Sentence-Length Informed Method for Active Learning Based Resource-Poor Statistical Machine Translation

Jinhua Du[1,2], Miaomiao Wang[1,2], and Meng Zhang[1]

[1] School of Automation and Information Engineering, Xi'an University of Technology
[2] Shaanxi Key Laboratory of Complex System Control and Intelligent Information Processing, Xi'an, 710048 China
jhdu@xaut.edu.cn

Abstract. This paper presents a simple but effective sentence-length informed method to select informative sentences for active learning (AL) based SMT. A length factor is introduced to penalize short sentences to balance the "**exploration**" and "**exploitation**" problem. The penalty is dynamically updated at each iteration of sentence selection by the ratio of the current candidate sentence length and the overall average sentence length of the monolingual corpus. Experimental results on NIST Chinese–English pair and WMT French-English pair show that the proposed sentence-length penalty based method performs best compared with the typical selection method and random selection strategy.

Keywords: active learning, SMT, sentence length penalty.

1 Introduction

The statistical-based or corpus-based machine translation (SMT) is intrinsically a kind of data-driven method, thus, the scale and quality of the parallel data are crucial for obtaining a good translation performance, especially the large-scale high quality parallel data.

However, it is not the case for many resource-poor language pairs. A number of methods have been presented to alleviate this problem, such as paraphrasing [1–3], related rich resources [4], etc. Considering the reality that a large scale of monolingual data can be easily acquired from the Web, digital media etc., active learning framework for SMT has been proposed to facilitate the shortage issue of parallel data [5–10]. Thus, less human cost could bring a significant improvement to the translation performance of resource limited language pairs.

The key issue in AL strategy is to choose rich-information sentences. As to the SMT, the basic idea of selecting sentences with high information is to find some sentences at each iteration to make the improvement of translation quality maximum [5]. In doing so, the sentences selected are of rich information. Intuitively, if more phrases or words in a sentence occur in the unlabeled (monolingual) data, then it might be more informative because it introduces more new

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 91–102, 2014.

knowledge [5]. In this paper, we presents a sentence length informed method to alleviate the tendency of choosing shorter sentences if the unlabeled data has a large variation in sentence length.

Using the state-of-the-art translation units based method and random selection method as baselines, our proposed method shows significant improvement in terms of translation quality compared with baselines on NIST Chinese-English pair and WMT French-English pair.

2 Related Work

In 2009, Haffari et al. (2009) firstly proposed a practical active learning framework for SMT where a number of high-quality parallel data are acquired from the large-scale monolingual data [5, 6]. 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, inverse model etc.

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 cost of label acquisition by prioritizing the most informative data for annotation, while crowd-sourcing reduces cost by using the power of the crowds to meet up the lack of expensive language experts. Their experiments showed significant improvements in translation quality even with less data [7, 9].

In 2012, Bakhshaei and Khadivi applied a pool-based AL strategy to improve Farsi-English SMT system. They increased n in the n-gram feature from 4 to 5, and verified that the sentence selection algorithms such as translation units, translation confidence, inverse model etc. perform better than the random selection method in the task of Farsi-English translation [10].

On the basis of previous work, this paper introduced a length penalty factor into the phrase-based sentence selection strategy to penalize the short sentences. The penalty is dynamically updated at each iteration of sentence selection by the ratio of the current candidate sentence length and the overall average sentence length of the monolingual corpus.

3 Active Learning Framework for SMT

The AL framework for SMT is to obtain parallel data from the large-scale monolingual corpus and add 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 to human translators.

Generally, the active learning framework has two prerequisites: (1) small-scale initial parallel corpus used to build a baseline SMT system; (2) large-scale monolingual corpus to acquire extra bilingual data. More importantly, there are two key issues in the AL strategy: (1) how to design an efficient algorithm to

evaluate the information that a sentence contains and select the rich-information sentences; (2) how to utilize the new parallel data to train and update the SMT system.

In our work, we mainly studied the first question, i.e., using translation units based methods, especially the phrase-based ones, on Chinese-English and French-English language pairs.

