The CASIA Machine Translation System for IWSLT 2013

Xingyuan Peng, Xiaoyin Fu, Wei Wei, Zhenbiao Chen, Wei Chen, Bo Xu

Interactive Digital Media Technology Research Center, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100190, China
{xingyuan.peng, xiaoyin.fu, wei.wei.media, zhenbiao.chen, xubo}@ia.ac.cn

Abstract

In this paper, we describe the CASIA statistical machine translation (SMT) system for the IWSLT2013 Evaluation Campaign. We participated in the Chinese-English and English-Chinese translation tasks. For both of these tasks, we used a hierarchical phrase-based (HPB) decoder and made it as our baseline translation system. A number of techniques were proposed to deal with these translation tasks, including parallel sentence extraction, pre-processing, translation model (TM) optimization, language model (LM) interpolation, turning, and post-processing. With these techniques, the translation results were significantly improved compared with that of the baseline system.

1. Introduction

This paper describes the machine translation (MT) system developed by the Institute of Automation Chinese Academy of Sciences (CASIA) for the evaluation campaign of I-WSLT 2013. We participated in the optional MT track with the Chinese-English and English-Chinese translation tasks. Our translation system is based on the hierarchical phrase-based translation model [1]. We used the state-of-the-art H-PB translation system as our baseline system.

Efforts have been made to improve the translation performance. To obtain high quality parallel sentences, we introduced a parallel sentence extraction method based on the lexical translation probabilities. Rule-based translation of named entities was proposed to deal with translations of time and numbers that are done incorrectly. For our translation model training, the forced alignment technique [2] was used for optimizing the translation rules and reducing the hierarchical phrase table size. In addition, we used the provided monolingual corpora to train different language models and interpolated the language model to adapt to the translation tasks. At last, we added word-to-word phrases to the hierarchical phrase table to reduce the number of untranslated words.

The remainder of the paper is structured as follows. Section 2 describes the resources used in our system. Section 3 gives an overview of the whole system. Section 4 discusses the improvement of translation performance in detail. Section 5 presents the experiments and evaluation results. Finally, section 6 concludes the paper.

2. Resources Used in IWSLT 2013

Training of the translation models and language models for MT systems is constrained to data supplied by the organizers. Since we only participated in the Chinese-English and English-Chinese translation tasks, we made full use of the parallel and monolingual training corpora in Chinese and English. The training corpora are divided into two parts: the parallel data for the TM training and the monolingual data for the LM training.

Parallel Data: There are two types of parallel data for our translation model training.

- WIT^3 (Web Inventory of Transcribed and Translated Talk), which redistributes the original content published by the TED Conference website [3]. After the pre-processing, there are 151,149 aligned English-Chinese parallel sentences in these data. We will use WIT^3 to represent the extracted parallel sentences from the WIT^3 training corpus.
- MultiUN (Multilingual UN Parallel Text 2000-2009), which provides parallel corpus extracted from the U-nited Nations website [4]. As the data has alignment problems, we realigned the parallel sentences (Details will be introduced in section 4.1). Finally, 7,819,291 parallel sentences were extracted in these data. We will use UN to represent the parallel sentences extracted from the MultiUN training corpus.

Monolingual Data (English): There are five English datasets used for the LM training in our experiment.

- News Commentary v7 from WMT 2012
- News Crawl from WMT 2012
- Europarl v7
- LDC2011T07 English Gigaword Fifth Edition
- Monolingual UN corpus (the English part of MultiUN)

Monolingual Data (Chinese): There are three Chinese datasets used for the LM training in our experiment.

- WIT^3 (the Chinese part of the parallel data mentioned above).

- Monolingual UN corpus (the Chinese part of MultiUN)

- Google book grams

3. System Overview

3.1. Chinese Word Segmentation System

The Chinese word segmentation (CWS) system is based on our in-house toolkit [5], which combines both CRF-based model and N-gram language model to segment Chinese words. CRF model treats the CWS task as a sequence tagging question. It overcomes the tagging bias problem in generative models. However, it tends to generate longer words, which is harmful to SMT system because it causes data sparseness. To overcome this drawback, we introduced N-gram language model as a supplement to CRF-based model. The N-gram language model generates significantly shorter words than the CRF-based model does, which can be helpful to distinguishing shorter words. Compared to the open source CWS toolkit ICTCLAS[1], the CRF++[2] training toolkit was used to train our CRF based model and the SRILM toolkit was used to train the N-gram language model with the annotated Chinese People's Daily News corpus as resources from February to June, 1998. We tested the performance on the news corpus in January, 1998. The results measured by precision (**P**), recall (**R**) and **F1** measure are listed in Table 1.

