The HIT-LTRC Machine Translation System for IWSLT 2012

Xiaoning Zhu, Yiming Cui, Conghui Zhu, Tiejun Zhao, Hailong Cao

Language Technology Research Center
Harbin Institute of Technology, China
{xnzhu,ymcui,chzhu,tjzhao,hailong}@mtlab.hit.edu.cn

Abstract

In this paper, we describe HIT-LTRC's participation in the IWSLT 2012 evaluation campaign. In this year, we took part in the Olympics Task which required the participants to translate Chinese to English with limited data.

Our system is based on Moses[1], which is an open source machine translation system. We mainly used the phrase-based models to carry out our experiments, and factored-based models were also performed in comparison. All the involved tools are freely available.

In the evaluation campaign, we focus on data selection, phrase extraction method comparison and phrase table combination.

1. Introduction

This paper describes the Statistical Machine Translation (SMT) system explored by the Language Technology Research Center of Harbin Institute of Technology (HIT-LTRC) for IWSLT 2012. Generally, our system was based on Moses, and phrase-based models were used.

In Olympics shared task, the training data was limited to the supplied data including HIT Olympic Bilingual Corpus (HIT)[2] and Basic Travel Expression Corpus (BTEC)[3]. Although the two corpora are both oral corpus, there are still some differences between them. For example, the BTEC corpus is travel-related, and the HIT corpus is mainly about the Olympic Games. Besides this, the organizer of IWSLT 2012 also provided two development sets which are selected from the HIT and BTEC corpus respectively. Because the training data is limited by the above corpus, in order to get a better performance, we need to excavate all the potential of the two corpora, including the development sets.

One key problem of the SMT system is how to extract the phrase. Giza++[4] is a popular word alignment tool which can produce word alignment information with parallel corpus. By using heuristic phrase extraction method, we can extract phrases with the alignment. Compared with heuristic phrase extraction method, Pialign[5] is an unsupervised model for joint phrase alignment and extraction using nonparametric Bayesian methods and inversion transduction grammars (ITGs). We compared the phrase table extracted by the two phrase extraction methods in many ways, such as the size, the quality, and the differences of two methods.

System combination has been approved to improve machine translation performance significantly. With several machine translation systems' outputs, researchers can get a better translation by combining the outputs. But in this paper, we didn't combine the outputs; instead we combine the models generated by Giza++ and Pialign. It is shown that we can get a better performance by model combination.

The following of the paper is organized as follows. Section 2 describes a phrase-based machine translation system which was used in our work. In section 3, we compared differences of two corpora. The result and phrase extraction are discussed in section 4. And in the last section, we give a conclusion and discuss the future work.

2. Phrase-based System

Our primary system is based on Moses with a phrase-based model. Under the log-linear framework[6], when given a source sentence f , we can get a translation e as follows:

$$p(e \mid f;\lambda) = \frac{\exp(\lambda \bullet h(f,e))}{Z(\lambda)}$$

with

$$Z(\lambda) = \sum \exp(\lambda \bullet h(f,e))$$

where $h(f,e)$ denotes the feature vector of the pair (f,e), and λ is its corresponding weight vector. $h(f,e)$ contains 14 features and they are divided into following categories:

- Bidirectional translation probabilities;
- Bidirectional lexical translation probabilities;
- MSD-reordering model;
- Distortion model;
- Language model;
- Word penalty;
- Phrase penalty.

2.1. Pre-processing

The Chinese sentences supplied by the organizer were not segmented, so we used the Stanford Word Segmenter[7] to segment the Chinese sentences with the PKU model. The English sentences were not tokenized, thus we used the open source tools supplied by Moses to tokenize them. We also lowercased all the English data for training. There are many English punctuation characters in Chinese sentences (and vice versa), so we wrote some scripts to change all the punctuation characters in order.

2.2. Training

In the training step, we used Giza++ to get alignments and combined the alignments with *grow-diag-final-and* method. With the alignments, we can extract phrases with heuristic phrase extraction method and generate the translation model. Besides, we also used Pialign to generate the alignments and phrases.

The language model was built with SRILM toolkit[8]. A 5-gram language model was used for decoding. The corpus we used to build the language model is all the supplied data, including training data and development data.

