ICIC International ©2017 ISSN 1881-803X
pp. 1425–1433

AN IMPROVED ACTIVE LEARNING FRAMEWORK FOR LOW-RESOURCE STATISTICAL MACHINE TRANSLATION

Guo Xie[1], Xinzhu Wang[1] and Jinhua Du[2]

[1]Shaanxi Key Laboratory of Complex System Control and Intelligent Information Processing
Xi'an University of Technology
No. 5, South Jinhua Road, Xi'an 710048, P. R. China

[2]ADAPT
School of Computing
Dublin City University
Dublin 9, D09 NA55, Ireland
jinhua.du@adaptcentre.ie

Received April 2017; accepted June 2017

Abstract. *An active learning-based (AL) framework for low-resource statistical machine translation (SMT) is an efficient and feasible way to acquire a number of high-quality parallel data to improve translation quality. One of the key issues in this framework is to select the most informative sentences from the monolingual corpus for human translation. However, when the sentence length in the monolingual corpus varies over a wide range, the basic selection algorithm performs worse than random selection. This paper proposes two improved algorithms, namely the sentence length-informed (SLI) and the weighted sentence length-informed (WSLI) methods, to alleviate the influence of the sentence length on selection performance by introducing a length penalty factor to dynamically penalize shorter sentences. Simulation experiments on small-scale NIST Chinese-English task, French-English WMT task and English-Spanish automatic post-editing (APE) task show that the proposed methods significantly improve translation quality compared to the random and the basic selection methods.*

Keywords: Active learning, Low resource language, Informative sentence, Monolingual data

1. **Introduction.** Statistical machine translation (SMT) is a data-driven method, where the scale and quality of the parallel data, and especially the large-scale high quality parallel data, are crucial for obtaining good translation performance.

However, it is not the case for many resource-poor language pairs, such as Urdu-English and Chinese-Italian, where large amounts of speakers exist for both languages. A number of methods have been presented to alleviate this problem, such as paraphrasing [1, 2, 3], and utilizing other rich resources [4, 16, 17]. Considering the reality that a large amount of monolingual data can be easily acquired from the Web or other digital media, the active learning framework for SMT has been proposed to facilitate the issue of parallel data scarcity [5, 6, 7, 8, 9, 10, 13, 14]. In this framework, only a little human involvement can bring a significant improvement to the translation performance via manually translating information-rich monolingual sentences.

One of the key issues in the AL framework is to choose the information-rich sentences. As to the SMT task, the basic idea of selecting sentences with high information is to find some sentences at each iteration to maximize the improvement in translation quality. In doing so, the sentences selected contain rich information. Intuitively, if phrases or words in a sentence occur frequently in the unlabeled (monolingual) data while appearing less in the labeled (parallel) data, this sentence might be more informative to the labeled (small-scale parallel) data because it introduces more new knowledge.

In this paper, we utilize three language pairs, namely Chinese-English, French-English and English-Spanish, with small-scale parallel data sets to study the AL-based framework for low-resource SMT.[1] Firstly, we studied the basic sentence selection algorithms, that is, the translation unit based strategies, and found in our experiments that these algorithms are apt to select shorter sentences compared to the random method, which makes the system performance worse [13, 14]. Based on our observations and analysis, we then proposed two improved algorithms, namely the sentence length-informed and the weighted sentence length-informed methods, to improve the performance and robustness of algorithms. Experimental results on these language pairs show the effectiveness of the proposed methods compared to the baseline.

The remainder of this paper is organized as follows. Section 2 briefs the related work on AL-based SMT. In Section 3, the basic and our modified AL frameworks for low-resource SMT are introduced. Section 4 formalizes the informative sentence selection question and describes three basic translation unit-based selection algorithms. Based on the findings and investigations, two improved sentence selection algorithms are proposed in Section 5 and comparative experiments with the baseline and the basic algorithm are carried out in Section 6. Section 7 concludes and gives avenues for future work.

2. **Related Work.** The topic of AL-based low-resource SMT has attracted a number of researchers. Haffari et al. [5] firstly proposed a practical active learning framework for SMT where a number of high-quality parallel data are acquired from large-scale monolingual data. Proper human costs are iteratively involved to translate information-rich sentences. Experimental results show that generally the translation unit-based selection strategies, namely phrases and n-grams, performed best compared to other methods such as random selection, translation confidence, and inverse model. They also applied the active learning framework to multilingual language pairs to acquire high-quality parallel data and verified its effectiveness [6].

