Technical Report of University of Macau for CWMT 2013

Longyue Wang, Liang Tian, Derek F. Wong, Lidia S. Chao, Francisco Oliveira, Shuo Li, Yiming Wang, Yi Lu
Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory
Department of Computer and Information Science
University of Macau, Macau S.A.R., China
{vincentwang0229, tianliang0123}@gmail.com,
{derekfw, lidiasc, olifran}@umac.mo,
{leevis1987, wang2008499, takamachi660}@gmail.com

澳门大学在第九届全国机器翻译评测中的技术报告

王龙跃，田亮，黄辉，周沁，Francisco Oliveira, 李硕，汪一鸣，卢义
自然语言处理与中葡机器翻译实验室
电脑与资讯科学系
澳门大学
{vincentwang0229, tianliang0123}@gmail.com,
{derekfw, lidiasc, olifran}@umac.mo,
{leevis1987, wang2008499, takamachi660}@gmail.com

Abstract

This paper describes our statistical machine translation (SMT) system and the evaluation results in the 9th China Workshop on Machine Translation (CWMT2013). Our Natural Language Processing & Portuguese-Chinese Machine Translation (NLP^{2}CT) laboratory of University of Macau participated in two shared tasks: English-to-Chinese (EC) news translation and Chinese-to-English (CE) news translation. We proposed a word pre-processing method which relies on statistical information extracted from the bilingual corpora for boosting word alignment relationships. In order to adapt our system to the domain sentences, a hybrid data selection approach was also applied to optimize the translation models. Finally, our primary system obtains 34.35 and 33.18 BLEU on EC-2009 testing and EC-2011 data separately, and 22.11 BLEU on CE-2009 task.

1 Introduction

This paper reports the techniques and performances of our NLP^{2}CT-SMT system in the CWMT 2013 English-to-Chinese and Chinese-to-English news translation tasks. We proposed a word pre-processing method which relies on statistical information extracted from the bilingual corpora, and a hybrid data selection approach for phrase-based SMT models (Koehn et al., 2007).

Chinese word segmentation (CWS), as a very first step for Chinese information processing, has a great impact on the results of Machine Translation (MT). Automatic word alignment can be defined as the problem of determining a translational correspondence at word level given a parallel corpus of aligned sentences (Ma et al., 2007). For Chinese, the sentence should be segmented into words before alignment. However, this segmentation is often performed in a monolingual context without considering any bilingual information. Although English sentences have natural delimiters, many terms have no meaning alone. Thus, they should be packed together to fit the linguistic phenomena of Chinese side. After investigating the English-Chinese MT via different CWS schemes and models that are learned from different benchmarking corpora (such as Penn Chinese Treebank, PKU People's Daily of China Corpus) as well as word packing, we relied on learned statistical information from the bilingual corpora in the word pre-processing approach. Firstly, we segment and tokenize the Chinese and English sentences, respectively. And then pack or decompose *N* consecutive Chinese or English words according to bilingual information such as auto-alignment, linguistic rules. Besides, we constructed an English-Chinese lexicon. The parallel sentences will be segmented and tokenized according to the lexicon when the words occur simultaneously. The objective of this step aims to maximize the *sure* alignments' quality by minimizing the *possible* alignments.

Data Selection aims at using the training data effectively by extracting sentences from large general-domain corpora in adapting SMT systems to domain-specific data. Given an in-domain test set or development set, the *N* most related sentences in general-domain corpus would be selected as a new pseudo in-domain subset. With this subset of the entire corpus, we re-train a better in-domain translation model. However, the kernel of this method is measuring the similarity between sentences. After investigating the state-of-the-art similarity criteria, we proposed a combination method (*i*CPE) (Wang et al., 2013), which combines **C**osine *tf-idf*, **P**erplexity and **E**dit distance techniques. Under the assumption that the CWMT news set is a kind of in-domain data, we adapted our translation system to the news data using *i*CPE (Wang et al., 2013).

