The ICT System Description for IWSLT 2008

Yang Liu, Zhongjun He, Haitao Mi, Yun Huang, Yang Feng, Wenbin Jiang, Yajuan Lu, and Qun Liu

Key Laboratory of Intelligent Information Processing

Institute of Computing Technology, Chinese Academy of Sciences

{yliu, zjhe, htmi, huangyun, fengyang, jiangwenbin, lvyajuan, liuqun}@ict.ac.cn

1. Systems

The ICT system for IWSLT 2008 is a combination of four systems.

1.1 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.

Unlike previous tree-to-string (Liu et al., 2006; Huang et al., 2006) or string-to-tree models (Galley et al., 2006; Marcu et al., 2006), we extract tree-to-string rules from aligned forest-string pairs. In decoding, the input is a source forest rather than a source tree.

Figure 1 demonstrates a pair of linked source forest and target string. Table 1 shows some tree-to-string rules learned from the example forest-string pair in Figure 1.
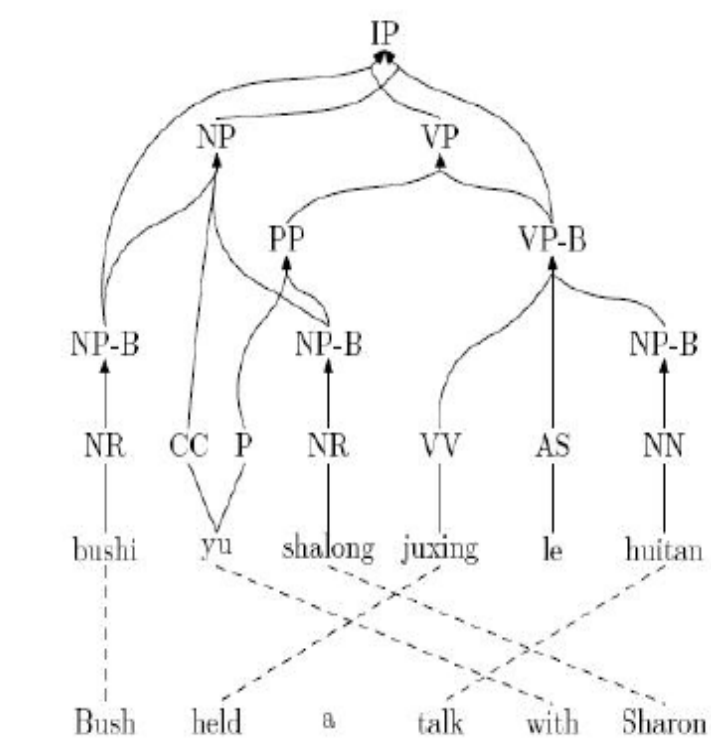

Figure 1: *A pair of linked source forest and target string. The solid lines denote hyperedges and the dashed lines denote word alignments.*

Table 1: *A minimal derivation of the example shown in Figure 1.*

(1)	$IP(x_1:NP\text{-}B, x_2:VP) \to x_1\ x_2$
(2)	$NP\text{-}B(x_1:NR) \to x_1$
(3)	$NR(bushi) \to Bush$
(4)	$VP(x_1:PP, x_2:VP\text{-}B) \to x_1\ x_2$
(5)	$PP(x_1:P, x_2:NP\text{-}B) \to x_1\ x_2$
(6)	$P(yu) \to with$
(7)	$NP\text{-}B(x_1:NR) \to x_1$
(8)	$NR(shalong) \to Sharon$
(9)	$VP\text{-}B(x_1:VV, AS(le), x_2:NP\text{-}B) \to x_1\ a\ x_2$
(10)	$VV(juxing) \to held$
(11)	$NN(huitan) \to talk$

1.2 Bruin

Bruin (Xiong et al., 2006) is a formally syntax-based system that implements a maximum entropy based reordering model on BTG rules (Wu 1997). Bruin employs the following three BTG rules to direct translation:

$$A \xrightarrow{[\ ]} (A^1, A^2)$$
$$A \xrightarrow{\langle\rangle} (A^1, A^2)$$
$$A \to (x, y)$$

The first two rules are used to merge two neighboring blocks into one 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.

