Edit Distance: A New Data Selection Criterion for Domain Adaptation in SMT

Longyue Wang[1], Derek F. Wong[1],
Lidia S. Chao[1], Junwen Xing[1], Yi Lu[1], Isabel Trancoso[2]
[1]Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory, University of Macau, Macau S.A.R., China
[2]L2F Spoken Language Systems Lab, INESC-ID, Lisboa, Portugal
{mb15505,derekfw,lidiasc,mb15470,mb25435}@umac.mo,
isabel.trancoso@inesc-id.pt

Abstract

This paper aims at effective use of training data by extracting sentences from large general-domain corpora to adapt statistical machine translation systems to domain-specific data. We regard this task as a problem of filtering training sentences with respect to the target domain[1] via different similarity metrics. Thus, we give new insights into when data selection model can best benefit the in-domain translation. Based on the investigation of the state-of-the-art similarity metrics, we propose edit distance as a new data selection criterion for this topic. To evaluate this proposal, we compare it with other methods on a large dataset. Comparative experiments are conducted on Chinese-English travel dialog domain and the results indicate that the proposed approach achieves a significant improvement over the baseline system (+4.36 BLEU) as well as the best rival model (+1.23 BLEU) using a much smaller training subset. This study may have a significant impact on mining very large corpora in a computationally-limited environment.

1 Introduction

A well-known problem of statistical machine translation (SMT) (Brown et al., 1993) is that the data-driven system is not guaranteed to perform optimally if the data for training and testing are not identically distributed. Domain adaptation for SMT has been explored at different component levels: word level, phrase level, sentence level and model level. For example mining unknown words from comparable corpora (Daume III and Jagarlamudi, 2011), weighted phrase extraction (Mansour and Ney, 2012), mixing multiple models (Civera and Juan, 2007; Foster and Kuhn, 2007; Eidelman et al., 2012), etc. Recently, data selection as a simple and effective way for this special task has attracted attention.

Under the assumption that there exists a large general-domain corpus (general corpus) including sufficient domains, the task of data selection is to translate a domain-specific text using the optimized translation model (TM) or language model (LM) trained by less but more suitable data retrieved from the general corpus. To state it formally, R is an abstract model of target domain and s_G is a sentence or a sentences pair in the general corpus G. The score of each s_G is given by

$$Score(s_G) \rightarrow Sim(s_G, R) \quad (1)$$

which means if we could find a better function to measure the similarity between s_G and R, G could be replaced by a new sub-corpus G_{sub} for training a domain-specific SMT system.

We focus on two data selection criteria that have been explored for domain adaptation. One comes from the realm of information retrieval (IR), which is defined as the cosine of the angle between two vectors based on term frequency-inverse document frequency (TF-IDF). Hildebrand et al. (2005) showed that it is possible to apply this standard IR technique for both TM adaptation and LM adaptation. It is also similar to the offline data optimization approach proposed by Lü et al. (2007), who re-sample and re-

[1] It could be modeled by an in-domain corpus or text to be translated.

Proceedings of Recent Advances in Natural Language Processing, pages 727–732,
Hissar, Bulgaria, 7-13 September 2013.

weight sentences in general corpus, achieving an improvement of about 1 BLEU point over the baseline system. This simple co-occurrence based matching only considers keywords overlap, which may result in weakness in filtering irrelevant data. Thus, it needs a large size of the selected subset (more than 50% of general corpus) to obtain an ideal performance. The other data selection criterion is a perplexity-based model which can be found in the field of language modeling. This has been explored by Gao et al. (2002) and more recently by Moore and Lewis (2010), who used cross-entropy to score text segments according to an additional in-domain LM. Axelrod et al. (2011) employed these perplexity-based variants for SMT adaptation and showed that the fast and simple technique allows to discard over 99% of the general corpus resulting in an increase of 1.8 BLEU points. By considering not only the distribution of terms but also the collocation, perplexity-based metrics perform better than the IR techniques in general.

We show that constraint factors in similarity measuring such as word overlap and word order may have a major impact on the quality of selected data as well as the translation quality. The stricter selection criteria may have stronger ability in filtering noises, resulting in a better domain-specific translation. Edit distance is much stricter than the former two criteria. The factors of words overlap, order and position are all comprehensively considered. This distance able to retrieve more similar sentences from the general corpus. Actually, edit distance has been widely used for example-based MT (EBMT) (Leveling et al., 2012) and convergence of translation memory (TM) and SMT (Koehn and Senellart, 2010), but it was not previously applied to this topic. This proposal is under the assumption that the general corpus is large and broad enough to cover highly similar sentences with respect to the target domain. We compared it with the baseline and other two state-of-the-art methods on a large Chinese-English general corpus. Using BLEU (Papineni et al., 2002) as an evaluation metric, we obtained a significant improvements over the baseline system and the best of other methods.