All of these approaches play important roles in i) improving the quality of word alignment, ii) preventing irrelevant phrase pairs, and iii) optimizing the re-ordering of output sentences. Using BLEU (Papineni et al., 2002) as an evaluation metric, results indicate that the proposed approach can achieve consistent and significant improvement over the baseline system.

This paper is organized as follows. Section 2 and 3 detail the proposed data pre-processing and domain adaptation strategies. The data sets and experiments are given in Section 4 and 5. Finally, we compare and discuss the results in Section 6 followed by the conclusions to end the paper.

2 Data Preprocessing

2.1 Tokenization and Segmentation

In English or other similar languages, it is easy to identify tokens according to the spaces. In order to have a better tokenization approach, we also considered the following issues:

- Separates punctuation like periods from the beginning or the end of other tokens;
- Splits off contractions (-*n*'*t*, -'*ll*, -'*ve*) and possessives (like -'*s*), and make them as tokens;
- Recognizes and handles punctuations, numbers, and special formats (i.e. emails and URLs);
- Looks for multi-word units (proper name like "*White House*"), as well as chunks, e.g. article followed by noun like "*an apple*".

Moreover, it often happens that English and Chinese words are not 1-to-1 alignments. For instance, the English word "*insulator*" is aligned to three Chinese characters "绝(jue)"+"缘(yuan)"+"体(ti)", and the word expression "*New York*" is corresponding to two Chinese characters "纽约(Niuyue)". Thus, we also considered the following in the Chinese segmentation:

- Words in the dictionary or annotated corpora (e.g. Chinese Treebank) can be segmented as one unit. For example, "教室 (jiaoshi)" means "*classroom*" (1-to-1). However, splitting it into two characters "教" and "室", it means "*teach something*" and "*room*", separately;
- All the adjective words should be grouped together. Take one English sentence "*She is a beautiful girl*" as an example. The word "*beautiful*" can be translated into three Chinese characters "漂亮的". However, it is usually segmented into "漂亮" and "的" by the conventional Chinese word segmenters. We suggest putting them together and treating it as a single word. Similar auxiliary words such as "了", "着" are also considered and processed in the same way.

2.2 Proposed Segmentation Scheme/Model

The proposed Chinese segmenter was implemented by an augmented maximum matching model guided by statistical constraint determine the best segmentation, which is the reference length of the corresponding English sentence. Our hypothesis is that, parallel sentences that have similar length tend to produce full sure alignments, and on the hand, the number of possible alignments can be reduced. In order to prove this could best benefit MT, we investigated various CWS models trained on corpora with different annotation schemes, such as the ICTCLAS (Zhang et al., 2003), two Stanford Chinese Segmentation Models trained on Penn Chinese Treebank (Stanford-CWS_{CTB}) and PKU's People's Daily of China Corpus (Stanford-CWS_{PKU}) (Tseng et al., 2005). This includes also the character-based baseline where each character is treated as an individual word. In setting up the experiment, the Chinese sentences of both the train and test data are tokenized by different segmenters, as well as those for building the language models. The results are shown in Table 1 (training data refer Section 4 and testing data is CMWT ec-2009-news). Character-based segmentation generates the worst translation result. Both the ICTCLAS and Stanford-CWS_{PKU} models give similar values both at the average sentence length and the BLEU scores. The model using Stanford-CWS_{CTB} gives an improvement of 0.27 BLEU values, while the translation model based on proposed segmentation scheme outperforms all the others. It brings about an improvement of up to 3.47 BLEU over the baseline Character-based model.

Model/Scheme	**Ave. Len.**	**BLEU**
Character-based	29.16	17.77
ICTCLAS	19.52	20.44
Stanford-CWS_{PKU}	19.04	20.73
Stanford-CWS_{CTB}	15.78	21.00
Proposed Model	18.11	21.24

Table 1. BLEUs based on different Chinese segmentation (The average length of English sentences is 19.37).

