Effective Selection of Translation Model Training Data

Le Liu Yu Hong* Hao Liu Xing Wang Jianmin Yao
School of Computer Science & Technology, Soochow University, China
{20124227052, hongy, 20134227035, 20114227047, jyao}@suda.edu.cn

Abstract

Data selection has been demonstrated to be an effective approach to addressing the lack of high-quality bitext for statistical machine translation in the domain of interest. Most current data selection methods solely use language models trained on a small scale in-domain data to select domain-relevant sentence pairs from general-domain parallel corpus. By contrast, we argue that the relevance between a sentence pair and target domain can be better evaluated by the combination of language model and translation model. In this paper, we study and experiment with novel methods that apply translation models into domain-relevant data selection. The results show that our methods outperform previous methods. When the selected sentence pairs are evaluated on an end-to-end MT task, our methods can increase the translation performance by 3 BLEU points.

1 Introduction

Statistical machine translation depends heavily on large scale parallel corpora. The corpora are necessary priori knowledge for training effective translation model. However, domain-specific machine translation has few parallel corpora for translation model training in the domain of interest. For this, an effective approach is to automatically select and expand domain-specific sentence pairs from large scale general-domain parallel corpus. The approach is named Data Selection. Current data selection methods mostly use language models trained on small scale in-domain data to measure domain relevance and select domain-relevant parallel sentence pairs to expand training corpora. Related work in literature has proven that the expanded corpora can substantially improve the performance of machine translation (Duh et al., 2010; Haddow and Koehn, 2012).

However, the methods are still far from satisfactory for real application for the following reasons:

- There isn't ready-made domain-specific parallel bitext. So it's necessary for data selection to have significant capability in mining parallel bitext in those assorted free texts. But the existing methods seldom ensure parallelism in the target domain while selecting domain-relevant bitext.
- Available domain-relevant bitext needs keep high domain-relevance at both the sides of source and target language. But it's difficult for current method to maintain two-sided domain-relevance when we aim at enhancing parallelism of bitext.

In a word, current data selection methods can't well maintain both parallelism and domain-relevance of bitext. To overcome the problem, we first propose the method combining translation model with language model in data selection. The language model measures the domain-specific generation probability of sentences, being used to select domain-relevant sentences at both sides of source and target language. Meanwhile, the translation model measures the translation probability of sentence pair, being used to verify the parallelism of the selected domain-relevant bitext.

2 Related Work

The existing data selection methods are mostly based on language model. Yasuda et al. (2008) and Foster et al. (2010) ranked the sentence pairs in the general-domain corpus according to the perplexity scores of sentences, which are computed with respect to in-domain language models. Axelrod et al. (2011) improved the perplexity-based approach and proposed bilingual cross-entropy difference as a ranking function with in- and general- domain language models. Duh et al. (2013) employed the method of (Axelrod et al.,

* Corresponding author

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 569–573,
Baltimore, Maryland, USA, June 23-25 2014.

2011) and further explored neural language model for data selection rather than the conventional n-gram language model. Although previous works in data selection (Duh et al., 2013; Koehn and Haddow, 2012; Axelrod et al., 2011; Foster et al., 2010; Yasuda et al., 2008) have gained good performance, the methods which only adopt language models to score the sentence pairs are sub-optimal. The reason is that a sentence pair contains a source language sentence and a target language sentence, while the existing methods are incapable of evaluating the mutual translation probability of sentence pair in the target domain. Thus, we propose novel methods which are based on translation model and language model for data selection.

3 Training Data Selection Methods

We present three data selection methods for ranking and selecting domain-relevant sentence pairs from general-domain corpus, with an eye towards improving domain-specific translation model performance. These methods are based on language model and translation model, which are trained on small in-domain parallel data.