In 2010, Ambati et al. proposed an active crowd translation (ACT) paradigm where active learning and crowd-sourcing come together to enable automatic translation for low-resource language pairs. Active learning is used to reduce costs of label acquisition by prioritizing the most informative data for annotation, while crowd-sourcing reduces costs by using the power of the crowd to replace expensive language experts. They compared active learning strategies with strong baselines, and showed significant improvements in translation quality even with less data [7]. In 2011, Ambati et al. came up with another active learning approach with multiple annotations for addressing the problem of building comparable corpora in low-resource scenarios. Experimental results also showed that the active learning strategy is effective [9].

In 2012, Bakhshaei and Khadivi applied a pool-based active learning strategy for improving Farsi-English MT. They increased n in the n-gram feature from 4 to 5, and verified that the sentence selection algorithms such as translation units, and translation confidence, perform better than the random selection method for this task [10].

In 2013, Du and Zhang applied the AL framework to Chinese-English SMT and found that the sentence length has a significant impact on system performance. They proposed a sentence length filtering strategy to filter out short sentences, and the experimental results showed that the translation unit-based method outperformed the random selection. However, in practical scenario, there are not enough parallel resources for low-resource language pairs, so this strategy of removing short sentences is not feasible [13].

[1] While for practical purposes, we acknowledge that these three language pairs hardly merit description as resource-poor, we utilize relative small amounts of training data to simulate SMT in low-resource applications. Having demonstrated the effectiveness of our proposed algorithms on these small data sets, in future work we intend to expand our experiments to resource-poor language pairs and tasks.

In 2014, Du et al. proposed a sentence length-informed method to improve the performance of AL-based SMT. This method introduced a dynamic factor to penalize short sentences in the process of informative sentence selection. Their experimental results on Chinese-English and French-English tasks showed that the proposed method significantly outperformed the random selection. However, the sentences selected by the sentence length informed method are remarkably longer than the random method, which would significantly improve the human cost [14].

This paper firstly reviews the active learning framework for SMT and basic informative sentence selection algorithms, and then proposes two improved informative sentence selection algorithms – a sentence length-informed and a weighted sentence length-informed – to better select informative sentences.

3. **Active Learning Framework for SMT.** The basic workflow of the AL framework for SMT is to obtain parallel data from large-scale monolingual corpora and add this to the initial small-scale parallel corpus for training.

We denote the initial parallel corpus as $L := \{(f_i, e_i)\}$, and the large-scale monolingual corpus as $U := \{f_j\}$. The key step is to design an algorithm to select highly informative sentences and submit them to human translators. A workflow of the AL framework is illustrated in Figure 1.