2.3. Decoder

The decoder used in our system is Moses.

2.4. Tuning

The parameters were tuned on the development set with standard trainer MERT[9]. When running MERT, the k-best-list-size was set as 100 and BLEU4[10] was selected as the evaluation metric.

2.5. Post-processing

The translations were post-processed after decoding.

- All the Chinese words in output were deleted. Because there are many names in the test set, and most of them can't be translated, so we deleted them;
- The English sentences were de-tokenized ;
- The English sentences were re-cased by the recaser tools provided by Moses.

3. Corpus

The IWSLT organizer provided two training corpus, including HIT corpus and BTEC corpus. HIT corpus is a multilingual oral corpus developed for the Beijing 2008 Olympic Games. There are five domains in HIT corpus, including traveling, dining, sports, traffic and business. The BTEC corpus is also an oral corpus containing tourism-related sentences. Besides the training corpus, they also provided two development corpus, which were extracted from the HIT corpus and BTEC corpus. In the following paper, we use HIT_train, HIT_dev, BTEC_train, BTEC_dev to denote four corpora respectively.

In our system, we used HIT_train, BTEC_train, BTEC_dev, HIT_dev as our training data. And HIT_dev was also used as our development set. We also random sampled 1000 sentences from HIT corpus as our test set.

The detail of the corpus is presented in Table 1.

Table 1: Corpus

	BTEC	HIT
Train	19975	52603
Dev	2977	2057
Total	22949	54660

We combined the four corpora as training data, and the new generated corpus is shown in Table 2.

Table 2: Training data

name	corpus	#
Corpus 1	BTEC_train+HIT_train	72575
Corpus 2	Corpus1+BTEC_dev	75552
Corpus 3	Corpus2+HIT_dev	77609

4. Experiments and Results

4.1. The comparison of Giza++ and Pialign

We first trained six models with Giza++ alignments and Pialign alignments. A comparison between the phrase table generated from Giza++ and Pialign is shown in Table 3. Table 4 shows the covering of the six phrase tables of the test set.

Table 3: Comparison between Giza++ and Pialign

Corpus	align	total	common	different
1	Giza++	1182913	409443	773470
	Pialign	1385520		976077
2	Giza++	1208128	418788	789340
	Pialign	1413367		994579
3	Giza++	1236688	428377	808306
	Pialign	1445577		1017200

Table 4: Covering of test set

Corpus	align	Chinese	English
1	Giza++	21.7%	36.0%
	Pialign	23.6%	38.3%
2	Giza++	21.7%	36.1%
	Pialign	23.8%	38.7%
3	Giza++	21.9%	36.6%
	Pialign	23.9%	38.9%

In Table 3, we showed the total number of phrase pairs, the common phrase pairs of Giza++ and Pialign, the different phrase pairs of Giza++ and Pialign. In Table 4, we show the covering capacity of the phrase table. The covering capacity c is defined as follows:

$$c = \frac{\#\text{ of phrases both in test set and in phrase table}}{\#\text{ of phrases in test set}}$$

To note that, the test set was divided into unigram to 5-gram phrases.

From Table 3 we can find that the phrase table generated by Pialign is a little bigger than Giza++. Because we use *-samps* parameters to sample the bilingual parser tree repeatedly. In this experiment, we tuned this parameters from 1(default) to 80. At first, with the increment of the phrase table size, the performance grows at the same time. But after 20th sampling, the bias of sampling adds too many noise phrase pairs. Finally, we set this value to 20. With default value, Pialign only generated 389,982 phrase pairs (32.28% as the Giza++ did), but the performances are still comparable. With the covering capacity, we can estimate the performance of the model. The result is the same with the translation result, which shows that Pialign is better than Giza++ in phrase extraction.

4.2. Results of translation

The result of translation outputs are shown in Table 5 and Table 6.

The result is confusing. After we tuned the parameters with HIT_dev, the result became worse. This may be caused by the mismatch between HIT_dev and HIT_train. The result also shows that although we continue to enlarge the size of

training data, the BLEU score may reduce on the contrary. These remind us that the model is also important.