The proposed Chinese segmenter heavily relies on a word list. It contains 1.2 million words and is collected from: (1) the modern Chinese encyclopedia (Ci Hai, 2003); (2) their English translations derived from the Oxford English-Chinese (Hornby, 1974); (3) the translation pairs derived from the word alignments of a four million parallel corpus trained with GIZA++ (Och and Ney, 2003), where the Chinese sentences are character-based tokenized. If multiple Chinese characters align to one English word, then the Chinese characters are treated as one word, and are added to the lexicon. During the segmentation, a word lattice is constructed to accommodate the possible words found from the lexicon. It takes both the contextual probabilities and reference length feature to determine the final segmentation result.

In particular, we found that when tokens of the Chinese sentence, which are segmented at character level, are equal to or approximately equal to the number of words in the English sentence, it can often obtain a better translation result in SMT compared to typical Chinese word segmentation methods.

3 Domain Specific Translation Model

3.1 Data Selection

Actually, data selection is one of the corpus weighting methods (Matsoukas et al., 2009). One of the dominant approaches is to select data suitable for the target domain from a large general-domain corpus (general corpus). Then a domain-adapted MT system could then be trained on these sub-corpora instead of the entire general corpus.

Three state-of-the-art data selection criteria are discussed below in different perspectives. The first is cosine *tf-idf* (term frequency-inverse document frequency) similarity. Hildebrand et al. (2005) applied this technique to construct Translation Memory (TM) and Language Model (LM) adaptation and they show that it is possible to adapt TMs for SMT by selecting similar sentences from general corpus. Furthermore, Lü et al. (2007) proposed re-sampling and re-weighting methods for online and offline TM optimization, which are closer to a real-life SMT system. The second one is perplexity-based approaches, which is used to score text segments according to an in-domain LM. Recently, Moore and Lewis (2010) derived the difference of the cross-entropy from a simple variant of Bayes rule. It was further developed by Axelrod et al. (2011) for SMT domain adaptation. The experimental results show that the fast and simple technique discard over 99% of the general corpus resulted in an increase of 1.8 in terms of BLEU score points. The third model is edit distance (ED), which is a widely used similarity measure for example-based MT (EBMT), known as Levenshtein distance (LD) (Levenshtein, 1966). Koehn and Senellart (2010) applied this method for convergence of TM and SMT. Then Leveling et al. (2012) investigated different approaches (e.g., LD and standard IR) to find similar sentences for EBMT. Therefore, we consider edit distance as a new similarity metric for this domain adaptation task.

After comparison (Wang et al., 2013), each individual retrieval model has its own advantages and disadvantages, which result in unclear or unstable performance. Instead of exploring any single individual models, we propose a hybrid model by performing linear interpolation on the three presented similarity metrics.

3.2 Proposed *i*CPE-M

Given the general-domain corpus which is the entire official data; the development and test set regarded as the in-domain corpus. We, firstly, used the three presented metrics to measure similarities between the general-domain data and in-domain data. Three subsets could be selected from the entire corpus. After training, three translation models could be obtained. Finally, we performed linear interpolation on these models. The phrase translation probability $\phi(\bar{f} \mid \bar{e})$ and the lexical weight $p_w(\bar{f} \mid \bar{e}, a)$ are estimated using Eq. 1 and Eq. 2, respectively.

$$\phi(\bar{f} \mid \bar{e}) = \sum_{i=0}^{n} \alpha_i \phi_i(\bar{f} \mid \bar{e}) \tag{1}$$

$$p_w(\bar{f} \mid \bar{e}, a) = \sum_{i=0}^{n} \beta_i p_{w,i}(\bar{f} \mid \bar{e}, a) \tag{2}$$

where i = 1, 2, 3 denote phrase translation probability and lexical weight trained with the sub-corpora retrieved by cosine *tf-idf*, perplexity-based and edit distance based approaches. α_i and β_i are the interpolation weights.

4 Data Sets

For training translation models, all the bilingual training data provided for the English-to-Chinese and Chinese-to-English news from the CWMT 2013 organizer are used (cwmt2013-corpora). The total number of the sentences after tokenization, normalization and filtering is approximately 3.3 million sentences.