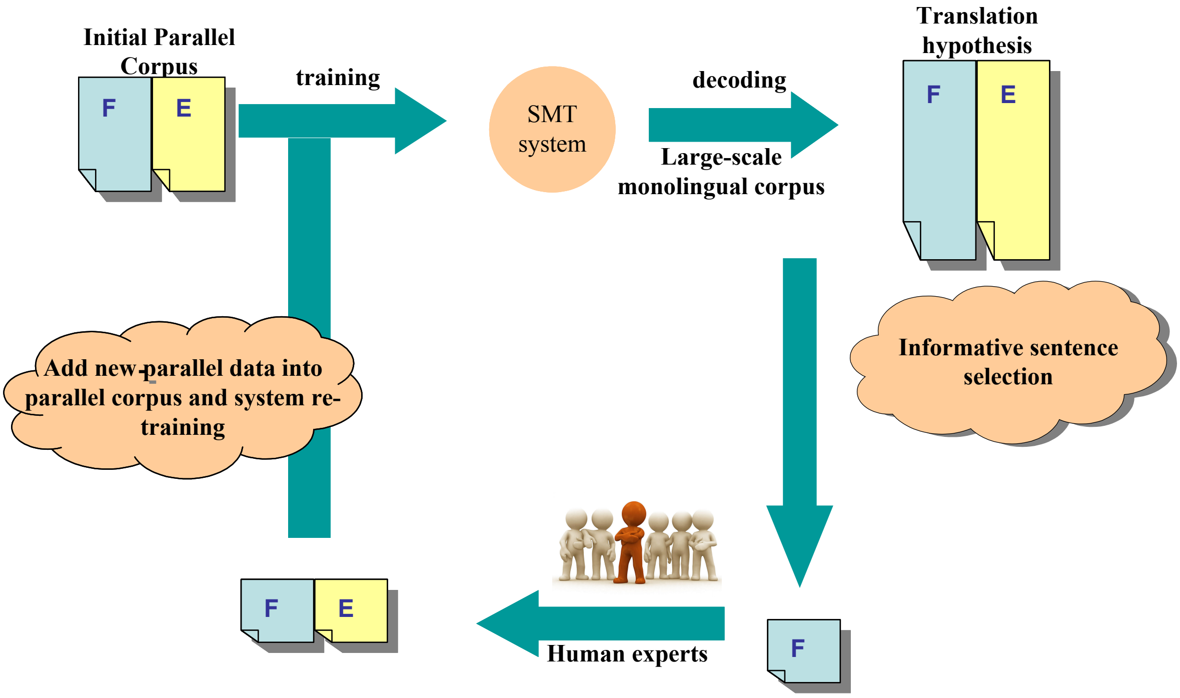
Initial Parallel Corpus
F
E
training
SMT system
decoding
Large-scale monolingual corpus
Translation hypothesis
F
E
Informative sentence selection
Add new parallel data into parallel corpus and system re-training
Human experts
F
E
F

FIGURE 1. The workflow of active learning framework for SMT

Generally, the active learning framework has two prerequisites:

- a small-scale initial parallel corpus used to build a baseline SMT system;
- a large-scale monolingual corpus to acquire extra bilingual data.

More importantly, there are two key research issues in the AL framework for SMT:

- how to design an efficient algorithm to evaluate the information that a sentence contains and select the information-rich sentences;
- how to utilize the new parallel data to train and update the SMT system.

In our work, we mainly focus on the first question, i.e., studying the informative sentence selection algorithms on different language pairs.

As to the second issue, different from the system used in Haffari & Sarkar [6], we only use the L-trained model to run our SMT system at each iteration without the extra translation model that is trained by U and its translation hypothesis [13]. Thus, the modified active learning framework using translation units based methods in our experiments is shown in "Algorithm 1".

Algorithm 1 Modified AL-SMT

```
1: Given bilingual corpus $L$, and monolingual corpus $U$.
2: $M_{F \to E} = \mathbf{train}(L)$
3: for $t = 1, 2, \ldots, N$ do
4:   Generate "Phrase Set" and compute sentence scores
5:   Select $k$ sentences from $U$, and acquire human translation
6:   Remove the $k$ sentences from $U$, and add the $k$
     sentence pairs to $L$.
7:   Update $M_{F \to E} = \mathbf{train}(L)$
8:   Evaluate the system performance on the test set.
9: end for
```

4. Basic Sentence Selection Algorithms.

4.1. Mathematical description.

We formalize the information-rich sentence selection algorithm as follows.

Given a monolingual corpus U, an initial parallel corpus L, and a sentence s consisting of m possible translation units $\{x | x \in X_s^m\}$ in U, the goal is to choose a sentence $\hat{s}$ with the highest score ϕ under a certain metric F as the most informative candidate. Therefore, the metric F to evaluate how much information a sentence has is the most important in a selection algorithm. This process can be defined as a *quadruple* in (1):

$$\phi(s) = F(X, s, U, L) \tag{1}$$

4.2. Geom-phrase and Arith-phrase algorithms.

In these two methods, the basic unit for computing the score $\phi(s)$ of a sentence s is the phrase. The Geom-Phrase algorithm is as in (2):

$$\phi(s) = \left[\prod_{x \in X_s^m} \frac{P(x|U)}{P(x|L)} \right]^{\frac{1}{|X_s^m|}} \tag{2}$$

where X_s^m is the set of possible phrases that the sentence s can offer, $P(x|U)$ and $P(x|L)$ are the probabilities of observing x in U and L, respectively, which are calculated as in (3) and (4):

$$P(x|U) = \frac{count(x) + \epsilon}{\sum_{x \in X_U^m} count(x) + \epsilon} \tag{3}$$

$$P(x|L) = \frac{count(x) + \epsilon}{\sum_{x \in X_L^m} count(x) + \epsilon} \tag{4}$$

where ϵ is the smoothing factor. X_U^m indicates the set of phrases that indeed occur in U, and X_L^m represents the set of phrases that truly appear in L.