This paper is organized as follows. Section 2 describes the related models for data selection. The resources and configurations of experiments for are detailed in Section 3. Finally, we compare and discuss the results in Section 4 followed by a conclusion to end the paper.

2 Model Description

This section will briefly describe the three data selection models to be considered: standard IR model, perplexity based model and the proposed model.

2.1 IR Model

Each document D_i is represented as a vector (w_{i1}, w_{i2},…, w_{in}), and n is the size of the vocabulary. So w_{ij} is calculated as follows:

$$w_{ij} = tf_{ij} \times \log(idf_j) \quad (2)$$

where tf_{ij} is term frequency (TF) of the j-th word in the vocabulary in the document D_i, and idf_j is the inverse document frequency (IDF) of the j-th word calculated. The similarity between two documents is then defined as the cosine of the angle between two vectors.

In practice, we only use the sentences in the source language for indexing and query generation. Each sentence in the general corpus is indexed as one document by Apache Lucene[2]. Every sentence without the stop words from the reference set is used as one separate query. As in (Hildebrand et al. 2005), we allow duplicated sentences during the selection which is similar with. All retrieved sentences with their corresponding target translations are ranked according to their similarity scores.

2.2 Perplexity-Based Model

The perplexity of a string s with empirical n-gram distribution p given a language model q is:

$$2^{-\sum_x p(x)\log q(x)} = 2^{H(p,q)} \quad (3)$$

in which $H(p, q)$ is the cross-entropy between p and q. Selecting segments based on a perplexity threshold is equivalent to selecting based on a cross-entropy threshold, which is more often used for this task (Moore and Lewis, 2010; Axelrod et al., 2011). Supposed that $H_I(s)$ and $H_O(s)$ are the cross-entropy of a string s according to an in-domain language model LM_I and non-in-domain LM_G respectively trained on in-domain data set I and a partition of general-domain data set G. Considering both source (*src*) and target (*tar*) side of parallel training data, there are three variants. The first is basic cross-entropy given by:

$$H_{I-src}(s) \quad (4)$$

[2] Available at http://lucene.apache.org.

and the second is cross-entropy difference (Moore and Lewis, 2010):

$$H_{I-src}(s) - H_{G-src}(s) \quad (5)$$

which tries to select the sentences that are more similar to the target domain but different to others in general corpus. The third one is to sum the cross-entropy difference over both source and target side of the corpus:

$$\begin{aligned} &\left[H_{I-src}(s) - H_{G-src}(s)\right] \\ &+\left[H_{I-tar}(s) - H_{G-tar}(s)\right] \end{aligned} \quad (6)$$

The third variant has been is proven to achieve the best result among the three cross-entropy variants (Axelrod et al., 2011).

2.3 Edit-Distance-Based Model

Given a sentence s_G from a general corpus and a sentence s_R from the test set or in-domain corpus, the edit distance for these two sequences is defined as the minimum number of edits, i.e. symbol insertions, deletions and substitutions, for transforming s_G into s_R. There are several different implementations of the edit-distance-based retrieval model. We used the normalized Levenshtein similarity score (fuzzy matching score, FMS) proposed by Koehn and Senellart (2010):

$$FMS = 1 - \frac{LED_{word}(s_G, s_R)}{Max(|s_G|, |s_R|)} \quad (7)$$

in which LED_{word} is a distance function and $|s|$ is the number of tokens of sentence s. In this study, we employed a word-based Levenshtein edit distance function instead of additionally using a letter-based one. If the score of a sentence exceeds a threshold, we will further penalize it according to space and punctuations edit differences.

3 Experimental Setup

3.1 Corpora

Two corpora are needed for the domain adaptation task. Our general corpus includes 5 million English-Chinese parallel sentences comprising various genres such as movie subtitle, law literature, news and novel. The in-domain corpus and test set are randomly selected from the IWSLT2010 (International Workshop on Spoken Language Translation) Chinese-English Dialog task[3], consisting of transcriptions of conversational speech in a travel setting. All of them were segmented [4] (Zhang, 2003) and tokenized [5] (Koehn, 2005). The sizes of the test set, in-domain corpus and general corpus we used are summarized in Table 1.