As out-of-list data from the organizer for the parallel corpora, 4,157,556 sentences of UM-Corpus (in-house English-Chinese parallel data) are added to the cwmt2013-corpora. After removing repeated and unparalleled sentences in the combined two parts, there are 7,445,190 sentences left and the statistics of the combined parallel corpus are presented in Table 2. The statistics of Chinese sentences are counted in character level (each Chinese character is treated as one token).

Lang.	Token	Av. Len.	Type
English	152,161,233	19.37	1,655,080
Chinese	229,110,265	29.16	397,442

Table 2. Statistics of cwmt2013-corpora + UM-Corpus.

5 Experiments

In the experiments described below, the phrase-based Moses decoder (Koehn et al., 2007) is used, GIZA++ is adopted to obtain bidirectional word alignment (Och and Ney, 2003), and the heuristic strategy of *grow-diag-final-and* (Koehn et al., 2007) is used to combine the word alignments of source-to-target and target-to-source directions. The combined word alignments are used to extract the phrase translation and the reordering tables. All the training parameters applied are default values used by Moses. There is no optimization step, such as tuning (Och and Ney, 2003; Bertoldi et al., 2009) and pruning (Johnson et al., 2007; Ling et al., 2012).

The English tokenization is based on the scripts *tokenizer.perl* in Moses and the Chinese segmentation is based on the *UM-CSegmenter*. The IRSTLM toolkit (Federico et al., 2008) with modified Kneser-Ney smoothing (Chen and Goodman, 1996) was used to train *5*-gram language models.

In data selection processing, we firstly build an in-domain model with the development set, which is regarded as an in-domain corpus. Then each sentence in the general-domain corpus is evaluated according to the similarity with the in-domain model. Finally, a subset of the entire corpus is built by selecting the most related sentence pairs.

6 Results and Discussions

We applied the two proposed approaches in our system for the CMWT 2013 English to Chinese (EC) and Chinese to English (CE) news evaluation task.

The baseline system was trained with the official pre-processed data. About our system, we firstly segment the entire CWMT released corpora. Secondly, we employed *i*CPE-M selection method to obtain a new subset of the entire data set. Finally, we use Moses to train an optimized translation model with the selected data set.

We evaluated these two systems with the CWMT testing data (ce-2009-news, ec-2009-

ce-2009-news	BLEU4-SBP	BLEU4	NIST6	GTM	mWER	mPER	ICT
UM	0.2113	0.2211	6.8197	0.6757	0.7011	0.5221	0.3278
Baseline	0.2088	0.2183	6.0349	0.5729	0.6425	0.5055	0.3136
Diff.	0.0025	0.0028	0.7848	0.1028	0.0586	0.0166	0.0142

Table 3. Translation results of CE test data.

ec-2009-news	BLEU5-SBP	BLEU5	NIST6	GTM	mWER	mPER	ICT
UM	0.3248	0.3435	9.6079	0.7846	0.6541	0.3875	0.3955
Baseline	0.3124	0.3369	9.5548	0.7858	0.6258	0.3735	0.3530
Diff.	0.0124	0.0066	0.0531	-0.0012	0.0283	0.014	0.0425

Table 4. Translation results of EC-2009 test data.

ec-2011-news	BLEU5-SBP	BLEU5	NIST6	GTM	mWER	mPER	ICT
UM	0.3164	0.3318	9.3382	0.7673	0.6387	0.3904	0.3685
Baseline	0.3072	0.3292	9.0422	0.7145	0.6191	0.3810	0.3471
Diff.	0.0092	0.0026	0.296	0.0528	0.0196	0.0094	0.0214

Table 5. Translation results of EC-2011 test data.

news and ec-2011-news) using multiple evaluation metrics, such as BLEU-SBP (Chiang et al., 2008), BLEU, NIST, GTM, mWER, mPER, and ICT. The evaluation results are shown in Table 3, 4 and 5 respectively.