As to the second issue, different from the method used in [6], we only use the L trained model to run our SMT system at each iteration. 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: $M_{F\to E} = \textbf{train}(L)$
3: **for** $t = 1, 2, \ldots, N$ **do**
4: Generate "Phrase Set" and compute sentence scores
5: Select k sentences from U, and ask human experts for true translations.
6: Remove the k sentences from U, and add the k sentence pairs to L.
7: Update $M_{F\to E} = \textbf{train}(L)$
8: Evaluate the system performance on the test set.
9: **end for**

4 Sentence-Length Informed Selection Strategy

The general mathematical description of a sentence selection algorithm is that 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 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 most important in a selection algorithm. We can see that this process can be defined as a *quadruple* in (1),

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

4.1 Geom-Phrase and Arith-Phrase Algorithms

In these two methods, the basic unit for computing scores of a sentence is phrase. The Geom-Phrase algorithm is as in (2),

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

where X_s^m is the set of possible phrases that the sentence s can offer, $P(x|U)$ and $P(x|L)$ is 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 smooth factor [1]. X_U^m indicates the set of phrases that indeed occur in U, and X_L^m represents the set of phrases that truely appear in L.

The Arith-Phrase method is defined as in (5):

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

In [6], the phrases in Eq. (2) and Eq. (5) are extracted from the k-best list of translations of a sentence s in U. In addition, the out-of-vocabulary (OOV) words appeared in the translations are also included as candidate phrases with a uniform probability. In order to make the phrase set approximately as a complete set, that is, 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 are not occurred in the phrase table.

4.2 Sentence-Length Informed Algorithm

The idea of presenting the sentence-length informed method is inspired by the findings and analysis in Section 5. The experimental results are not consistent with the conclusions in [5]. Thus, we carried out a comprehensive investigation and analysis, and found that

- the sentences selected by Arith-Phrase algorithm are generally shorter than the random selection;
- the sentence length varies in a wide range in corpus U (1$\sim$ 100 words in a sentence in our experiments).

We consider that the sentence length might have a significant impact on the selection performance. Therefore, we introduce a brevity penalty to prevent very short sentences as in [12]. The modified Arith-Phrase algorithm which we call it "Arith-Phrase-Penalty" is as in Eq. (6),

$$\phi(s) = [\frac{1}{|X_s^m|} \sum_{x \in X_s^m} \frac{P(x|U)}{P(x|L)}] \times BP \tag{6}$$

[1] We set $\epsilon = 0.5$ in the experiments.

where BP is the brevity penalty and defined as follows,

$$BP = \begin{cases} 1 & \text{if} \quad c > r \\ e^{1-\frac{r}{c}} & \text{if} \quad c \leq r \end{cases} \tag{7}$$

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

5 Experiments, Findings and Analysis

5.1 Experiment Setup

The language pairs in our experiments are Chinese–English and French–English. English is the target language both in these two pairs. The initial parallel data and monolingual data are randomly selected respectively from 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 Chinese–English task is NIST 2006 current set (1,664 sentences with four references for each source sentence), the test sets are NIST 2005 current set (1,083 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 French–English task is WMT Newstest 2013 (3,000 sentences with one reference for each source sentence), and the test set is WMT Newstest 2014 (3,003 sentences with one reference for each source sentence).

We utilize Moses [11] to indirectly evaluate the performance of sentence selection algorithms in terms of BLEU scores. The language model is five-gram built on the English part of the bilingual corpus.

As in [5], the iteration times in the AL framework is set to 25, and at each iteration, 200 informative sentences are chosen from the corpus U; the smooth factor ϵ in Eq. (3) and Eq. (4) is set to 0.5.

5.2 Experiments on Arith-Phrase and Random Methods

We set the random selection method as the basic baseline. In this Section, we carried out a comparison experiments between the typical Arith-Phrase algorithm and the baseline on Chinese–English and French–English pairs, and found that the typical Arith-Phrase method did not beat the random selection method as shown shown in Figure 4.

In Figure 4, three top figures demonstrate the true BLEU scores at each iteration of the random, typical Arith-Phrase and our proposed methods for two language pairs, and three bottom figures use the polynomial fitting to demonstrate the trends of these methods so that the differences in terms of BLEU can be obviously observed.

We can see that in our experiments BLEU scores of the typical Arith-Phrase method are significantly lower than the random method.

A comprehensive investigation and analysis were carried out whereafter, and found that the sentences selected by the typical Arith-Phrase algorithm are generally shorter than those of the baseline. The observation drives us to consider the following questions,

- What extent does the sentence length affect the performance of algorithms?
- How could we alleviate the impact of sentence length so that we can find out the true informative sentences?

With these questions, we proceeded a data analysis as described below.