The Arith-Phrase method is the logarithmic form of "Geom-phrase" as defined in (5):[2]

$$\phi(s) = \frac{1}{|X_s^m|} \sum_{x \in X_s^m} \frac{P(x|U)}{P(x|L)} \tag{5}$$

Haffari et al. [5] extracted the phrases used in Equation (2) and Equation (5) from the k-best list of translations of a sentence s in U. Additionally, they obtained the out-of-vocabulary (OOV) from the translations and gave them a uniform probability. However, the k-best list cannot provide all possible phrases that a source sentence can offer. In order to make the phrase set approximate to a complete set, i.e., include all possible phrases that a sentence can offer, we utilize the phrase table generated by L to retrieve all possible phrases and collect OOVs that do not occur in the phrase table.

4.3. **Geom *n*-grams.** More generally, n-grams are often used as an alternative to phrases to measure the importance score. The resulting score is the weighted combination of the n-gram-based scores, as in (6):

$$\phi(s) = \sum_{n=1}^{N} \frac{\omega_n}{|X_s^n|} \sum_{x \in X_s^n} \log \frac{P(x|U,n)}{P(x|L,n)} \tag{6}$$

where $X_s^n\{n = 1, \ldots, N\}$ denotes n-grams in the sentence s, and $P(x|U,n)$ and $P(x|L,n)$ are the probabilities of x occurring in the set of n-grams in U and L, respectively. ω_n is the weight that adjusts the importance of the scores of n-grams with different lengths.

In previous work, the "Geom n-grams" performed worse than the "Geom-Phrase" and "Arith-Phrase" methods on many language pairs [6, 13], so we use the "Arith-Phrase" as the state-of-the-art translation unit-based method to carry out our study.

5. **Improved Arith-Phrase Algorithms.** The basic informative sentence selection algorithms cannot handle the data with sentence length varying in a wide range. Thus, we propose two different strategies described below.

5.1. **Improved Algorithm 1: Sentence length-informed method.** The ideas behind our improved algorithms are inspired by the problem of the basic Arith-Phrase selection algorithm which is prone to select shorter sentences. In order to alleviate these problems, we intuitively introduce a brevity penalty to penalize these short sentences to reduce the possibility of their being selected. We denote this method as "sentence length-informed (SLI)" algorithm (we use "Arith-Phrase-Penalty" for short in our experiments).

Since the basic Arith-Phrase is used as the typical selection algorithm in our AL framework, we modify it by adding a brevity penalty factor into Equation (5), as rewritten in (7):

$$\phi(s) = \left[\frac{1}{|X_s^m|} \sum_{x \in X_s^m} \frac{P(x|U)}{P(x|L)} \right] \times BP \tag{7}$$

where BP is the brevity penalty as defined in (8):

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-\frac{r}{c}} & \text{if } c \le r \end{cases} \tag{8}$$

where r is the average sentence length of the monolingual corpus U at each iteration, and c is the length of the sentence to be selected. r is dynamically updated at each iteration with the change of the monolingual corpus U after the informative sentences are selected.

[2] "Arith-Phrase" is more accurate than "Geom-Phrase" in program implementation due to the problem of multiplication of very small probabilities, so we mainly use it as the basic algorithm in our experiments.