Table 1: *The CWS results on the Chinese People's News corpus.*

System	**P**	**R**	**F1**
ICTCLAS	98.1%	98.7%	98.4%
CASIA	97.5%	97.7%	97.6%

3.2. Hierarchical Phrase-based System

For our HPB translation system, we employed an in-house implementation of the state-of-the-art MT decoder, which is mainly based on the work of [7]. In HPB translation system, a weighted synchronous context-free grammar is induced. There are two types of phrases distinguished by the non-terminals in HPB rules. Phrases without non-terminals are the initial phrases and those with up to two non-terminals are the hierarchical ones. Both of them were heuristically extracted from the aligned parallel sentences. The search was carried out on a CKY parser with beam search together with a post-processor for mapping source language derivations to target ones. The standard features integrated into our decoder include: phrase translation probabilities and lexical probabilities in both translation directions, word and phrase penalty, glue rules, and N-gram language model, all of which are assigned by the log-linear model [8]. Besides, we used the cube pruning [9] to speed up our decoder, and the standard MERT [10] to tune the weights of our features on the 100-best translation assumptions on IWSLT 2010 development set.

3.3. Forced Alignment System

Usually, the original HPB phrases can be extracted heuristically from the aligned words of parallel sentences, as proposed in [7]. However, the heuristical phrases extraction suffers from a large amount of redundant rules and meets difficulties in probability estimation. To avoid these, we employed the idea of force-aligning training data with the heuristically trained HPB rules [2]. Instead of directly applying these HPB rules in decoding, we used the original HPB rule to align the parallel training sentences and generated the bilingual derivation trees that represent both the source and target sentences. Then, HPB rules were extracted from the derivation trees with a threshold pruning. The translation probabilities of HPB rules were updated.

It should be noted that we only re-estimated the phrasal translation probabilities, and kept the lexical translation probabilities estimated with the method of [11]. After generating the optimized HPB rules, we tested our forced alignment (FA) method on the IWSLT 2012 and 2013 Chinese-English MT test set with the translation models trained from the WIT^3 parallel corpus. The phrase table sizes and translation results are listed in Table 2.

Table 2: *Forced Alignment results on the IWSLT 2012 and 2013 Chinese-English translation tasks. The translation performances are measured by BLEU and TER.*

System	tst2012		tst2013		#Phrases
	BLEU	TER	BLEU	TER	
WIT^3	12.5	67.0	14.3	68.4	22.7M
WIT^3+FA	12.6	66.7	14.4	67.7	11.4M

The result showed that the total HPB phrase table was reduced by 50% and the performances in both translation tasks are slightly better compared to that of the baseline HPB translation system. Besides, a large number of phrases have been dropped out by our forced alignment, speeding up the HPB decoder.

4. Improvements

4.1. Parallel Sentence Extraction

The MultiUN Chinese-English parallel corpus provided by the IWSLT2013 Evaluation Campaign is aligned by chapter instead of sentence. It is difficult to train word alignment using this corpus. By investigating the MultiUN dataset, we found two alignment problems. First, instead of one to one sentence alignment, the sentence on the source side may

[1] http://www.ictclas.org/

[2] http://crfpp.sourceforge.net/

align to two or even more sentences on the target side. Second, the sentence on the source side may have no aligned sentences on the target side. The simple introduction of the MultiUN corpus may not help to improve the translation performance. Therefore, we proposed a method to extract parallel sentences from the MutiUN dataset.

Given the source sentence f with m words and target sentence e with n words, we suppose that words on the source side can be aligned to any words on the target side. The similarity between the two sentences is calculated as

$$\begin{aligned} sim(f,e) = \lambda_1 P(f|e) + \lambda_2 P(e|f) \\ +\lambda_3 L(e) + \lambda_4 L(f) + \lambda_5 R(f,e) \end{aligned} \tag{1}$$

where $P(f|e)$ is the average weights of the words in target sentences that are aligned to those in source sentences. It can be calculated as

$$P(f|e) = \frac{1}{m}\sum_{i=1}^{m}(\frac{1}{n}\sum_{j=1}^{n} log(p_{ij})) \tag{2}$$

where p_{ij} is the lexical translation probability and i and j are the position of words in source and target sentences. If there is no lexical translation probability between the aligned words, we set p_{ij} to be a minimal probability with e^{-10}. $P(e|f)$ can be calculated in the similar way.

$L(f)$ and $L(e)$ are used to punish the sentence length, which can be calculated as:

$$L(f) = log(m) \tag{3}$$

$$L(e) = log(n) \tag{4}$$

$R(f,e)$ is used to punish the length difference between the aligned sentences:

$$R(f,e) = log(max\{m,n\}/min\{m,n\}) \tag{5}$$

In our experiment, we supposed that the source sentence could align to at most 10 target sentences. All the possible alignments were scored by Equation 1, and the aligned sentences with the highest score were selected as the parallel sentence pairs. For the MultiUN parallel data, we finally obtained 7,819K sentence pairs to train our translation model.