5.3 Data Statistics and Analysis

5.3.1 The Contradiction

Here we define "new words" as occurring in U but not in L before, and "existing words" as appearing both in U and L.

When selecting high-information sentences, there is a pair of contradiction that is **exploration** and **exploitation**, i.e. selecting sentences to discover new phrases vs estimating accurately the phrase translation probabilities [5]. Specifically,

- the more new words a sentence has in terms of the parallel corpus, the more informative the sentence is, but a lower word alignment accuracy to the added parallel data (c.f. Section 5.3.4);
- while the more existing words a sentence has to the parallel data, the more accurate the phrase probability is estimated, but a lower coverage to the test set (or unknown data)(c.f. Section 5.3.3 and Section 5.3.4).

Thus, we need to make a tradeoff between the ratio of new words and existing words when selecting out sentences from the monolingual corpus U.

Accordingly, regarding the negative results, we conducted a data analysis to investigate the hidden reasons from three aspects:

- statistics of the sentence length in different situations;
- the coverage rates of test sets by the parallel data in different situations;
- the increments of existing words and new words in different situations.

5.3.2 Average Sentence Length

Figure 1 illustrates the average sentence length of selected sentences at each iteration of the Random, Arith-Phrase and Arith-Phrase-Penalty methods for two language pairs.

In Fig. 1, "X" in "Random:X", "Arith-Phrase:X" and "Arith-Phrase-Penalty:X" indicates the overall average sentence length of all selected sentences in terms of three different methods; "Average Sentence Length of U" is the average sentence length of the monolingual corpus U at each iteration that is in fact the parameter r in Eq. (7).

It can be seen that

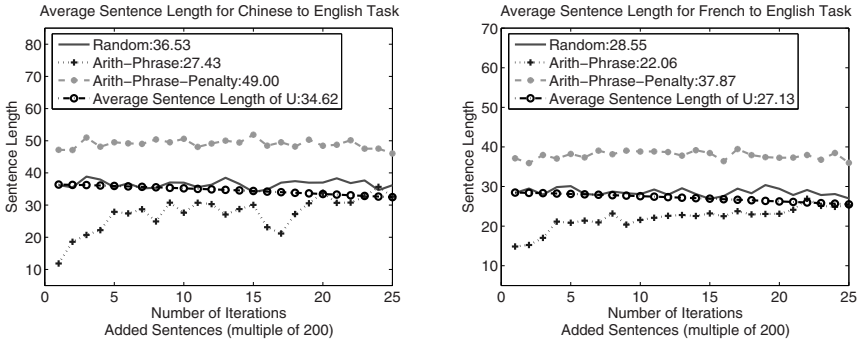

Fig. 1. Statistics of sentence length for different selection methods

- the average length of Arith-Phrase is shortest at each iteration both for two language pairs;
- the variance of the average sentence length of Arith-Phrase is biggest while the average lengths of the other two methods change slightly at each iteration.

The main purpose of active learning based SMT is to find and select the most informative sentences from the monolingual corpus, then the observations above drive us to ask that do the short sentences selected by Arith-Phrase algorithm really contain more information? If so, why does it perform worse than the Random method?

To answer the questions above, we might investigate the deep relationship between "existing words" and "new words" – the contradiction of the active learning SMT.

5.3.3 Coverage of Testsets by Parallel Data

An investigation is carried out to compute the distribution of the coverage rates of the test set by selected sentences at each iteration to see whether the Arith-Phrase method can find out more informative sentences. Results are shown in Figure 2[2].

We can see that the coverage of Arith-Phrase is significantly higher than that of Random. This indicates that Arith-Phrase tends to select sentences containing more "new words", i.e., tends more to the "exploration" side. Intuitively, the increase of new words would improve the translation performance because it is useful to reduce the out-of-vocabulary (OOVs) in the translation hypotheses. However, higher coverage rate did not improve the translation quality!

We analyze that this might be: the Arith-Phrase algorithm is potentially to look for highly informative sentences featured by more new words. Intuitively, the shorter a sentence is, the greater the proportion of news words in this sentence is, the more likely it can be chosen compared to a longer sentence.

[2] Due to the limitation of the paper space, we take ZH-EN NIST 2005 and FR-EN WMT 2014 test sets as examples. It is the same situation for ZH-EN NIST 2008.