5.2. **Improved Algorithm 2: Weighted sentence length-informed method.** This algorithm is proposed due to the findings that our SLI algorithm penalizes over-harshly the short sentences which will be discussed in Section 6. Therefore, we propose a "weighted sentence length-informed (WSLI)" algorithm (we use "Arith-Penalty-Weight" for short) to better balance the sentence length and the human cost, as in (9):

$$\phi(s) = \left[\frac{1}{|X_s^m|} \sum_{x \in X_s^m} \frac{P(x|U)}{P(x|L)} \right] \times WBP \tag{9}$$

where WBP is the weighted brevity penalty as defined in (10):

$$WBP = \begin{cases} 1 & \text{if } \omega \cdot c > r \\ e^{1-\frac{r}{\omega \cdot c}} & \text{if } \omega \cdot c \leq r \end{cases} \tag{10}$$

where r is the average sentence length of the monolingual corpus U, c is length of the sentence to be selected, and ω is the weight to control the penalty factor to balance the sentence length and human cost of translation. ω is set empirically as follows.

- The purpose of introducing the weight ω is to select informative sentences that are expected to be no longer than those selected by the Random method.
- We use the ratio between the two overall average sentence lengths of sentences selected by Random and Arith-Phrase, respectively, to decide the weight ω. Empirically, we set ω to 1.5 in our experiments.

6. Experiments on Improved AL Framework.

6.1. **Experiment setup.** The language pairs in our experiments are Chinese-English, French-English and English-Spanish. For Chinese-English and French-English pairs, the initial parallel data and monolingual data are randomly selected respectively from the NIST Chinese-English FBIS corpus and WMT News Commentary corpus, where the parallel data contains 5k pairs and the monolingual data includes 20k sentences.

The development set for the Chinese-English task is the NIST 2006 current set (1,664 sentences with four references for each source sentence), and the test sets are the NIST 2005 current set (1,082 sentences with four references for each source sentence) and 2008 current set (1,357 sentences with four references for each source sentence). The development set for the French-English task is the WMT Newstest 2013 (3,000 sentences with one reference for each source sentence), and the test set is the WMT Newstest 2014 (3,003 sentences with one reference for each source sentence).

For the English-Spanish pair, we use the data from the WMT Automatic Post-Editing (APE) Shared Task. The parallel data are edited by humans from the translations of an unknown MT system, so it can reflect human cost to some extent. The APE task only contains 11,272 parallel sentences as the training data, and 1,000 parallel sentences as the development set. In order to adapt these data to our task, we extract 1,272 parallel sentences as the initial parallel corpus L, and the remainder as the monolingual corpus U. We split the original development set into two parts (500 sentences for each part), where the first half is used as the "devset" and the other as the "testset" in our experiments.

We utilize Moses (Koehn et al., [11]) to indirectly evaluate the performance of sentence selection algorithms in terms of BLEU score [12]. The language model is 5-gram built from the target side of the bilingual corpus.

We use the same system set-up as in Haffari et al. [5], i.e., the iterations in the AL framework is set to 25, and at each iteration, 200 informative sentences are chosen from the corpus U; the smoothing factor ϵ in Equation (3) and Equation (4) is set to 0.5.

6.2. **Experimental results.** In our previous experiments of comparing the Arith-Phrase method with the Random method, we found that Arith-Phrase performs worse than Random. Therefore, we set the Random method as the baseline to compare with the proposed two algorithms, namely the Arith-Phrase-Penalty and the Arith-Penalty-Weight, in terms of BLEU score. The comparison curves for all iterations and the BLEU scores of the last iteration for all tasks are shown in Figure 2 and Table 1, respectively. In Table 1, all the significance tests use bootstrap method [15].[3]

