An Improved Patent Machine Translation System Using Adaptive Enhancement for NTCIR-10 PatentMT Task*

Hai Zhao[1,2†] Jingyi Zhang[1,2]
(1) MOE-Microsoft Key Laboratory of Intelligent Computing and Intelligent System;
(2) Department of Computer Science and Engineering, Shanghai Jiao Tong University,
#800 Dongchuan Road, Shanghai, China, 200240
zhaohai@cs.sjtu.edu.cn,
zhangjingyi1990@yeah.net

Masao Utiyama Eiichro Sumita
Multilingual Translation Laboratory, MASTAR Project
National Institute of Information and Communications Technology
3-5 Hikaridai, Keihanna Science City, Kyoto,
619-0289, Japan
{mutiyama, eiichiro.sumita}@nict.go.jp

ABSTRACT

This paper describes the work that we conducted for the Chinese-English (CE) task of the NTCIR-10 patent machine translation evaluation. We built standard phrase-based and hierarchical phrase-based statistical machine translation (SMT) systems with optimized word segmentation, adaptive language model and improved parameter tuning strategy. Our systems outperform official baselines by approximate 2 BLEU points.

Keywords

Statistical machine translation

Team Name

SJTU

Subtasks

Chinese-to-English patent MT

1. INTRODUCTION

This paper describes our translation systems for the Chinese-to-English subtask of the NTCIR-10 patent machine translation evaluation [2]. We built two SMT systems based on the phrase-based [7] and hierarchical phrase-based [1] models, respectively, then applied three enhanced techniques to improve translation performance for both systems.

Based on our previous work in [14], we used an optimized dictionary to perform word segmentation over source language.

Our adaptation for language model training is based on the fact that n-gram language model gives higher probability for a word sequence that has higher frequency. We selected a small set of English sentences from parallel corpus whose corresponding Chinese parts are similar with test set and expanded the corpus for language model (LM) training to improve translation performance on test set. [10] adopted an LM adaptation approach that interpolates a general LM with "bias LM" to prefer correct translations for test set. It is different from our approach that we focus on selecting corpus for LM training.

Tuning model parameters of machine translation relies on a fundamental assumption that model parameters that give better performance on development set will do so on test set as well. However, two groups of model parameters that give best performance on two different sets respectively are probably different and the fundamental assumption only works for similar datasets. So for better performance over test set, it is reasonable to tune model parameters on a development set that is as similar with test set as possible. In our systems, development set was selectively constructed according to test set as done in [8] but using different construction method.

In this paper, when we say similarity between a Chinese-to-English parallel dataset and a Chinese dataset, we mean similarity between the Chinese side of parallel dataset and Chinese dataset.

2. OUR SYSTEMS

2.1 Basic Systems

In our experiments for Patent Machine Translation Chinese-to-English (CE) subtask at NTCIR-10 [3] shared task, only training, development and test sets officially provided were used. We trained standard phrase-based and hierarchical phrase-based SMT systems with a lexicalized reordering model [5] for phrase-based model on training set and tuned model parameters on development set using Moses [6]. A 5-gram language model (LM) was trained on the English side of the training set by IRST LM Toolkit[1] for both systems. GIZA++ [13] and the *grow-diag-final-and* heuristic [7] were used to obtain symmetric word alignment model. There are four kinds of test data in NTCIR-10 PatentMT CE subtask: Intrinsic Evaluation (IE), Patent Examination Evaluation (PEE), Chronological Evaluation (ChE), Multilingual Evaluation (ME). We submitted translation results from both phrase-based (SJTU-2) and hierarchical phrase-based (SJTU-1) SMT systems for IE test data. For the rest three test datasets, we submitted translation results by hierarchical phrase-based SMT system since usually hierarchical

*This work was partially supported by the National Natural Science Foundation of China (Grant No. 60903119, Grant No. 61170114, and Grant No. 61272248), and the National Basic Research Program of China (Grant No. 2009CB320901 and Grant No.2013CB329401).

†This work was partially done as the first author was at NICT with support of MASTAR project.

[1]http://hlt.fbk.eu/en/irstlm

phrase-based model gives better performance than phrase-based model.