3.1 Data Selection with Translation Model

Translation model is a key component in statistical machine translation. It is commonly used to translate the source language sentence into the target language sentence. However, in this paper, we adopt the translation model to evaluate the translation probability of sentence pair and develop a simple but effective variant of translation model to rank the sentence pairs in the general-domain corpus. The formulations are detailed as below:

$$P(e|f) = \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i) \quad (1)$$

$$R = \sqrt[l_e]{P(e|f)} \quad (2)$$

Where $P(e|f)$ is the translation model, which is IBM Model 1 in this paper, it represents the translation probability of target language sentence e conditioned on source language sentence f. l_e and l_f are the number of words in sentence e and f respectively. $t(e_j|f_i)$ is the translation probability of word e_j conditioned on word $f_i$and is estimated from the small in-domain parallel data. The parameter ϵ is a constant and is assigned with the value of 1.0. R is the length-normalized IBM Model 1, which is used to score general-domain sentence pairs. The sentence pair with higher score is more likely to be generated by in-domain translation model, thus, it is more relevant to the in-domain corpus and will be remained to expand the training data.

3.2 Data Selection by Combining Translation and Language model

As described in section 1, the existing data selection methods which only adopt language model to score sentence pairs are unable to measure the mutual translation probability of sentence pairs. To solve the problem, we develop the second data selection method, which is based on the combination of translation model and language model. Our method and ranking function are formulated as follows:

$$P(e,f) = P(e|f) \times P(f) \quad (3)$$

$$R = \sqrt[l_e]{P(e|f)} \times \sqrt[l_f]{P(f)} \quad (4)$$

Where $P(e,f)$ is a joint probability of sentence e and f according to the translation model $P(e|f)$ and language model $P(f)$, whose parameters are estimated from the small in-domain text. R is the improved ranking function and used to score the sentence pairs with the length-normalized translation model $P(e|f)$ and language model $P(f)$. The sentence pair with higher score is more similar to in-domain corpus, and will be picked out.

3.3 Data Selection by Bidirectionally Combining Translation and Language Models

As presented in subsection 3.2, the method combines translation model and language model to rank the sentence pairs in the general-domain corpus. However, it does not evaluate the inverse translation probability of sentence pair and the probability of target language sentence. Thus, we take bidirectional scores into account and simply sum the scores in both directions.

$$R = \sqrt[l_e]{P(e|f)} \times \sqrt[l_f]{P(f)} + \sqrt[l_f]{P(f|e)} \times \sqrt[l_e]{P(e)} \quad (5)$$

Again, the sentence pairs with higher scores are presumed to be better and will be selected to incorporate into the domain-specific training data. This approach makes full use of two translation models and two language models for sentence pairs ranking.

4 Experiments

4.1 Corpora

We conduct our experiments on the Spoken Language Translation English-to-Chinese task. Two corpora are needed for the data selection. The in-domain data is collected from CWMT09, which consists of spoken dialogues in a travel setting, containing approximately 50,000 parallel sentence pairs in English and Chinese. Our general-domain corpus mined from the Internet contains 16 million sentence pairs. Both the in- and general- domain corpora are identically tokenized (in English) and segmented (in Chinese)[1]. The details of corpora are listed in Table 1. Additionally, we evaluate our work on the 2004 test set of "863" Spoken Language Translation task ("863" SLT), which consists of 400 English sentences with 4 Chinese reference translations for each. Meanwhile, the 2005 test set of "863" SLT task, which contains 456 English sentences with 4 references each, is used as the development set to tune our systems.

Bilingual Corpus	#sentence		#token	
	Eng	Chn	Eng	Chn
In-domain	50K	50K	360K	310K
General-domain	16M	16M	3933M	3602M

Table 1. Data statics

4.2 System settings

We use the NiuTrans [2] toolkit which adopts GIZA++ (Och and Ney, 2003) and MERT (Och, 2003) to train and tune the machine translation system. As NiuTrans integrates the mainstream translation engine, we select hierarchical phrase-based engine (Chiang, 2007) to extract the translation rules and carry out our experiments. Moreover, in the decoding process, we use the NiuTrans decoder to produce the best outputs, and score them with the widely used NIST mteval131a[3] tool. This tool scores the outputs in several criterions, while the case-insensitive BLEU-4 (Papineni et al., 2002) is used as the evaluation for the machine translation system.