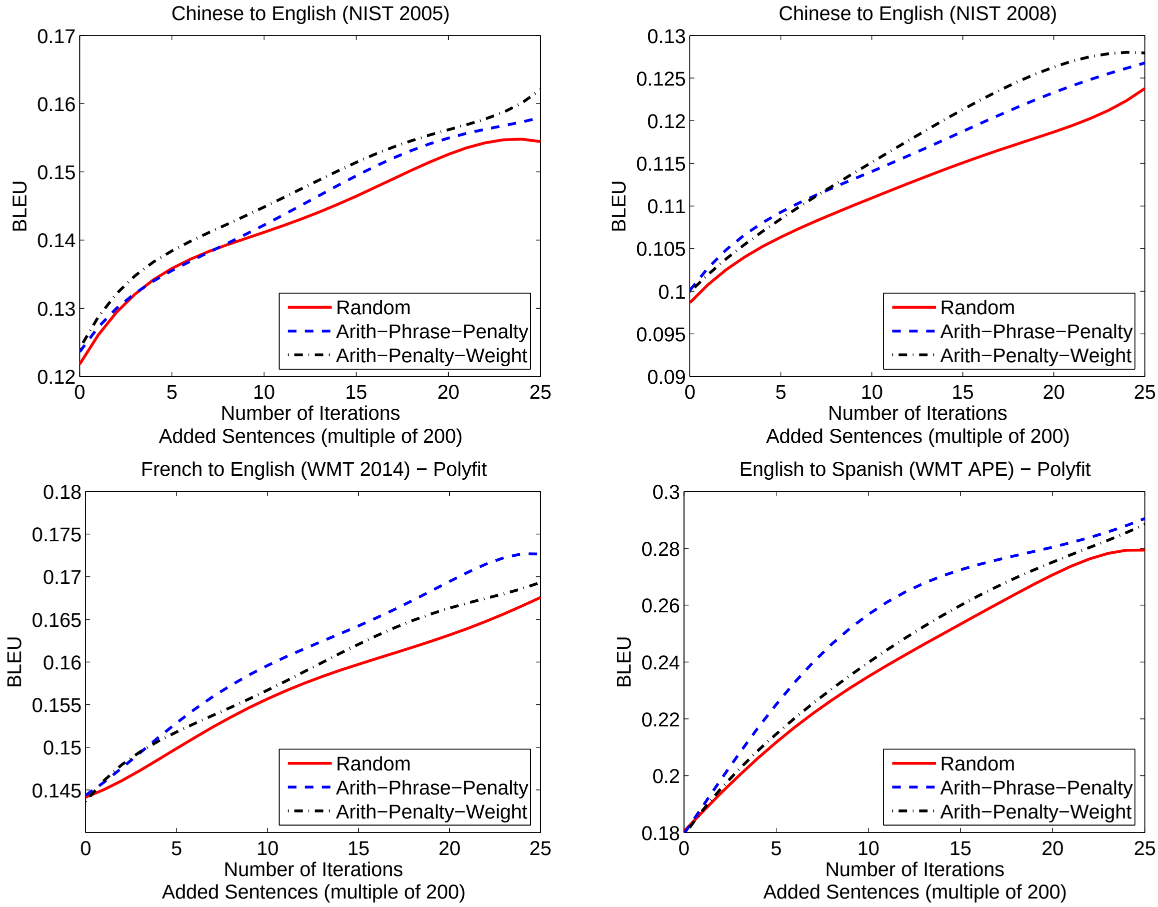

FIGURE 2. Experimental results of sentence length informed (Arith-Phrase-Penalty) and weighted sentence length informed (Arith-Penalty-Weight) methods compared to the baseline (Random)

TABLE 1. Results of three methods at the last iteration

Method/Task	ZH-EN NIST2005	ZH-EN NIST2008	FR-EN WMT	EN-ES APE
Random	15.44	12.47	16.68	27.93
Arith-Phrase-Penalty	15.73	12.64	**17.31**	**29.21**
Arith-Penalty-Weight	**16.01**	**12.81**	17.01	29.00

In Figure 2, the first figure shows the BLEU scores of the Random, Arith-Phrase-Penalty and Arith-Penalty-Weight methods on the Chinese-English NIST 2005 test set. The second shows the results of the Chinese-English NIST 2008 test set, the third shows the results of the French-English WMT test set, and the fourth shows the results of the

[3] http://projectile.sv.cmu.edu/research/public/tools/bootStrap/tutorial.htm.

English-Spanish APE data. All the curves come from the polynomial fitting of real BLEU scores at each iteration.

From Figure 2 and Table 1, we can see that in all tasks, our two improved informative sentence selection algorithms outperform the Random method in terms of BLEU score, which demonstrates the effectiveness of the proposed algorithms. In addition, the Arith-Penalty-Weight performs better than the Arith-Phrase-Penalty on the Chinese-English task and a little of worse than the Arith-Phrase-Penalty on the French-English and English-Spanish tasks in terms of BLEU score. We hypothesize that this might be the case as currently ω is a fixed weight and is not discriminatively optimized in our experiments, so it is not optimal for all three language pairs. This will be part of our future work.