2.2 Chinese Word Segmentation

We used an empirical dictionary optimization (more precisely, pruning) algorithm to improve the related dictionary-based segmenters that was proposed in [14]. The algorithm is motivated by the empirical observation that most words in a given dictionary provide poor information for aligning and decoding in a specific SMT task. As a dictionary with n words is given, the algorithm is to find a subset of the dictionary to maximize the machine translation performance. Both alignment model and BLEU scores given by minimum-error-rate-training (MERT) on the development set are exploited to determine the optimized dictionary, and aligning counter is adopted as the metric to evaluate how well a word inside the dictionary individually contributes to machine translation. This detailed algorithm is given in Algorithm 1. There are two layers of loops in the algorithm, however, this algorithm usually ends after running the MT routine less than 12 times. In addition, against existing dictionary optimization approaches [12, 11], the algorithm is non-parametric, which is more convenient and practical for use. Using the above algorithm on NTCIR-9 patent MT parallel corpus, we finally obtain a very small dictionary with only about 10K words. To enhance out-of-vocabulary word recognition, we further merge the obtained dictionary and other online lexicon with about 100K words from Beijing University to work for the actual segmentation task [9].

2.3 Improved Tuning Strategy

We adopt an improved tuning strategy based on the assumption that using a development set which is more similar with test set can improve the performance of tuning. Note it is different from our previous work [4] that is about how to perform parameter tuning with an optimal tuning schedule. In our systems, an appropriate subset extracted from a range of datasets is selected as development set for a given test set according to a predefined similarity.

Given test set T and candidate set C with size m and l respectively, we use edit distance to select a part of C with size k that have most similarity with T according to Algorithm 2. Only Chinese side of parallel corpus is used in Algorithm 2. One development set and two test sets (ChE and ME) with reference that contain 2,000 sentence pairs respectively are officially provided. We selected 2,000 sentence pairs from the original official development set and ME test set to tune for ChE subtask. For IE, PEE and ME subtasks, 2,000 sentence pairs for tuning were selected from the original development set and ChE test set.

2.4 Language Model Adaptation

To improve the LM, we retrain an improved LM on a revised dataset. For the English side of training set that the original LM is trained on, we perform a slight modification by duplicating a small part of the dataset whose corresponding Chinese side in training set are more similar with the input test set. We use the same algorithm for development set optimization to select the duplicated part of data. Additional experiments were done to determine the size of the duplicated part. We translated the officially provided development set using the standard phrase-based SMT system with default model parameters and LM adapted by different sized duplicated part. The results of experiments to determine the size of duplicated set are shown in Table 1. The

Algorithm 1 Dictionary optimization

INPUT An initial dictionary, D
while do
  Segment the MT corpus with D.
  Run GIZA++ for alignment model M.
  Run MERT and receive BLEU score(on the dev set) b.
  Rank all words in D according to aligning times.
  Let $counter$=0 and n=2
  while $counter$ <2 **do**
    Extract top $1/n$ words from D according to aligning times to build dictionary D_n.
    Run GIZA++, MERT and receive BLEU score b_n.
    if $b_n < b_{n-1}$ **then**
      $counter = counter + 1$.
    end if
    $n = n + 1$
  end while
  if max $\{b_i\} < b$ **then**
    return D
  end if
  Let $D_0 =$ arg max$_{D_i}$ b_i and b= max $\{b_i\}$
  Let $D' = D$ - D_0
  According to aligning times in M, divide D' into $2n$ dictionaries, D'_1,...,D'_n, ..., D'_{2n}.
  for top n most-aligned dictionaries, $D'_i, i = 1, ..., n$ **do**
    Segment the MT corpus with D_0+D'_i.
    Run GIZA++, MERT and receive BLEU score b'_i.
  end for
  if max $\{b'_i\} < b$ **then**
    return D_0
  end if
  Let $D =$ arg max$_{D_0+D'_i}$ b'_i and b= max $\{b'_i\}$
end while

Algorithm 2 Development set optimization

Input: $T = \{t_i | i = 1, 2, \cdots m\}$, $C = \{c_i | i = 1, 2, \cdots l\}$, k, where t_i or c_i represents a sentence.
Output: Development set D.
Initialize a queue q: $\{(a_i, b_i) | i = 1, 2, \cdots\}$
for $i := 1$ **to** l **do**
  $ed := \_MAX\_VALUE$
  for $j := 1$ **to** m **do**
    $ed' := edit\_distance(t_j, c_i)$
    if $ed' < ed$ **then**
      $ed := ed'$
    end if
  end for
  Add a queue element with $a = ed$ and $b = i$ in q at an appropriate position that keeps a increasingly sorted.
  if the size of q is larger than k **then**
    Remove the last element of q.
  end if
end for
for $i := 1$ **to** k **do**
  add b_i into D
end for

size of the duplicated set "40,000" with the best BLEU score on development set was chosen at last.