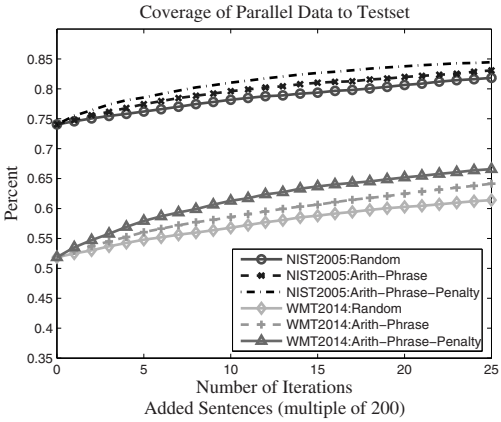

Fig. 2. Comparison of coverage rates of parallel data to different test sets at each iteration

The possible reason is that although Arith-Phrase tries to make a tradeoff between the "**exploitation**" and "**exploration**", it is difficult when the sentence length varies in a wide range of the monolingual corpus. From another viewpoint, the increase of new words implies that the relative decrease of frequencies of existing words that might be more important to improve the accuracy of phrase probability estimation for the resource-limited small-scale SMT system.

Based on the data statistics and analysis of sentence length and coverage, we refer that the variation between the frequencies of existing words and new words might provide a reasonable explanation.

5.3.4 Statistics of Existing Words and New Words

As mentioned that "existing words" appear both in U and L, and "new words" occur only in U, we analyze the relationship between existing words and new words from three aspects: 1) the incremental frequencies of existing words at each iteration; 2) the incremental new words at each iteration; 3) the incremental frequencies of new words at each iteration. Statistics are shown in Fig. 3.

It can be seen in Fig.3 that

- in terms of incremental frequencies of existing words for two language pairs, the Random method is significantly higher than the Arith-Phrase algorithm. This would improve the accuracy of probability estimation of existing words.
- in terms of incremental new words for two language pairs, the Arith-Phrase is higher than the Random method. This is consistent with the coverage comparison in Section 5.3.3.
- in terms of incremental frequencies of new words, the Arith-Phrase is also higher than the Random method. However, we can find that the frequency increments of new words are nearly the same as the increments of new words, i.e., the new words have quite low frequencies. For example, at iteration 5 of

