Pseudo In-Domain Data Selection from Large-Scale Web Corpus for Spoken Language Translation

Shixiang Lu, Xingyuan Peng, Zhenbiao Chen, and Bo Xu

Interactive Digital Media Technology Research Center (IDMTech),
Institute of Automation, Chinese Academy of Sciences, Beijing, China
{shixiang.lu,xingyuan.peng,zhenbiao.chen,xubo}@ia.ac.cn

Abstract. This paper is concerned with exploring efficient domain adaptation for the task of statistical machine translation, which is based on extracting sentence pairs (*pseudo in-domain subcorpora*, that are most relevant to the in domain corpora) from a large-scale general-domain web bilingual corpus. These sentences are selected by our proposed unsupervised phrase-based data selection model. Compared with the traditional bag-of-words models, our phrase-based data selection model is more effective because it captures contextual information in modeling the selection of phrase as a whole, rather than selection of single words in isolation. These pseudo in-domain subcorpora can then be used to train small domain-adapted spoken language translation system which outperforms the system trained on the entire corpus, with an increase of 1.6 BLEU points. Performance is further improved when we use these pseudo in-domain corpus/models in combination with the true in-domain corpus/model, with increases of 4.5 and 3.9 BLEU points over single in- and general-domain baseline system, respectively.

Keywords: domain adaptation, phrase-based data selection, pseudo in-domain subcorpora, spoken language translation.

1 Introduction

Statistical machine translation (SMT) system performance is dependent on the quantity and quality of available training data. It seems to be a universal truth that translation performance can always be improved by using more training data, but only if the training data is reasonably well-matched with the current translation task [14]. It is also obvious that among the large training data the topics or domains of discussion will change [3], which causes the mismatch problems with the translation task. For these reasons, one would prefer to use more in-domain data for training, and this would empirically better target the translation task at hand [12,11]. However, parallel in-domain data is usually hard to find, and so performance is assumed to be limited by the quantity of domain-specific training data used to build the model. Additional bilingual data can be

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 116–126, 2013.

readily acquired, but at the cost of specificity: either the data is entirely unrelated to the task at hand, or the data is from a broad enough pool of topics and styles, such as the web, that any use this corpus may provide is due to its size, and not its relevance [1].

Domain adaptation task in SMT is to translate a text in a particular (target) domain for which only a small amount of training data is available, using a SMT system trained on a larger set of data that is not restricted to the target domain. We call this larger set of data a *general-domain* corpus, which allows a large uncurated corpus to include some text that may be relevant to the target domain.

Many existing domain adaptation methods fall into two broad categories. First, adaptation can be done at the corpus level, by selecting or weighting the data sets upon which the models are trained [1,5,13]. Second, it can be also achieved at the model level by mixing multiple translation models together [1,4,7], often in a weighted manner. In this paper, we explore both of the two above categories.

Firstly, we propose two types (monolingual and bilingual) phrase-based data selection models, and assume that data selection should be performed at the phrase level. Compared with the traditional bag-of-words models that account for data selection of single words in isolation [5,13], our two phrase-based data selection model are potentially more effective because they captures some contextual information in modeling the selection of phrase as a whole. More precise selection can be determined for phrases than for words, as we will show in the experiments.

Nextly, we use the phrase-based data selection models for ranking the sentence pairs in a large-scale general-domain web bilingual corpus with respect to an in-domain corpus. A cutoff can then be applied to produce a very small but useful subcorpus, which in turn can be used to train a domain-adapted SMT system. We show that it is possible to use our data selection models to subselect less than 18% of a large general training corpus and still increase translation performance by nearly 1.6 BLEU points on the IWSLT task.

Finally, we explore how best to use these selected subcorpora. We test their combination with the in-domain corpora, followed by examining the subcorpora to see whether they are actually in-domain, out-of-domain, or something in between. Based on this, we compare translation model combination methods. We show that these tiny translation models for model combination can improve system performance even further over the current standard way of producing a domain-adapted SMT system. The resulting process is lightweight, simple, and effective. Performance is further improved when we use these domain-adapted corpus/models in combination with the true in-domain corpus/model, with increases of 4.5 and 3.9 BLEU points over single in- and general-domain baseline system, respectively.

