The ICT,CAS MT Systems for the IWSLT09 Evaluation

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

1. Overview
ICT,CAS participated in three tasks:
1. BTSC task, Chinese-English direction;
2. Challenge task, English-Chinese direction;
For each task, we finally submitted a single system who achieved a maximum BLEU score on development set among four different single 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.
2.2 Moses
Moses (Xiong et al., 2008) 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 provided the data provider 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 sentences. Then, we converted all alphabetic characters to 2-byte representation. Finally, we extract GIZA++-used the “grow-diagonal” heuristic to get many-to-many word alignments.

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 sentences. Our method works as follows: First, we gather every n-gram (up to 10) in the test into a map W, and assign a score S_w for each n-gram w in W, which is calculated as:

\[ S_w = \sum_{i=1}^{n} \sum_{j=1}^{m} \max \{ \theta_{i,j} \} \]

where \( \theta_{i,j} \) is the number of occurrence of \( w_i \) in \( w_j \) test set. Then, we assign a sentence score \( S_s \) for each sentence in development set, which is calculated as:

\[ S_s = \sum_{w \in S} S_w \]

Finally, we reranked the merged best lists of all single systems with our sentence-level combination system, which is global linear model with a series of word alignment.

5. Additional Experiments
We also conducted several experiments of system combination after the Evaluation Campaign. Firstly, we applied two kinds of word level combination systems, which are based on the techniques of HMKe, (Mi et al., 2008) and TRENoiser, (Mi et al., 2008) respectively. But both systems are failed due to the poor hypothesis alignment, on which the Oracle BLEU score is only 6 points higher than the score of single best system. Finally, we reranked the merged best lists of all single systems with our sentence level combination system, which is global linear model with a series of simple heuristics, obtained significant improvements of +4 BLEU on BTSC, CT task.

6. Conclusion
In this paper, we describe the ICT,CAS MT systems for the evaluation campaign of IWSLT09. 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 rescoring or system combination techniques for the final submission, we got a 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 SM 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 these 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 combination on spoken language translation.

Table 1: BLEU scores of our single systems on Dev sets

<table>
<thead>
<tr>
<th>Task</th>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTSC</td>
<td>CR</td>
<td>0.5063</td>
</tr>
<tr>
<td>CT_CE</td>
<td>CR</td>
<td>0.3078</td>
</tr>
<tr>
<td>CT_EC</td>
<td>ASR 20</td>
<td>0.2859</td>
</tr>
<tr>
<td>Silenus</td>
<td>ASR 20</td>
<td>0.2901</td>
</tr>
</tbody>
</table>

Table 2: The BLEU scores of each primary single system on test sets