7. **Conclusion and Future Work.** This paper studies the active learning framework and different information-rich sentence selection algorithms for resource-poor SMT. Based on previous work on the Arith-Phrase method in which the sentence length is an important factor to affect system performance when the length of sentences in the monolingual corpus varies over a wide range. Therefore, we proposed two improved algorithms, namely the SLI and the WSLI to penalize shorter sentences which often are selected as the most informative sentences. Experimental results on three language pairs demonstrate that the proposed methods significantly outperform the Random method, and are effective for the active learning based resource-poor SMT.

In future work, we intend to carry out further studies on the AL framework regarding 1) presenting improved sentence selection algorithms that contain rich knowledge to better quantize the information in a sentence; 2) proposing novel solutions that can better balance the tradeoff between **exploration** and **exploitation**, and further decrease the human cost in the AL framework; 3) having successfully demonstrated the effectiveness of our two new algorithms on somewhat simulated resource-poor tasks, we will apply them to actual resource-poor scenarios.

Acknowledgment. This work is partially supported by the Shaanxi Science and Technology Plan (No. 2015JM6328, No. 2016KJXX-79 and No. 2013KCT-04). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] C. Callison-Burch, P. Koehn and M. Osborne, Improved statistical machine translation using paraphrases, *Proc. of NAACL-06*, pp.17-24, 2006.

[2] P. Nakov, Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing, *Proc. of WMT-08*, pp.147-150, 2008.

[3] J. Du, J. Jiang and A. Way, Facilitating translation using source language paraphrase lattices, *Proc. of EMNLP-10*, pp.420-429, 2010.

[4] P. Nakov and H. Ng, Improved statistical machine translation for resource-poor languages using related resource-rich languages, *Proc. of EMNLP-09*, pp.1358-1367, 2009.

[5] G. Haffari, M. Roy and A. Sarkar, Active learning for statistical phrase-based machine translation, *Proc. of NAACL-09*, pp.415-423, 2009.

[6] G. Haffari and A. Sarkar, Active learning for multilingual statistical machine translation, *Proc. of ACL and the 4th IJCNLP*, pp.181-189, 2009.

[7] V. Ambati, S. Vogel and J. Carbonell, Active learning and crowd-sourcing for machine translation, *Proc. of LREC-10*, pp.2169-2174, 2010.

[8] V. Ambati, S. Vogel and J. Carbonell, Multi-strategy approaches to active learning for SMT, *Proc. of MT Summit XIII*, pp.122-129, 2011.

[9] V. Ambati, S. Hewavitharana, S. Vogel and J. Carbonell, Active learning with multiple annotations for comparable data classification task, *Proc. of the 4th Workshop on Building and Using Comparable Corpora*, pp.69-77, 2011.

[10] S. Bakhshaei and S. Khadivi, A pool-based active learning method for improving Farsi-English MT system, *Proc. of IST*, pp.822-826, 2012.

[11] P. Koehn, H. Hoang, C. Callison-Burch et al., Moses: Open source toolkit for statistical machine translation, *Proc. of ACL*, pp.177-180, 2007.

[12] K. Papineni, S. Roukos, T. Ward and W. J. Zhu, BLEU: A method for automatic evaluation of machine translation, *Proc. of ACL*, pp.311-318, 2002.

[13] J. Du and M. Zhang, Findings and considerations in active learning based framework for resource-poor SMT, *Proc. of IALP*, pp.95-98, 2013.

[14] J. Du, M. Wang and M. Zhang, Sentence-length informed method for active learning based resource-poor statistical machine translation, *Communications in Computer and Information Science*, pp.91-102, 2014.

[15] Y. Zhang and S. Vogel, Measuring confidence intervals for the machine translation evaluation metrics, *Proc. of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pp.85-94, 2004.

[16] P. Nakov and H. T. Ng, Improving statistical machine translation for a resource-poor language using related resource-rich languages, *Journal of Artificial Intelligence Research*, vol.44, pp.179-222, 2012.

[17] M. R. Costa-jussa, Q. C. A. Henriquez and R. E. Banchs, Evaluating indirect strategies for Chinese-Spanish statistical machine translation, *Journal of Artificial Intelligence Research*, vol.45, pp.761-780, 2012.