Size of added set	0	10K	20K	40K	80K
BLEU(%)	30.85	31.10	31.12	31.19	31.14

Table 1: BLEU scores with different sized duplicated sets.

3. EVALUATION RESULTS

System	baseline1	baseline2	SJTU-1	SJTU-2
BLEU	0.3252	0.3134	0.3437	0.3396
NIST	8.3027	8.2076	8.6372	8.6137

Table 2: Automatic evaluation results of Intrinsic Evaluation.

Table 2 shows the automatic evaluation results for IE subtask including our two systems and two baseline systems in NTCIR-10 PatentMT [2]. The "baseline1" system was referred to the hierarchical phrase-based system while the "baseline2" the phrase-based system. It can be seen that the additional techniques improved performances of both models by approximate 2 BLEU points. Besides the automatic evaluation, the NTCIR-10 organizer also carried out manual evaluation about adequacy and acceptability of translations. The adequacy were divided into five levels: 1, 2, 3, 4 and 5, from the worst to the best. And the acceptability also had 5 levels: AA, A, B, C and F, from the best to the worst. Table 3 shows the adequacy scores of our system "SJTU-1" and the baseline system "baseline1" while Table 4 gives the acceptability scores of "SJTU-1" since the NTCIR-10 organizer did not provide the acceptability scores of the baseline systems. As can be seen, our system "SJTU-1" also gave better translation compared to the baseline systems according to the manual measure.

System	Average adequacy	level scores 5	4	3	2	1
baseline1	3.23	46	73	91	84	6
SJTU-1	3.32	63	60	93	79	5

Table 3: Manual evaluation results (adequacy) of Intrinsic Evaluation.

Table 5 shows BLEU scores for ChE and ME provided by the organizer. The improvements of our system compared to the baseline systems on ChE and ME are accordant with the ones on IE.

4. CONCLUSIONS

Our systems described in this paper participated in the Chinese-to-English subtask of the NTCIR-10 PatentMT task and the reported results outperform official baselines. It is worth noting that the additional techniques implemented for standard SMT systems may be applied to other types of translation tasks as they do not make use of any characteristics of patent documents.

System	AA	A	B	C	F
SJTU-1	21	22	29	34	194

Table 4: Manual evaluation results (acceptability) of Intrinsic Evaluation.

System	baseline1	baseline2	SJTU-1
ChE	0.3074	0.2934	0.3274
ME	0.1796	0.1805	0.1933

Table 5: BLEU scores of ChE and ME.

5. REFERENCES

[1] D. Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics, 2005.

[2] I. Goto, K. Chow, B. Lu, E. Sumita, and B. Tsou. Overview of the patent machine translation task at the ntcir-10 workshop. In *Proceedings of NTCIR*, 2012.

[3] I. Goto, B. Lu, K. Chow, E. Sumita, and B. Tsou. Overview of the patent machine translation task at the ntcir-9 workshop. In *Proceedings of NTCIR*, volume 9, pages 559–578, 2011.

[4] C. Hui, H. Zhao, Y. Song, and B.-L. Lu. An empirical study on development set selection strategy for machine translation learning. In *Proceedings of WMT-2010*, pages 67–71, Uppsala, Sweden, July 2010.

[5] P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *International Workshop on Spoken Language Translation*, 2005.

[6] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Annual meeting-association for computational linguistics*, volume 45, page 2, 2007.

[7] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics, 2003.

[8] M. Li, Y. Zhao, D. Zhang, and M. Zhou. Adaptive development data selection for log-linear model in statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 662–670. Association for Computational Linguistics, 2010.

[9] J. K. Low, H. T. Ng, and W. Guo. A maximum entropy approach to Chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 161–164, Jeju Island, Korea, 2005.

[10] J. Ma and S. Matsoukas. Bbn's systems for the chinese-english sub-task of the ntcir-9 patentmt evaluation. In *Proceedings of NTCIR*, volume 9, pages

579–584, 2011.

[11] J. Ma and S. Matsoukas. BBN's systems for the Chinese-English sub-task of the NTCIR-9 PatentMT evaluation. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 579–584, Tokyo, Japan, December 2011.

[12] Y. Ma and A. Way. Bilingually motivated domain-adapted word segmentation for statistical machine translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 549–557, Athens, Greece, April 2009. Association for Computational Linguistics.

[13] F. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51, 2003.

[14] H. Zhao, M. Utiyama, E. Sumita, and B.-L. Lu. An empirical study on word segmentation for chinese machine translation. In *Proceedings of CICLing 2013, Part II, LNCS Vol.7817*, pages 248–263, Samos, Greece, March 2013.