This is a typical two-class classification. We build a maximum entropy model to predict the merging order so as to reorder phrases.
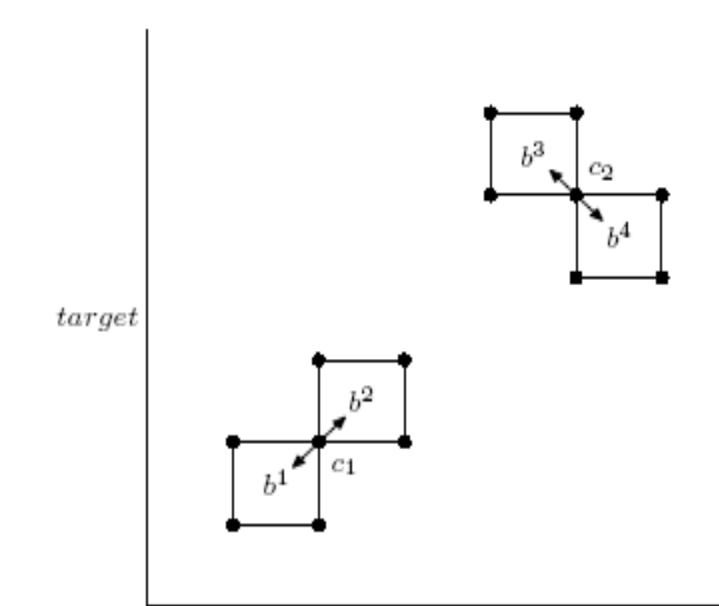

Figure 2: *Blocks and orders. The bold dots are corners and the arrows from the corners are their links.*

1.3 Mencius

Mencius (He et al., 2008) is a phrase-based system that is very similar to Moses. The major difference is that we introduce similarity-based partial matching for bilingual phrases to alleviate data sparseness problem.

If two source phrases share with the same length and the same parts-of-speech sequence, we refer to them as similar phrase pair. Their similarity is computed as the phrase length divided by the number of shared words. Our hope is that similar bilingual phrases can be used to create translation templates if one source phrase cannot find translations in the phrase table.

For example, suppose that we cannot find translations for a source phrase "*yu zuotian dida taiguo*" in a phrase table, in we find a similar phrase "*yu zuowan dida bulage*" with its translation "*arrived in Prague last evening*". According to the alignment information, we obtain a translation template "<*yu X₁ dida X₂, arrived in X₂ X₁*>". Then, the unmatched source substrings "*zuotian*" and "*taiguo*" can be translated into "*yesterday*" and "*Thailand*", respectively. As a result, the translation for "*yu zuotain dida taiguo*" is "*arrived in Thailand yesterday*".

1.4 Change

Change is an implementation of the state-of-the-art hierarchical phrase-based model (Chiang, 2007). Considered as an extension of standard phrase-based model, hierarchical phrase-based model allows non-contiguous parts of source sentence to be translated into possibly non-contiguous parts of target sentence. The model can formalized as a synchronous context-free grammar.

Our implementation faithfully follows Chiang's work. The only exception is the condition for terminating cube pruning. Chiang's implementation quits upon considering the next item if its score falls outside the beam by more than a margin. We find that large number of items will often be enumerated under this condition in our experiments. To tackle this problem, we further limit the number of items taken from the heap.

1.5 System Combination

We combine the outputs of single SMT systems at sentence level, similarly to the work by Macherey and Och (2007). Global linear models are used as a framework for reranking a merged *n*-best list:

$$\hat{y} = \operatorname*{argmax}_{y \in GEN(x)} \mathbf{f}(x, y) \cdot \mathbf{W}$$

Three types of features are used: (1) relative BLEU scores against 1-best translations from other candidates, (2) language models scores, and (3) length of translation. The feature weights are tuned using minimum-error-rate training (Och, 2003). In this year's evaluation, each single SMT system generated 200-best list translations, which were merged and served as the input to the combiner.

2. Data

Besides the data provided by the organizer, we used the following additional data:

(1) Chinese LDC (CLDC-LAC-2003-004)

(2) Chinese LDC (CLDC-LAC-2003-006)

(3) Chinese LDC (2004-863-008)

(4) Chinese LDC (2004-863-009)

(5) LDC2002L27 "Chinese-English Translation Lexicon Version 3.0"

(6) LDC2005T34 "Chinese-English Named Entity Lists Version 1.0"

(7) Tanaka's Corpus

The training corpus contains about 8.1M Chinese words and 8.6M English words. We used SRILM to train 5-gram language models.