The remainder of this paper is organized as follows. Section 2 describes our proposed monolingual and bilingual phrase-based data selection methods. Section 3 presents the large-scale general-domain web corpus, domain adaptation results and experimental analysis, and followed by conclusions and future work in section 4.

2 Phrase-Based Data Selection

For the phrase-based translation model [6], the basic translate unit is phrase, that is to say, a continuous word sequence. It is a natural idea to use the phrase to measure the similarity between the sentence pairs in in- and general-domain corpus. If the sentence pair in general-domain corpus which are selected contain more phrases in in-domain corpus, the sentence pair is more similar to the in-domain corpus. Then we try to select the bilingual sentence pairs from the general-domain corpus which can cover more phrases of the in-domain corpus as the similar sentence pair for domain adaptation. Next, we will first describe the monolingual phrase-based data selection, and then extend it to bilingual data selection.

2.1 Monolingual Phrase-Based Data Selection

In our monolingual phrase-based data selection model, the phrases play a vital role. Inspired by the work of [9,10], we assume the following generative process. Firstly, we extract all the phrases from the source-side sentences in the in-domain bilingual corpus and assign them different weights. We take two aspects into account to estimate the weight of phrase: the information it contains and the length of the phrase.

In information theory, the information contained in a statement is measured by the negative logarithm of the probability of the statement [2,8]. Therefore, we should estimate the probability of each phrase firstly. We class the phrases with their lengths and only use the phrases whose length is not longer than five[1] in order to avoid the sparse data problem. We calculate the probabilities of the phrases based on their lengths. For a phrase p, $|p|$ represents its length, and the probability $P(p)$ is estimated by the following formula:

$$P(p) = \frac{count(p)}{\sum_{|p_i|=|p|} count(p_i)} \tag{1}$$

where the numerator $count(p)$ is the total number of phrase p appearing in the source-side sentences of the in-domain bilingual corpus, and the denominator is the total number of the phrases whose length is equal to $|p|$. It is worth to notice that letting the phrase length be one reduces the model from phrase to word, and we get word frequency. Though this is somewhat similar to TF-IDF, our approach is based on information theory, they are different in essence and get different performances.

Then, the information contained in phrase p is calculated as follows,

$$I(p) = -\log P(p) \tag{2}$$

In this way, we get the information contained in each phrase. Because the translation model is based on phrase, the longer phrase will lead to better translation. Therefore, we take $|p|$, the length of phrase, into account. We use the

[1] In our experiments, when the phrase length is large than five, the phrase become sparse sharply, and the performance of selected sentences decreases consistently.

square root of length, but not the length directly because of the data smoothing problem. The formula used to calculate the weight of each phrase is shown as follows,

$$W(p) = \sqrt{|p|} \cdot I(p) \tag{3}$$

Next, we get the weight for each phrase in the source-side sentences of the in-domain bilingual corpus based on the length of the phrase and the information it contains. Then we can estimate the average weight of a source-side sentence in the sentence pair of general-domain bilingual corpus by the weights of all the phrases it contains. For a source-side sentence s_{src} in the bilingual sentence pair, if more phrases it contains appear in the source-side sentences of the in-domain bilingual corpus, we assign it a larger score. Thus, the score of the source-side sentence s_{src} can be calculated by the following formula:

$$Score_1^{mono} = \frac{\sum_{p \in P_{src}^I} W(p)}{|s_{src}|} \tag{4}$$

where $|s_{src}|$ represents its length, and P_{src}^I is the set of all the phrases contained in the source-side sentences of the in-domain bilingual corpus.

We extract all the phrases whose length is not longer than five in sentence s_{src}, and add all the weights of phrases together. If a phrase does not appear in the source-side sentences of the in-domain corpus, the weight of the phrase is set to zero. Then, the sentence pairs are sorted by their source-side sentence's score $Score_1^{mono}$ in a descending order, and we select the sentence pair whose $Score_1^{mono}$ higher as the similar sentence pairs and add into pseudo in-domain corpus.