The improvements show that our proposed methods could be used to boost a state-of-the-art SMT system. In all tasks, our system has better results than the baseline system. For example, in Table 4, our system outperforms the baseline by 0.0124 BLEU5-SBP points. However, the improvements in other tasks are not very clear. There are two main reasons for this:

- The segmentation may have different impacts on different translation direction. This method has a better benefit for English-Chinese direction.
- Data selection is a domain adaptation method. Our experiments are conducted based on the assumption that news is a kind of domain. However, it may be inaccurate. The news may also contain sport, political events, entertainment, etc.

7 Conclusion

In this paper, we proposed two models in application to the SMT system. They are the task oriented segmentation for SMT and hybrid data selection and combination model. We not only report their performance respectively but also explore the combination method for the domain specific Chinese-English translation. From the in-house experiments, the results are quite promising. However, the final results are not as good as expected. The problem should be further investigated.

Acknowledgments

The authors are grateful to the Science and Technology Development Fund of Macau and the Research Committee of the University of Macau for the funding support for our research, under the reference No. 017/2009/A and MYRG076(Y1-L2)-FST13-WF. The authors also wish to thank Prof. Derek Fai Wong and Dr. Francisco Oliveira who gave us many helpful comments.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* pp. 355–362.

Nicola Bertoldi, Barry Haddow, and Jean-Baptiste Fouet. 2009. Improved Minimum Error Rate Training in Moses. *The Prague Bulletin of Mathematical Linguistics*, Vol. 91, pp. 7–16

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation* pp. 224–232.

David Chiang, Steve DeNeefe, Yee Seng Chan, and Hwee Tou Ng. 2008. Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proceedings of EMNLP 2008*, pp. 610-619.

Stanley F. Chen, and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics* pp. 310–318.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech* pp. 1618–1621.

Ci Hai. 2003. *Modern Chinese Encyclopedia*. China Research Institute of the Publishing Science.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of EAMT* Vol. 2005, pp. 133–142.

Albert Sydney Hornby, Anthony Paul Cowie, Alfred Charles Gimson, and J. Windsor Lewis. 1974. *Oxford advanced learner's dictionary of current English* Vol. 1428. Cambridge Univ Press.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of EMNLP-CoNLL*, pp. 967–975.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 177–180.

Philipp Koehn, and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *Proceedings of AMTA Workshop on MT Research and the Translation Industry* pp. 21–31.

Johannes Leveling, Debasis Ganguly, Sandipan Dandapat, and Gareth Jones. 2012. Approximate Sentence Retrieval for Scalable and Efficient Example-based Machine Translation. In *COLING 2012* pp. 1571–1586.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady* Vol. 10, p. 707.

Wang Ling, Nadi Tomeh, Guang Xiang, Alan Black, and Isabel Trancoso. 2012. Improving Relative-Entropy Pruning using Statistical Significance. In *COLING 2012*, pp. 713–722.

Yajuan Lü, Jin Huang, and Qun Liu. 2007. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. In *EMNLP-CoNLL* pp. 343–350.

Yanjun Ma, Nicolas Stroppa, and Andy Way. 2007. Bootstrapping word alignment via word packing. In *Annual Meeting-Association for Computational Linguistics* Vol. 45, p. 304.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2* pp. 708–717.

Robert C. Moore, and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers* pp. 220–224.

Franz Josef Och, and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* pp. 311–318.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing* Vol. 171.

Longyue Wang, Derek F. Wong, Lidia S. Chao, Yi Lu, and Junwen Xing. 2013. iCPE: A Hybrid Data Selection Model for SMT Domain Adaptation. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data* pp. 280–290. Springer.

Longyue Wang, Derek F. Wong, Lidia S. Chao, and Junwen Xing. 2012. CRFs-Based Chinese Word Segmentation for Micro-Blog with Small-Scale Data. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing* pp. 51–57. Tianjin, China.

Hua-Ping Zhang, Qun Liu, Xue-Qi Cheng, Hao Zhang, and Hong-Kui Yu. 2003. Chinese lexical analysis using hierarchical hidden markov model. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17* pp. 63–70.