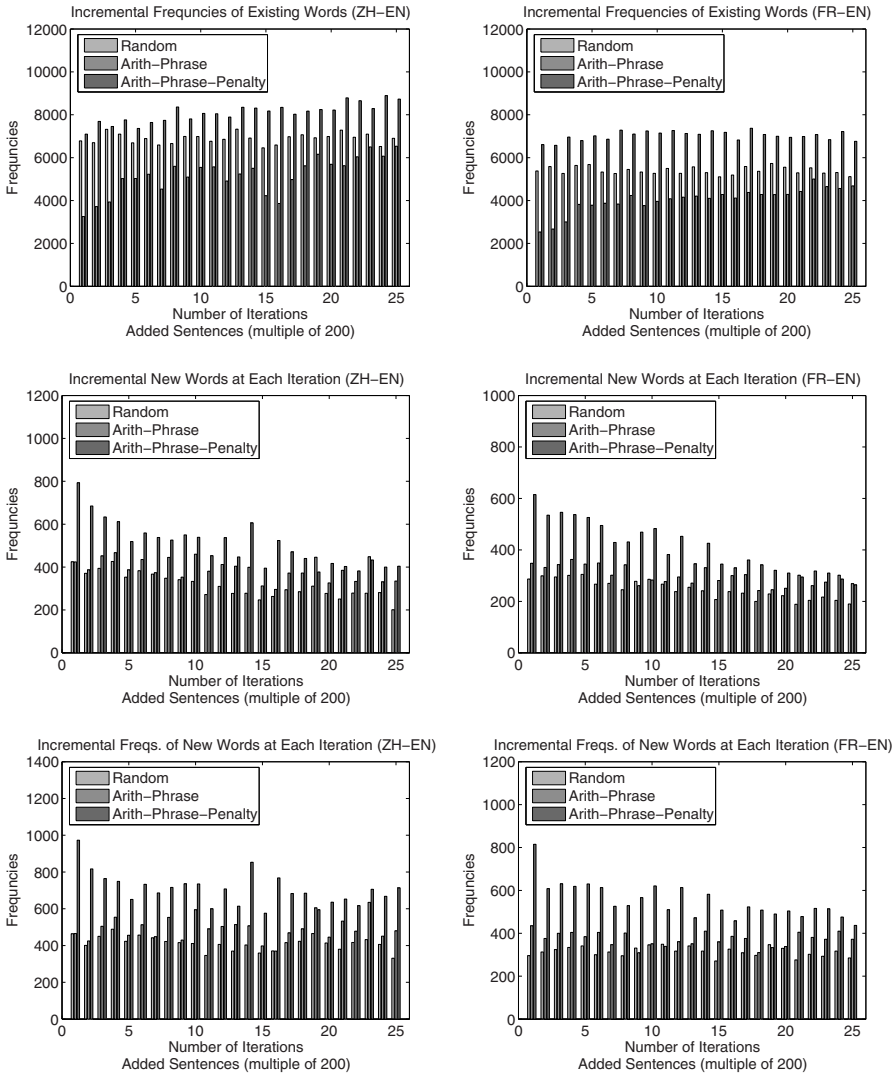

Fig. 3. Statistics of existing words and new words at each iteration

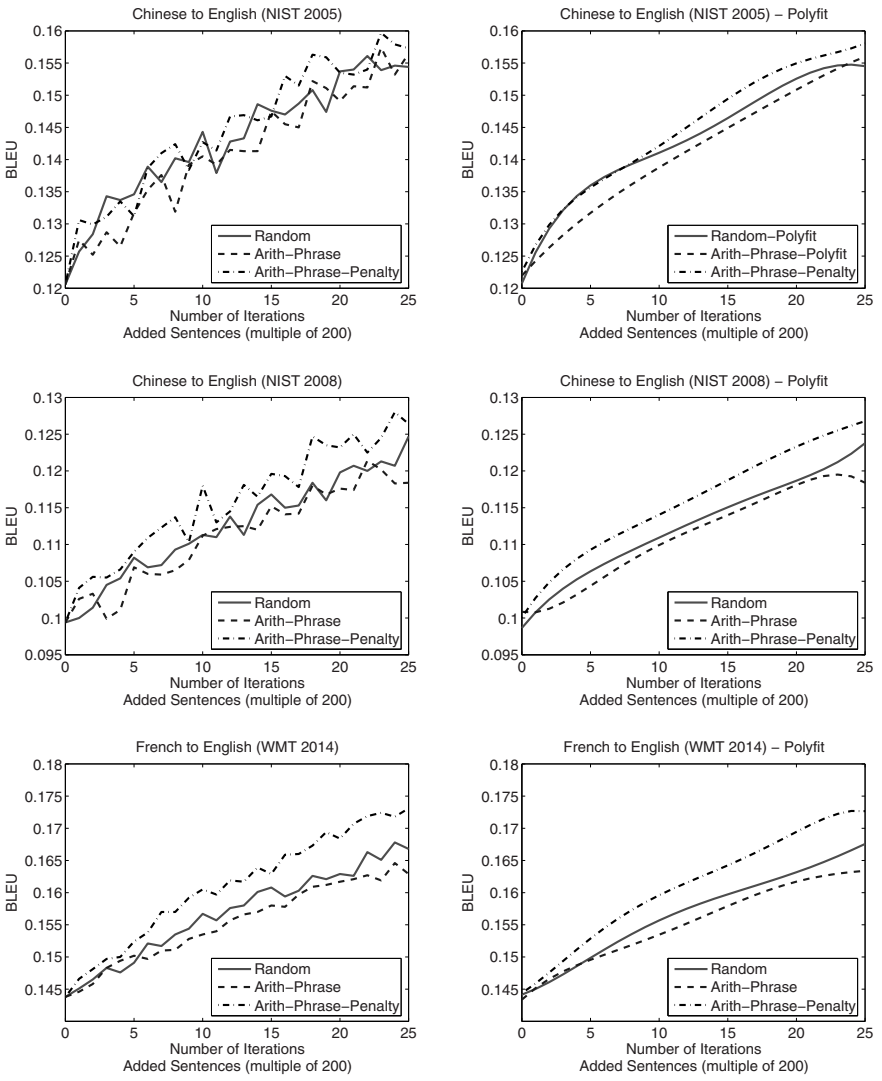

Fig. 4. Experimental results of Arith-Phrase, Sentence-length Informed (Arith-Phrase-Penalty) methods compared to the baseline (Random)

Arith-Phrase for FR-EN task, the number of incremental new words is 345, the number of incremental frequencies is 384, the average frequency for each new word is about 1.11 that is quite low. Thus, this might explain that for Arith-Phrase method, although the coverage increases, the accuracy of the probability estimation of existing words and new words relatively decreases due to low frequencies compared to the Random method.