To further improve the performance, we also define another formula to estimate the weight of sentence s_{src}, as follows,

$$Score_2^{mono} = \frac{\sum_{p \in P_{src}^I} W(p) - \sum_{p \in (P_{src}^G - P_{src}^I)} W(p)}{|s_{src}|} \tag{5}$$

where, P_{src}^G is the set of all the phrases contained in the source-side sentences of the general-domain bilingual corpus.

Compared with $Score_1^{mono}$, in this formula, we consider the phrases which have occurred in P_{src}^G but not occurred in P_{src}^I as the unseen phrases, assume these unseen phrases have negative information to similarity measure, and assign lower score to the source-side sentence s_{src} which has more unseen phrases. This means the sentence pair in the general-domain bilingual corpus whose source-side sentence contains more unseen phrases, would not like to be selected as the similar sentence pair. The weights of the unseen phrase are calculated as Equation (1) to Equation (3) on the source-side sentences of a random subset[2] from the general-domain bilingual corpus.

[2] In our experiments, the size of the random subset is equal to the size of in-domain corpus.

2.2 Bilingual Phrase-Based Data Selection

To further use the above monolingual criteria for data selection, we propose another new model that takes into account the bilingual nature of the problem. To this end, we sum monolingual phrase-based similarity score over each side of the bilingual sentence pair, both source- and target-side,

$$Score_1^{bi} = \frac{\sum_{p \in P_{src}^I} W(p)}{|s_{src}|} + \frac{\sum_{p \in P_{tgt}^I} W(p)}{|s_{tgt}|} \tag{6}$$

$$Score_2^{bi} = \frac{\sum_{p \in P_{src}^I} W(p) - \sum_{p \in (P_{src}^G - P_{src}^I)} W(p)}{|s_{src}|} + \frac{\sum_{p \in P_{tgt}^I} W(p) - \sum_{p \in (P_{tgt}^G - P_{tgt}^I)} W(p)}{|s_{tgt}|} \tag{7}$$

Again, the sentence pair in the general-domain bilingual corpus which has higher sum scores are presumed to be better. These two models reuse the two extract phrase sets from the source-side sentences in in- and general-domain bilingual corpus, respectively, but requires the corresponding similarly-trained twos over the English side.

3 Experiments and Results

3.1 Corpora

We conduct our experiments on the International Workshop on Spoken Language Translation (IWSLT) Chinese-to-English task. Two corpora are needed for the domain adaptation task. Our in-domain bilingual corpus consists of the Basic Traveling Expression corpus and China-Japan-Korea corpus, which contains 0.38M parallel sentence pairs with 3.5/3.82M words of Chinese/English. Our general-domain bilingual corpus are collected from web data (Baidu[3], Youdao[4], Huajian[5] and Shooter[6]), which contains 11M parallel sentences pairs with 123/135M words of Chinese/English, and they are most relevant to the spoken language domain. The test set is IWSLT 2007 test set which consists of 489 sentences with 4 English reference translations each, and the development set is IWSLT 2005 test set which consists of 506 sentences with 4 English reference translations each.

[3] The example bilingual sentence pairs in `http://dict.baidu.com/`
[4] The example bilingual sentence pairs in `http://dict.youdao.com/`
[5] The example bilingual sentence pairs in `http://www.hjtrans.com/`
[6] The bilingual subtitles in `http://www.shooter.cn/xml/list/sub`

3.2 System Description

We use an out-of-the-box Moses[7] *(2010-8-13 version)* framework to implement the phrase-based machine translation system. GIZA++ [17] is used to get word alignments from the bilingual corpus with *grow-diag-final-and* option. Using the English side of the corresponding bilingual corpus, we estimate the 4-gram language models (LM) by the SRILM toolkit [19] with interpolated modified Kneser-Ney discounting. We perform minimum error rate training [16] to tune the feature weights on the development set. The translation quality is evaluated by case-insensitive BLEU-4 metric [18] using the script *mteval-v13a.pl*.

3.3 Baseline System

Using the corresponding corpus, the baseline translation models (in- and general-domain) are generated by Moses with default parameter settings. The BLEU scores of the baseline single-corpus systems are in Table 3. The results show that a translation system trained on the general-domain corpus outperforms a system trained on the in-domain corpus by over 0.5 BLEU points.