4.2. Rule-based Translation of Named Entities

Although some named entities as time and numbers can be well translated by translation models, a majority of them cannot be correctly translated. Therefore, we introduced rule-based translation of named entities toolkit, which identifies time and number entities from the source sentences, and then translates it into the target language. We did not treat the named entities in the post-processing, but introduced them as normal translation rules. In our translation tasks, we first built a phrase table containing the named entities along with the translation results. Then we added it to the hierarchical translation model with a higher probability during the decoding process.

Take the phrase “26.5 million” in English-Chinese translation tasks as an example. Our toolkit will give us a parallel phrase as “26.5 million ||| 2,650 万”. By adding it to the translation model, the named entity of “26.5 million” can be translated correctly.

4.3. Translation Model Optimization

In TM training steps, we used the open source toolkit Giza++ [12] to get the bidirectional word alignments and combined them with $grow\text{-}diag\text{-}final\text{-}and$ method. Then we extracted the initial phrases and hierarchical phrases with heuristic extraction method to generated our original HPB model. We did forced alignment with the original HPB model on our training data and re-estimated the translation probabilities with the extracted phrases. At last, we used these refined phrases to generate our TM model for translation.

4.4. Language Model Interpolation

The language models used in our system are obtained by interpolating individual language models trained on the corpora of a different domain.

For the English language model, these training data sources are mentioned in section 2. First, the 5-gram modified Kneser-Ney discounted LMs are trained by using the SRILM toolkit [6]. Then the optimal interpolation weights for each LMs are estimated by using the tst2011 as the perplexity calculation text. The perplexities of each individual LMs and the final English LM are shown in table 3.

Table 3: *Perplexity and interpolation weights of the 5-gram English Language Models.*

data	tst2011	tst2013	weight
News Com	145.9	143.5	0.127
News Cra	88.8	93.0	0.697
Europarl	291.7	271.0	0.065
LDC Gigaword	403.7	345.7	0.109
UN	114.8	108.9	0.002
interpolate	84.3	84.1	-
prune	103.8	90.1	-

For the Chinese language model, four 4-gram modified Kneser-Ney discounted LMs are trained firstly. Then the optimal interpolation weights are estimated by using the tst2011. During weight estimating, the tst2010-2012 set dose not include the tst2011. However, during interpolating, the tst2010-2012 set includes dev2010, tst2010, tst2011 and tst2012, making the final Chinese LM contain the data of tst2011. Table 4 presents the perplexities of each LM.

4.5. Translated Rule Addition

In our HPB translation system, some of the words in source language are untranslated as no matched rules are available. However, these words actually have translations which can-

Table 4: *Perplexity and interpolation weights of the 4-gram Chinese Language Models.*

data	tst2011	tst2013	weight
WIT^3	188.9	217.5	0.630
UN	575.5	595.5	0.119
Google book grams	4553.1	4444.5	0.110
tst2010-2012	48.8	395.9	0.141
interpolate	83.4	205.2	-

not be extracted because of the restriction during phrase extraction. To avoid the non-translated phenomenon, we extracted word-to-word translation rules for these untranslated words from the lexical translations from word alignment.

For each untranslated word w_f in the source language, we looked up all of its target word w_e from the lexical probability table. The joint probability for each word pair is scored as:

$$P(w_f, w_e) = \frac{1}{2}(logP(w_f|w_e) + logP(w_e|w_f)) \tag{6}$$

where $P(w_f|w_e)$ and $P(w_e|w_f)$ are the bidirectional lexical probabilities.

We chose 3-best joint probabilities with the corresponding translations and added them to the hierarchical phrase table by a very lower probability. With the help of these additional phrase rules, some of the untranslated words could be translated correctly.

5. Experimental Results

We first trained our baseline HPB system (WIT^3) using the extracted WIT^3 parallel corpus. Then we did forced alignment with the baseline HPB model on the WIT^3 training data and obtained the optimized translation model (WIT^3+FA). We added the extracted parallel sentences from UN to WIT^3 and trained a larger translation model (WIT^3+UN). Considering the huge amount of parallel sentences in UN, we copied the WIT^3 corpus five times when combining these two types of parallel sentences. We also used the new translation model to do forced alignment on WIT^3 (WIT^3+UN+FA), which helps to generate more useful translation rules than the smaller translation model. At last, we added the translation rules, which were generated by the named entities toolkit and untranslated words, to our final HPB model as the translated template (WIT^3+UN+FA+Template). Both of Chinese-English and English-Chinese translation models are trained following the same way as described above. The results for Chinese-English and English-Chinese translation tasks on IWSLT 2012 and 2013 test sets are listed in Tables 5 and 6.

