can conclude is that sentence-level combination method is more suitable than word-level approach on spoken language translation.

6. Conclusion
We also used the ICT,SMT systems for the evaluation campaign of IWSLT 2008. For each task, we first used the selection method to construct a development set, on which we tuned all the single systems with MERT. Finally, we chose the system with maximum BLEU score as our primary system. Since we didn’t use any reranking or system combination techniques for the final submission, we got relatively lower rank. Another problem we doubt is the small training set, which includes only 30k sentence pairs. The small training set will inevitably introduce much errors to SMT pipeline, such as word segmentation, parsing, word alignment. As a result, on one hand, good translation models are failed to explore their potential strengths; on the other hand, the pre- and post-processing techniques will attract more and more attentions, since they can reduce the negative effects of noise significantly. As last, the additional experiments, we carried out after the Evaluation Campaign, also suggested that our sentence-level combination system performs better than the word-level combinations of tasks of spoken language translation.