Table 1. Baseline translation results for in- and general-domain corpus

Corpus	BLEU	
	Development	Test
In	51.94	40.62
General	48.32	41.15

3.4 Selecting Subset from the General-Domain Corpus

The baseline results show that a translation system trained on the general-domain corpus outperforms a system trained on the in-domain corpus by over 0.5 BLEU points. However, this can be further improved. In our experiments, we consider the following methods for extracting domain targeted sentence pairs from the general-domain bilingual corpus:

TF-IDF is the foundation of our experiment since it has gained significant performance for data selection based translation model adaptation [5,13]. We use it as the source-side monolingual formula for data selection.

Bilingual Cross-Entropy Difference (BCED) [1] is chosen to be compared with our approach, because it also captures contextual information when selecting similar data, and it is used to select data from large-scale general-domain corpus for SMT. It sum cross-entropy difference over each side of the sentence pair in general-domain bilingual corpus, both source- and target-side:

$$[H_{I-src}(s_{src}) - H_{G-src}(s_{src})] + [H_{I-tgt}(s_{tgt}) - H_{G-tgt}(s_{tgt})] \quad (8)$$

[7] `http://www.statmt.org/moses/index.php?n=Main.HomePage`

where, in our implementation, the in-domain source- and target-side LM are estimated by the corresponding side of in-domain bilingual corpus, the general-domain source- and target-side LM are estimated on the corresponding side of a random subset of 0.38M sentences pairs[8] from the general-domain corpus, respectively.

$\mathbf{Score}_1^{mono}$ and $\mathbf{Score}_2^{mono}$ are our proposed source-side monolingual phrase-based data selection model, and $\mathbf{Score}_1^{bi}$ and $\mathbf{Score}_2^{bi}$ are our proposed bilingual phrase-based data selection model, respectively.

Regardless of method, the overall procedure is the same. Using the scoring method, we rank the individual sentence pairs of the general-domain corpus. The net effect is that of domain adaptation via threshhold filtering. New SMT systems are then trained solely on these small subcorpora, and compared against the baseline model trained on the entire 11M sentence pairs of the general-domain corpus.

We select the top N sentence pairs using each scoring method, varying N from 0.5M to 6M, and then train the corresponding translation models on these subcorpora. These translation models are then used to test the performance on the development set, as shown in Fig. 1. These subcorpora outperforms the entire general-domain corpus, and yet, no of them are anywhere near the performance of the in-domain corpus. From this, it can be deduced that data selection methods are not finding data that is strictly in-domain. Rather they are selecting *pseudo in-domain data* which is relevant, but with a different distribution than the original in-domain corpus. The results show that the top 2M pseudo in-domain sentence pairs works best. From now, we use this top 2M sentence pairs out of the 11M general-domain corpus for the next experiments.

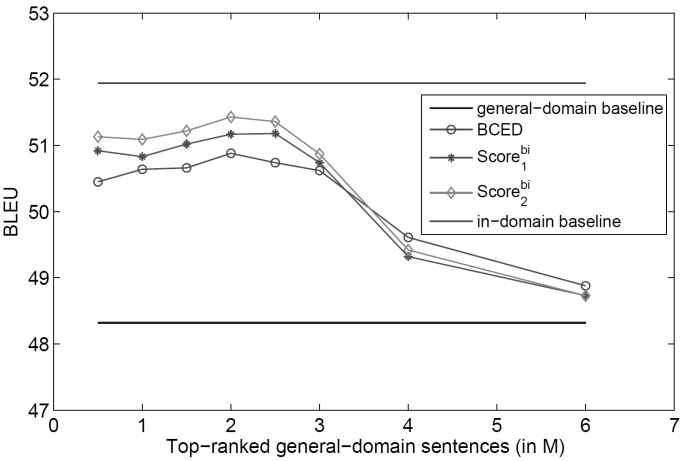

Fig. 1. The translation results of pseudo in-domain sentence pairs selection from the large-scale general-domain corpus on the development set

[8] Which is equal to the size of in-domain corpus.