Table 5: Result without tuning

Corpus	align	BLEU%
1	Giza++	20.76
	Pialign	20.80
2	Giza++	20.62
	Pialign	21.20
3	Giza++	20.51
	Pialign	20.54

Table 6: Result with tuning

Corpus	align	BLEU%
1	Giza++	19.97
	Pialign	19.70
2	Giza++	18.40
	Pialign	19.66
3	Giza++	15.52
	Pialign	15.10

4.3. Combination of two phrase table

We explored Giza++ and Pialign to extract phrases. In this section, we want to combine the two methods by merging two phrase tables using a linear interpolation method. For Giza++, the best result was achieved when we used Corpus1. For Pialign, the best result was achieved when we used Corpus2. So we combined the two phrase tables. The result without tuning is shown in Table 7. The parameter means the weight of Pialign.

Table 7: Phrase Table Combination

parameter	BLEU%
0.4	20.69
0.5	20.78
0.6	20.62

Compared with Table 7 and Table 5, we can draw a conclusion that phrase table combination can improve the performance of machine translation systems a little. Maybe due to the size of the training data, the result is not very clear to see the increment. And our combination method is only a linear interpolation method, which is naive for phrase table combination. We believe that a more complex strategy, such as some machine learning algorithms can improve the phrase table combination results.

4.4. Linguistic knowledge

In recently years, many researchers have focused on how to integrate linguistic knowledge into machine translation systems. In this work, part of speech was introduced to improve the machine translation systems. We used Stanford Log-linear Part-Of-Speech Tagger[11] to get the POS tag. Factored-based model of Moses was used to train a translation model. The result is shown in Table 8.

Table 8: Linguistic features

system	With tuning	Without tuning
baseline	19.97	20.76
+pos tag	18.53	16.63

As we can see that the result with POS tag is also not better than the baseline. We think that linguistic knowledge is a good research field to improve machine translation performance.

4.5. Official Results

We took part in the Olympics task(OLY)[12], and the final translations we submitted was generated by Pialign with corpus 2. And because of the bad performance of tuning, we submit out results without tuning. The final result was shown in Table 9.

Table 9: Official results in BLEU

system	case+punc	no_case+no_punc
Pialign-2	19.10	18.76

5. Conclusions and Future Work

In this paper, we explained our work in the IWSLT 2012 evaluation campaign. We compared two phrase extraction methods and tried to combine the two methods. The results show that the combination method can improve the result of MT systems.

In future, we will still try to study some other advanced combination methods to modify our system.

6. Acknowledgements

The work of this paper is funded by the project of National Natural Science Foundation of China (No. 61100093) and the project of National High Technology Research and Development Program of China (863 Program) (No. 2011AA01A207).

7. References

[1] P. Koehn et al., “Moses: Open Source Toolkit for Statistical Machine Translation”, in Proceedings of the ACL Demo and Poster Sessions, Prague, Czech Republic, 2007, pp.177–180.

[2] Muyun Yang, Hongfei Jiang, Tiejun Zhao and Sheng Li, “Construct Trilingual Parallel Corpus on Demand”, *Chinese Spoken Language Processing*, vol. 4274, pp. 760-767, 2006

[3] Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. “Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World”, in Proceedings of LREC 2002, Las Palmas, Spain, 2002

[4] Franz Josef Och, Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models", *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.

[5] G. Neubig, T. Watanabe, E. Sumita, S. Mori, and T. Kawahara, “An unsupervised model for joint phrase

alignment and extraction," in Proceedings of ACL, 2011, pp. 632–641.

[6] Franz Josef Och and Hermann Ney. 2002. "Discriminative training and maximum entropy models for statistical machine translation", in Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pages 295–302, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

[7] Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In Fourth SIGHAN Workshop on Chinese Language Processing.

[8] Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In Proceedings of the International Conference on Spoken Language Processing, volume 2, pages 901–904.

[9] Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pages160–167, Morristown, NJ, USA. Association for Computational Linguistics.

[10] Kishore Papineni, Salim Roukos, Todd Ward, andWei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

[11] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of HLT-NAACL, pages 252–259.

[12] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, S. Stüker. Overview of the IWSLT 2012 Evaluation Campaign, In Proc. of IWSLT, Hong Kong, HK, 2012.