Based on the analysis, we consider that if a sentence length informed factor can be introduced into the Arith-Phrase algorithm, it might balance the contradiction of existing words and new words. Thus, the brevity penalty based Arith-Phrase algorithm is proposed. The next sections will carry out comparison experiments between our method and baselines.

5.3.5 Experiments on the Proposed Method

It can be found in Figures 1, 2, 3 and 4 that

- the proposed Arith-Phrase-Penalty method significantly outperforms the Random and Arith-Phrase in all tasks in terms of translation performance (BLEU scores)(See Fig. 4);
- the incremental frequencies of existing words of Arith-Phrase-Penalty is far higher than those of Random and Arith-Phrase, which could further improve the probability estimation accuracy of existing words (See Fig. 3);
- the incremental new words of Arith-Phrase-Penalty is higher than Random and Arith-Phrase, which could bring a broader coverage to test sets. Figure 2 shows the consistency;
- the average sentence length of Arith-Phrase-Penalty is larger than Random and Arith-Phrase and the variance is small (See Fig. 1). The longer sentences bring more existing words and introduce new words into the parallel data, which not only increases the coverage to test sets, but also improves the accuracy of probability estimation of phrases. However, the potential problem is that the human cost would be risen.

It can be said that the proposed algorithm indeed improved the performance of typical Arith-Phrase method, and achieved best results.

6 Conclusions

This paper studies the active learning framework and different high-information sentence selection algorithms for resource-poor SMT. Based on the negative experimental results on Arith-Phrase method, we found that the sentence length is an important factor to affect the system performance when the length of sentences in the monolingual corpus varies in a wide range. Based on the analysis, a simple but effective method – sentence length informed Arith-Phrase – is proposed to penalize sentences that are shorter than the overall average length of the monolingual corpus U at each iteration. Experimental results demonstrate

that the proposed method significantly outperforms the typical Arith-Phrase and Random method.

In future, we intend to carry out further study on the AL framework in the respects of 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 decrease the human cost in the AL framework.

Acknowledgments. This work is supported by NSF project (61100085), the Open Projects Program of National Laboratory of Pattern Recognition, and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry. Thanks the reviewers for their insightful comments and suggestions.

References

[1] Callison-Burch, C., Koehn, P., Osborne, M.: Improved statistical machine translation using paraphrases. In: Proceedings of HLT-NAACL 2006: Proceedings of the NAACL, pp. 17–24 (2006)
[2] Nakov, P.: Improving English-Spanish statistical machine translation: experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In: Proceedings of WMT, pp. 147–150 (2008)
[3] Du, J., Jiang, J., Way, A.: Facilitating Translation Using Source Language Paraphrase Lattices. In: Proceedings of EMNLP, pp. 420–429 (2010)
[4] Nakov, P., Ng, H.: Improved statistical machine translation for resource-poor languages using related resource-rich languages. In: Proceedings of EMNLP, pp. 1358–1367 (2009)
[5] Haffari, G., Roy, M., Sarkar, A.: Active learning for statistical phrase-based machine translation. In: Proceedings of NAACL, pp. 415–423 (2009)
[6] Haffari, G., Sarkar, A.: Active Learning for Multilingual Statistical Machine Translation. In: Proceedings of ACL and the 4th IJCNLP, pp. 181–189 (2009)
[7] Ambati, V., Vogel, S., Carbonell, J.: Active learning and crowd-sourcing for machine translation. In: Proceedings of LREC, pp. 2169–2174 (2010)
[8] Ambati, V., Vogel, S., Carbonell, J.: Multi-strategy approaches to active learning for smt. In: Proceedings of the MT Summit XIII, pp. 122–129 (2011)
[9] Ambati, V., Hewavitharana, S., Vogel, S., Carbonell, J.: Active learning with multiple annotations for comparable data classification task. In: Proceedings of the Fourth Workshop on Building and Using Comparable Corpora, pp. 69–77 (2011)
[10] Bakhshaei, S., Khadivi, S.: A Pool-based Active Learning Method for Improving Farsi-English MT system. In: Proceedings of IST, pp. 822–826 (2012)
[11] Koehn, P., Hoang, H., Callison-Burch, C., et al.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of ACL, pp. 177–180 (2007)
[12] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of ACL, pp. 311–318 (2002)