Table 2. Translation results of using only a subset of the general-domain corpus

Method	Sentence Pairs	BLEU	
		Development	Test
General	11M	48.32	41.15
TF-IDF	2M	50.15	41.92
BCED	2M	50.68	42.43
$Score_1^{mono}$	2M	50.42	42.21
$Score_1^{mono}$	2M	50.56	42.13
$Score_1^{bi}$	2M	50.97	42.51
$Score_2^{bi}$	2M	51.23	42.77

Table 2 contains BLEU scores of the systems trained on subsets (pseudo in-domain sentence pairs) of the general-domain corpus. Using only the source-side monolingual phrase-based score ($Score_1^{mono}$ and $Score_1^{mono}$) are able to outperform the general-domain model when selecting 2M out of the entire 11M sentence pairs. The previous BCED (bilingual cross-entropy difference) works better. The bilingual phrase-based method ($Score_1^{bi}$ and $Score_2^{bi}$) proposed in this paper work best, especially $Score_2^{bi}$ consistently boosting performance by +1.6 BLEU points while using less than 18% of the available training data (2M sentence pairs). Consider the unseen phrases can further improve the performance of the phrase-based data selection model ($Score_2^{bi}$ vs. $Score_1^{bi}$; $Score_2^{mono}$ vs. $Score_1^{mono}$).

3.5 Mixing Corpus

As further evidence, consider the results of mixing the in-domain corpus with the best extracted sub pseudo in-domain corpus to train a single translation system in Table 3.

Table 3. Translation results of mixing the in-domain and pseudo in-domain data to train a single model

Method	Sentence Pairs	BLEU	
		Development	Test
In	0.38M	51.94	40.62
General	11M	48.32	41.15
BCED	2M	50.68	42.43
$Score_1^{bi}$	2M	50.97	42.51
$Score_2^{bi}$	2M	51.23	42.77
In+BCED	2.38M	53.51	44.24
In+$Score_1^{bi}$	2.38M	53.83	44.46
In+$Score_2^{bi}$	2.38M	54.16	44.52

The change in both the development and test scores appears to reflect dissimilarity in the underlying data. Were the two data sets more alike, one would expect the models to reinforce each other rather than cancel out. Mixing the pseudo in-domain data with in-domain data outperforms the in- and general-domain data, and with increases of 3.9 ("In+$Score_2^{bi}$" vs. "In") and 3.4 ("In+$Score_2^{bi}$" vs. "General") BLEU points, respectively.

3.6 Mixing Models

Finally, we test the approach in [4,7], passing the two phrase tables directly to the decoder and tuning a system using both phrase tables in parallel. Each phrase table receives a separate set of weights during tuning, thus this mixed translation model has more parameters than a normal single-table system. Unlike the previous work [15], we explicitly did not attempt to resolve any overlap between the two phrase tables, as there is no need to do so with the multiple decoding paths. Any phrase pairs appearing in both models will be treated separately by the decoder. However, the exact overlap between the phrase tables was tiny, minimizing this effect.

It is well to use the in-domain data to select pseudo in-domain data from the general-domain corpus, but given that this requires access to an in-domain corpus, one might as well use it. As such, we used the in-domain translation model alongside the pseudo in-domain translation models. The detail translation results are in Table 4.

Table 4. Translation results from mixing in-domain and pseudo in-domain translation models together

Method	BLEU	
	Development	Test
In	51.94	40.62
General	48.32	41.15
In,General	53.61	43.57
In,BCED 2M	54.77	44.82
In,$Score_1^{bi}$ 2M	55.03	44.99
In,$Score_2^{bi}$ 2M	55.17	45.16

A translation system trained on the pseudo in-domain subset of the general-domain corpus, can be further improved by combining with an in-domain model. Furthermore, this system combination works better than the conventional mixing multi-model approach by up to 0.6 BLEU points ("In,$Score_2^{bi}$ 2M" vs. "In+$Score_2^{bi}$") on the test set. Thus a domain-adapted system mixing two phrase tables trained on a total of 2.38M sentences outperforms the standard multi-model system which is trained on 11M sentences. This model-combined system is also 4 BLEU points better than the general-domain system by itself, and 4.5 BLEU points ("In,$Score_2^{bi}$ 2M" vs. "In") better than the in-domain system alone.