3. Annotation

We used the Chinese lexical analysis system ICTCLAS for splitting Chinese characters into words and the tokenizer provided by IWSLT for tokenizing English sentences. After that, we convert all alphanumeric characters to their 2-byte representation. Then, we ran GIZA++ and used the "grow-diag-final" heuristic to get many-to-many word alignments. We observe that in a sentence some phrases are more likely to appear at the beginning, while other phrases are more likely to be located at the end. Inspired by the literature in language modeling, we mark the beginning and ending of word aligned sentences with two tags, "<s>" and "</s>", to capture such reordering information. The sentences to be translated will also be annotated with the two tags, which will be removed after decoding. To get packed forests for Silenus, we used the Chinese parser (Xiong et al., 2005) modified by Haitao Mi and the English parser (Charniak and Johnson, 2005) modified by Liang Huang to produce entire parse forests. Then, we ran the Python scripts (Huang, 2008) provided by Liang Huang to output packed forests. To prune the packed forests, Huang uses inside and outside probabilities to compute the distance of the best derivation that traverses a hyperedge away from the globally best derivation. A hyperedge will be pruned if the difference is greater than a threshold. Nodes with all incoming hyperedges pruned are also pruned.

Table 2: *BLEU scores of five systems on IWSLT 2007 Chinese-English development set.*

System	provided	provided + additional
Silenus	0.3413	0.3919
Bruin	0.3572	0.4213
Mencius	0.3692	0.4341
Change	0.3792	0.4287
Combination	0.3894	0.4483

Table 3: *BLEU scores of five systems on test sets.*

System	BTEC_CE		CT_CE		CT_EC	
	CRR	ASR.1	CRR	ASR.1	CRR	ASR.1
Silenus	0.4639	0.4061	0.4048	0.3541	0.4988	0.4119
Bruin	0.4796	0.4097	0.4114	0.3280	0.4746	0.3950
Mencius	0.4689	0.3960	0.3850	0.3050	0.4700	0.3864
Change	0.4712	0.3946	0.4066	0.3308	0.4815	0.3961
Combination	0.4943	0.4403	0.4370	0.3703	0.5045	0.4249

4. Results

Table 2 presents the BLEU scores (case-sensitive, with punctuations) of our five systems achieved on the IWSLT 2007 Chinese-English development set. Prior to the evaluation, we used the development sets from 2003 to 2006 as development sets to tune model scaling factors and used 2007 development set as test set. "provided" denotes the training data provided by the organizer that consist of about 30K pairs of sentences. "provided+additional" denotes all the training data we have, as listed at the beginning of Section 3.1. We observe that using more data results in substantial improvements of about 5 BLEU points.

Table 3 gives the BLEU scores (case-sensitive, with punctuations) of our five systems achieved on the test sets "BTEC CE" denotes Chinese-English direction of BTEC task, "CT CE" denotes Chinese-English direction of challenge task, and "CT EC" denotes English-Chinese direction of challenge task. "CRR" denotes correct recognition results and "ASR.1" denotes using 1-best ASR output.

Our sentence-level system combiner outperformed single systems consistently on all tasks. While system combination benefited Chinese-English direction significantly, the improvements on English-Chinese direction were relatively small. One possible reason might be that fewer development sets are available for English-Chinese direction for system combiner to optimize the parameters automatically.

For single SMT systems, Bruin got better results than the others on Chinese-English direction. Interestingly, Silenus surpassed other systems significantly on English-Chinese direction. There are two findings worth noting:

(1) Silenus uses packed forests instead of 1-best parses, minimizing the negative effect of parsing errors. As the amount and domain of data used for training parsers are comparatively limited, parsers will inevitably output ill-formed trees when handle real-world text. Guided by such noisy syntactic information, syntax-based models that rely on only 1-best parses are prone to produce degenerate translations. The results suggest that packed forests do help syntax-based systems to achieve comparable performance with phrase-based systems on tourism-related sentences.

(2) Parsing accuracy has a substantial effect on syntax-based models. Silenus obtained better results on English-Chinese direction than Chinese-English direction. We believe the major reason is that parsing on English is more accurate than Chinese.

5. Conclusion

In this paper, we give a brief introduction to our four single SMT systems and one system combiner. We report the resources used, annotation techniques, and results achieved on the test sets. We find that our implementation of sentence-level system combination works for all tasks. Another interesting finding is that syntax-based models could produce translations as good as phrase-based systems on tourism-related text if packed forests are used.