In Tables 5 and 6, the first two systems are trained using only the WIT^3 corpus. By adding the UN corpus to WIT^3 corpus, the translation performance was improved on both of the Chinese-English translation tasks, indicating that our extraction method can effectively get parallel sentences from MultiUN training corpus. However, the improvement on English-Chinese translation tasks is not significant as that on Chinese-English tasks. The results also show that our forced alignment can get better performance on the tasks with much smaller translation models. Moreover, the introduction of translation rules for named entities and untranslated words gives us the best results on the IWSLT 2013 translation tasks.

It is noteworthy that the satisfactory English-Chinese translation results on tst2012 are not attributed to our high quality translation system, but the incorrectly trained LM interpolated with data from tst2012.

Table 5: *Results for the Chinese-English MT task on IWSLT 2012 and 2013 test sets. The primary submission is the system combination of all the training methods.*

System	tst2012		tst2013	
	BLEU	TER	BLEU	TER
WIT^3	12.5	67.0	14.3	68.4
WIT^3+FA	12.6	66.7	14.4	67.7
WIT^3+UN	12.8	67.4	14.7	68.1
WIT^3+UN+FA	12.9	66.0	14.8	66.8
WIT^3+UN+FA+Template	13.0	65.8	15.0	66.4

Table 6: *Results for the English-Chinese MT task on IWSLT 2012 and 2013 test sets. The primary submission is the system combination of all the training methods.*

System	tst2012		tst2013	
	BLEU	TER	BLEU	TER
WIT^3	12.7	71.2	11.9	70.1
WIT^3+FA	12.8	70.6	11.9	69.6
WIT^3+UN	12.9	71.3	12.1	69.9
WIT^3+UN+FA	12.9	70.9	12.2	69.7
WIT^3+UN+FA+Template	13.0	70.7	12.3	69.6

6. Conclusion

In this paper, we presented our submission runs to the IWSLT 2013 Evaluation Campaign for the optional MT track on Chinese-English in both directions. We did our translation tasks by using the in-house hierarchical phrase-based decoder and Chinese word segmentation system as well as other open source toolkits. In particular, we used the forced alignment to optimize the HPB rules, which obtained the same translation results with a much smaller phrase table. To get better translation performances, we introduced the template rules into our decoder to deal with time and number entities and the words that exist in training data but do not have translation rules.

In future work, we plan to add some other features to our log-linear model and use the system combination methods to modify our system.

7. Acknowledgements

This work was supported by the National High Technology Research and Development Program (863 Program) of China under Grant No.2011AA01A207.

8. References

[1] D. Chiang, “A hierarchical phrase-based model for statistical machine translation”, in Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 2005, pp.263–270.

[2] X. Fu, W. Wei, S. Lu, B. Xu, “Filtration and Optimization for Hierarchical Phrase-based Model with Forced Alignment”, in Proceedings of the 12th China National Conference on Computational Linguistics (CCL), 2013.

[3] M. Cettolo, C. Girardi and M. Federico, “WIT3: Web Inventory of Transcribed and Translated Talks”, In Proceedings of EAMT, 2012, pp.261–268.

[4] A. Eisele and Y. Chen, “MultiUN: A Multilingual corpus from United Nation Documents”, in Proceedings of LREC, 2010.

[5] W. Chen, W. Wei, Z. Chen and B. Xu, “Integrating Multi-source Bilingual Information for Chinese Word Segmentation in Statistical Machine Translation”, in Proceedings of the 12th China National Conference on Computational Linguistics (CCL), 2013.

[6] A. Stolcke, “SRILM: An Extensible Language Modeling Toolkit”, in Proceedings of the 7th International Conference on Spoken Language Processing, 2002, pp.901–904.

[7] D. Chiang, “Hierarchical phrase-based translation” Computational Linguistics. 33(2), 2007, pp.201–228.

[8] F. J. Och and H. Ney, “Discriminative training and maximum entropy models for statistical machine translation”, in Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp.295–302.

[9] L. Huang and D. Chiang, “Forest Rescoring: Faster Decoding with Integrated Language Models”, in Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, 2007, pp.144–151.

[10] F. J. Och, “Minimum Error Rate Training in Statistical Machine Translation”, in Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics, 2003, pp.160–167.

[11] P. Koehn, F. J. Och, D. Mareu, “Statistical Phrase-Based Translation”, in Proceedings of the 2003 Conference of the NAACL, 2003, pp.48–54.

[12] F. J. Och and H. Ney, “A Systematic Comparison of Var-ious Statistical Alignment Models”, Computational Linguistics , 29(1), 2003, pp.19–51.