4 Conclusions and Future Work

To improve the performance of spoken language translation, we have collected large-scale general-domain web parallel corpus, such as example bilingual sentence pairs and bilingual subtitles. However, sentence pairs from these general-domain web bilingual corpus that seem similar to an in-domain corpus may not actually represent the same distribution of language. Nonetheless, we have shown that relatively tiny amounts of the pseudo in-domain data can prove more useful than the entire general-domain corpus for the purposes of domain-targeted translation tasks. A translation model trained on any of these subcorpora can be comparable or substantially better than a translation system trained on the entire corpus.

We have also proposed two types phrase-based data selection methods to extract these pseudo in-domain sentence pairs from the general-domain corpus. Compared with the traditional bag-of-words models, our proposed methods are more effective in that they can capture contextual information instead of selecting single words in isolation, and are shown to be more efficient and stable for SMT domain adaptation. Translation models trained on data selected in this way consistently outperform the general-domain baseline while using as few as 18% (2M out of the entire 11M sentence pairs) and result in an increase of 1.6 BLEU points. Next, we have shown that mixing pseudo in-domain corpus/model with the true in-domain corpus/model significantly outperforms the two state-of-the-art translation systems trained on in- and general-domain corpus, with increases of 4.5 and 3.9 BLEU points, respectively.

In the future, it will be instructive to explore other approaches for bilingual data selection, such word-based translation model [12], bilingual topic model [11]. Besides improving translation performance, this work also provides a way to mine very large corpora in a computationally-limited environment in the future, such as on a mobile terminal. The maximum size of a useful general-domain corpus is now limited only by the availability of data, rather than by how large a translation model can be fit into memory at once.

Acknowledgments. This work was supported by 863 program in China (No. 2011AA01A207). We thank the anonymous reviewers for their insightful and helpful comments.

References

1. Axelrod, A., He, X., Gao, J.: Domain adaptation via pseudo in-domain data selection. In: Proceedings of EMNLP, pp. 355–362 (2011)
2. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley, New York (1991)
3. Eck, M., Vogel, S., Waibel, A.: Language model adaptation for statistical machine translation based on information retrieval. In: Proceedings of LREC, pp. 327–330 (2004)

4. Foster, G., Kuhn, R.: Mixture-model adaptation for SMT. In: Proceedings of ACL, pp. 128–135 (2007)
5. Hildebrand, A.S.: Adaptation of the translation model for statistical machine translation based on information retrieval. In: Proceedings of EAMT, pp. 133–142 (2005)
6. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of NAACL, pp. 48–54 (2003)
7. Koehn, P., Schroeder, J.: Experiments in domain adaptation for statistical machine translation. In: Proceedings of WMT (2007)
8. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of ICML, pp. 296–304 (1998)
9. Liu, P., Zhou, Y., Zong, C.: Approach to selecting best development set for phrase-base statistical machine translation. In: Proceedings of PACLIC, pp. 325–334 (2009)
10. Liu, P., Zhou, Y., Zong, C.: Data selection for statistical machine translation. In: Proceedings of NLP-KE, pp. 232–236 (2010)
11. Lu, S., Fu, X., Wei, W., Peng, X., Xu, B.: Joint and coupled bilingual topic model based sentence representations for language model adpataiton. In: Proceedings of IJCAI, pp. 2141–2147 (2013)
12. Lu, S., Wei, W., Fu, X., Xu, B.: Translation model based cross-lingual language model adaptation: from word models to phrase models. In: Proceedings of EMNLP-CoNLL, pp. 512–522 (2012)
13. Lv, Y., Huang, J., Liu, Q.: Improving statistical machine translation peformance by training data selection and optimization. In: Proceedings of EMNLP, pp. 343–350 (2007)
14. Moore, R., Lewis, W.: Intelligent selection for language model training data. In: Proceedings of ACL, pp. 220–224 (2010)
15. Nakov, P.: Improving English-Spanish statistical machine translation: experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In: Proceedings of WMT (2008)
16. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of ACL, pp. 160–167 (2003)
17. Och, F.J., Ney, H.: Improved statistical alignment models. In: Proceedings of ACL, pp. 440–447 (2000)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of ACL, pp. 311–318 (2002)
19. Stolcke, A.: SRILM - An extensible language modeling toolkit. In: Proceedings of ICSLP, pp. 901–904 (2002)