4.3 Translation and Language models

Our work relies on the use of in-domain language models and translation models to rank the sentence pairs from the general-domain bilingual training set. Here, we employ ngram language model and IBM Model 1 for data selection. Thus, we use the SRI Language Modeling Toolkit (Stolcke, 2002) to train the in-domain 4-gram language model with interpolated modified Kneser-Ney discounting (Chen and Goodman, 1998). The language model is only used to score the general-domain sentences. Meanwhile, we use the language model training scripts integrated in the NiuTrans toolkit to train another 4-gram language model, which is used in MT tuning and decoding. Additionally, we adopt GIZA++ to get the word alignment of in-domain parallel data and form the word translation probability table. This table will be used to compute the translation probability of general-domain sentence pairs.

4.4 Baseline Systems

As described above, by using the NiuTrans toolkit, we have built two baseline systems to fulfill "863" SLT task in our experiments. The In-domain baseline trained on spoken language corpus has 1.05 million rules in its hierarchical-phrase table. While, the General-domain baseline trained on 16 million sentence pairs has a hierarchical phrase table containing 1.7 billion translation rules. These two baseline systems are equipped with the same language model which is trained on large-scale monolingual target language corpus. The BLEU scores of the In-domain and General-domain baseline system are listed in Table 2.

Corpus	Hierarchical phrase	Dev	Test
In-domain	1.05M	15.01	**21.99**
General-domain	1747M	27.72	**34.62**

Table 2. Translation performances of In-domain and General-domain baseline systems

The results show that General-domain system trained on a larger amount of bilingual resources outperforms the system trained on the in-domain corpus by over 12 BLEU points. The reason is that large scale parallel corpus maintains more bilingual knowledge and language phenomenon, while small in-domain corpus encounters data sparse problem, which degrades the translation performance. However, the performance of General-domain baseline can be improved further. We use our three methods to refine the general-domain corpus and improve the translation performance in the domain of interest. Thus, we build several contrasting systems trained on refined training data selected by the following different methods.

1http://www.nlplab.com/NiuPlan/NiuTrans.YourData.ch.html
2http://www.nlplab.com/NiuPlan/NiuTrans.ch.html#download
3 http://ww.itl.nist.gov/iad/mig/tools

- **Ngram**: Data selection by 4-gram LMs with Kneser-Ney smoothing. (Axelrod et al., 2011)
- **Neural net**: Data selection by Recurrent Neural LM, with the RNNLM Tookit. (Duh et al., 2013)
- **Translation Model (TM)**: Data selection with translation model: IBM Model 1.
- **Translation model and Language Model (TM+LM)**: Data selection by combining 4-gram LMs with Kneser-Ney smoothing and IBM model 1(equal weight).
- **Bidirectional TM+LM**: Data selection by bidirectionally combining translation and language models (equal weight).

4.5 Results of Training Data Selection

We adopt five methods for extracting domain-relevant parallel data from general-domain corpus. Using the scoring methods, we rank the sentence pairs of the general-domain corpus and select only the top N = {50k, 100k, 200k, 400k, 600k, 800k, 1000k} sentence pairs as refined training data. New MT systems are then trained on these small refined training data. Figure 1 shows the performances of systems trained on selected corpora from the general-domain corpus. The horizontal coordinate represents the number of selected sentence pairs and vertical coordinate is the BLEU scores of MT systems.

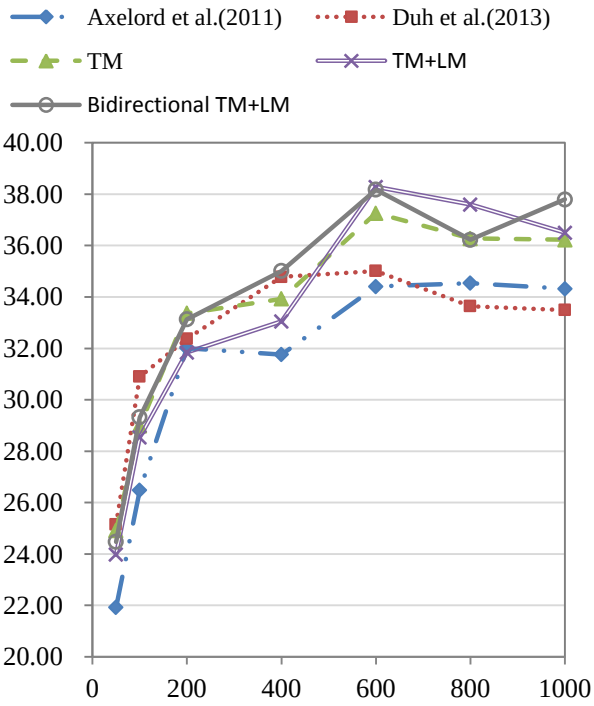

Figure 1. Results of the systems trained on only a subset of the general-domain parallel corpus.