Data Set	Sentences	Tokens	Ave. Len.
Test Set	3,500	34,382	9.60
In-domain	17,975	151,797	9.45
Training Set	5,211,281	53,650,998	12.93

Table 1: Corpora statistics.

In practice, we followed the experiments conducted by L ü et al. (2007) and Hildebrand et al. (2005), where the test set was used to select in-domain data from general corpus. The only difference is that an additional in-domain corpus is employed to build the LM for perplexity-based retrieval (Moore and Lewis, 2010; Axelrod et al., 2011).

3.2 System Description

The experiments presented in this paper are carried out with the Moses toolkit (Koehn et al., 2007), a state-of-the-art open-source phrase-based SMT system. The translation and the reordering model relied on "*grow-diag-final*" symmetrized word-to-word alignments built using GIZA++ (Och and Ney, 2003) and the training script of Moses. A 5-gram language model was trained on the target side of the training parallel corpus using the IRSTLM toolkit (Federico et al., 2008), exploiting improved Modified Kneser-Ney smoothing, and quantizing both probabilities and back-off weights.

3.3 Baseline System

The baseline system was trained on the general corpus with toolkits and settings as described above. The baseline BLEU is **29.34** points. This low value is occurred by the fact that he general corpus does not consist of enough sentences on the travel domain and has a lot of out-of-domain data, which can be regarded as noise for this task.

4 Results and Discussions

A number of experiments have been conducted to investigate five data selection methods: standard IR (IR), source-side cross-entropy (CE),

[3] http://iwslt2010.fbk.eu/node/33.

[4] IC-TCLAS2013 is available at http://ictclas.nlpir.org/.

[5] Scripts are available at http://www.statmt.org/europarl/.

source-side cross-entropy difference (CED), bilingual cross-entropy difference (B-CED) and the fuzzy matching (FMS_{ours}) methods. Supposed that M is the size of the test set or in-domain corpus and N is the number of sentences retrieved from the general corpus according to each query. Thus, the size of the subset we selected is $M \times N$.

We investigate each method in a step of 2x starting from 0.25% of the general corpus (0.29%, 0.52%, 1.00%, 2.30%, 4.25% and 12.5%) where K% means K percentage of general corpus are selected as a subset.

Firstly, we evaluated IR which improves by at most 1.03 BLEU points when using 4.25% of the general corpus as shown in Fig. 1. Then the performance begins to drop when the size is more than 4.25%. This shows that keyword overlap plays a significant role in retrieving sentences in a similar domain. However, it still needs a large amount of selected data to obtain an ideal performance due to its weakness in filtering noise.

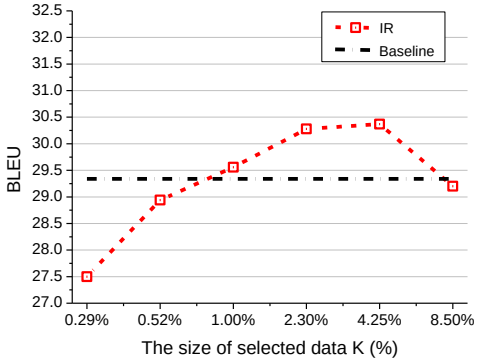

Figure 1: Translation results using subset of general corpus selected by standard IR model.

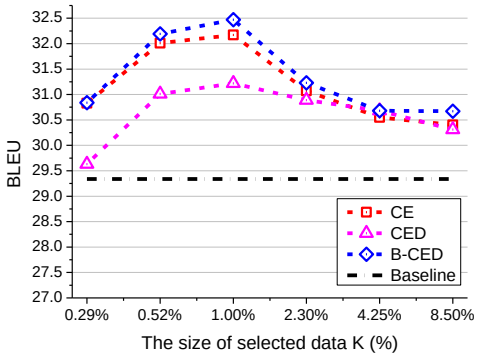

Figure 2: Translation results using subset of general corpus selected by three perplexity-based variants.

Secondly, we compared three perplexity-based methods. As illustrated in Fig. 2, all of them were able to significantly outperform the baseline system using only 1% of the entire training data. The size threshold is much smaller than the one of IR when obtaining the equivalent performance. Moreover, the curve drops slowly and is always over the baseline. This shows a better ability of filtering noises. Among the perplexity-based variants, the B-CED works best, which is similar to the conclusion drawn by Axelrod et al. (2011). It proves that bilingual resources are helpful to balance OOVs and noises. Next we will use B-CED to stand for perplexity-based methods and compare with other selection criteria.