From Figure 1, we conclude that these five data selection methods are effective for domain-specific translation. When top 600k sentence pairs are picked out from general-domain corpus to train machine translation systems, the systems perform higher than the General-domain baseline trained on 16 million parallel data. The results indicate that more training data for translation model is not always better. When the domain-specific bilingual resources are deficient, the domain-relevant sentence pairs will play an important role in improving the translation performance.

Additionally, it turns out that our methods (**TM**, **TM+LM** and **Bidirectional TM+LM)** are indeed more effective in selecting domain-relevant sentence pairs. In the end-to-end SMT evaluation, **TM** selects top 600k sentence pairs of general-domain corpus, but increases the translation performance by 2.7 BLEU points. Meanwhile, the **TM+LM** and **Bidirectional TM+LM** have gained 3.66 and 3.56 BLEU point improvements compared against the general-domain baseline system. Compared with the mainstream methods (**Ngram** and **Neural net)**, our methods increase translation performance by nearly 3 BLEU points, when the top 600k sentence pairs are picked out. Although, in the figure 1, our three methods are not performing better than the existing methods in all cases, their overall performances are relatively higher. We therefore believe that combining in-domain translation model and language model to score the sentence pairs is well-suited for domain-relevant sentence pair selection. Furthermore, we observe that the overall performance of our methods is gradually improved. This is because our methods are combining more statistical characteristics of in-domain data in ranking and selecting sentence pairs. The results have proven the effectiveness of our methods again.

5 Conclusion

We present three novel methods for translation model training data selection, which are based on the translation model and language model. Compared with the methods which only employ language model for data selection, we observe that our methods are able to select high-quality domain-relevant sentence pairs and improve the translation performance by nearly 3 BLEU points. In addition, our methods make full use of the limited in-domain data and are easily implemented. In the future, we are interested in applying

our methods into domain adaptation task of statistical machine translation in model level.

Acknowledgments

This research work has been sponsored by two NSFC grants, No.61373097 and No.61272259, and one National Science Foundation of Suzhou (Grants No. SH201212).

Reference

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK, July. Association for Computational Linguistics.

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*, 1993, 19(2): 263-311.

Stanley Chen and Joshua Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. *Technical Report 10-98, Computer Science Group, Harvard University.*

Moore Robert C, Lewis William. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*. Association for Computational Linguistics, 2010: 220-224.

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

Kevin Duh, Graham Neubig, Katsuhito Sudoh and Hajime Tsukada. Adaptation Data Selection using Neural Language Models: Experiments in Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 678-683, Sofia, Bulgaria, August 4-9 2013.

Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010.Analysis of translation model adaptation for statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) - Technical Papers Track*.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. *Empirical Methods in Natural Language Processing*.

Barry Haddow and Philipp Koehn. 2012. Analysing the effect of out-of-domain data on smt systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 422–432, Montreal, Canada, June. Association for Computational Linguistics.

Och, Franz Josef, and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational linguistics* 29.1 (2003): 19-51.

Och, Franz Josef. Minimum error rate training in statistical machine translation. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*-Volume 1. Association for Computational Linguistics, 2003.

Philipp Koehn and Barry Haddow. 2012. Towards effective use of training data in statistical machine translation. In *WMT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL*.

Andreas Stolcke. 2002. SRILM - An extensible language modeling toolkit. *Spoken Language Processing*.

Tong Xiao, Jingbo Zhu, Hao Zhang and Qiang Li. NiuTrans: an open source toolkit for phrase-based and syntax-based machine translation. In *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, 2012: 19-24.

Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. 2008. Method of selecting training data to build a compact and efficient translation model. *International Joint Conference on Natural Language Processing.*