Finally, we evaluated FMS and compared it with IR, B-CED and the baseline system, which are shown in Fig. 3. FMS seems to give an outstanding performance on most size thresholds. It always outperforms B-CED over at least 1 point under the same settings. Even using only 0.29% data, the BLEU is still higher than baseline over 0.66 points. In addition, FMS is able to conduct a better in-domain SMT system using less data than other selection methods. This indicates that it is stronger to filter noises and keep in-domain data when considering more constrain factors for similarity measuring.

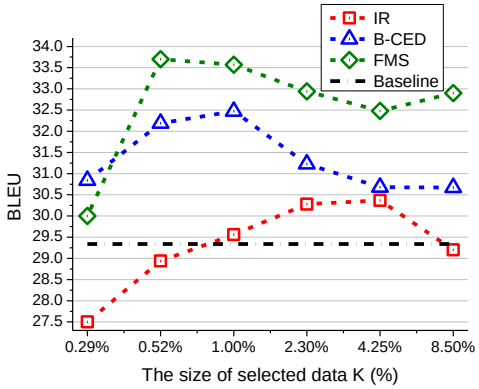

Figure 3: Translation results using subset of general corpus selected by different methods.

Corpus	Size (%)	BLEU
Baseline	100	29.34
IR	4.25	30.37 (+1.03)
CE	1.00	32.17 (+2.83)
CED	1.00	31.22 (+1.88)
B-CED	1.00	32.47 (+3.13)
FMS_{ours}	**0.52**	33.70 (+**4.36**)

Table 2: Best result of each method with corresponding size of selected data.

To give a better numerical comparison, Table 2 lists the best result of each method. As expected, FMS could use the smallest data (0.52%) to achieve the best performance. It outperforms the baseline system trained on the entire dataset

over 4.36 BLEU points and B-CED over 1.23 points.

5 Conclusions

In this paper, we regard data selection as a problem of scoring the sentences in a general corpus via different similarity metrics. After revisiting the state-of-the-art data selection methods for SMT adaptation, we propose edit distance as a new selection criterion for this topic. In order to evaluate the proposed method, we compare it with four other related methods on a large data set. The methods we implemented are standard information retrieval model, source-side cross-entropy, source-side cross-entropy difference, bilingual cross-entropy difference as well as a baseline system. We can analyze the results from two different aspects:

Translation Quality: The results show a significant performance of the proposed method with increasing 4.36 BLEU points than the baseline system. And it also outperforms other four methods over 1-3 points.

Filtering Noises: Fuzzy matching could discard about 99.5% data of the general corpus without reducing translation quality. However, other methods will drop their performance when using the same size of data. The proposed metric has a very strong ability to filter noises in general corpus.

Finally, we can draw a composite conclusion that edit distance is a more suitable similarly model for SMT domain adaptation.

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 the anonymous reviewers for many helpful comments.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In: *Proceedings of EMNLP*. pp. 355–362.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*. 19:263–311.

Jorge Civera and Alfons Juan. 2007. Domain adaptation in statistical machine translation with mixture modelling. *Proceedings of the Second ACL Workshop on Statistical Machine Translation*. pp. 177–180.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In: *Proceedings of ACL-HLT*. pp. 407–412.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. *Proceedings of ACL*: *Short Papers*. Vol. 2. pp. 115–119.

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

G. Foster and R. Kuhn. 2007. Mixture-model adaptation for SMT. *Proceedings of the Second ACL Workshop on Statistical Machine Translation*. pp. 128 – 136.

Jianfeng Gao, Joshua Goodman, Mingjing Li, and Kai-Fu Lee. 2002. Toward a unified approach to statistical language modeling for Chinese. *ACM Transactions on Asian Language Information Processing* (TALIP). 1:3–33.

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

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit*. Vol. 5. pp. 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran et al. 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of ACL*. pp. 177–180.

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

Yajuan Lü, Jin Huang, and Qun Liu. 2007. Improving statistical machine translation performance by

training data selection and optimization. *Proceedings of EMNLP-CoNLL*. pp. 343–350.

Saab Mansour and Hermann Ney. 2012. A Simple and Effective Weighted Phrase Extraction for Machine Translation Adaptation. *Proceedings of IWSLT*. pp. 193-200.

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

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*. 29:19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of ACL*. pp. 311–318.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. *Proceedings of the 2nd SIGHAN Workshop on Chinese Language processing*. Vol. 17. pp. 184–187.