Connecting Phrase based Statistical Machine Translation Adaptation

Rui Wang[1,2], **Hai Zhao**[1,2] *, **Bao-Liang Lu**[1,2], **Masao Utiyama**[3] * **and Eiichro Sumita**[3]
[1]Department of Computer Science and Engineering,
[2]Key Lab of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering,
Shanghai Jiao Tong University, Shanghai, China
[3]National Institute of Information and Communications Technology, Kyoto, Japan
wangrui.nlp@gmail.com, {zhaohai, blu}@cs.sjtu.edu.cn,
{mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

Although more additional corpora are now available for Statistical Machine Translation (SMT), only the ones which belong to the same or similar domains of the original corpus can indeed enhance SMT performance directly. A series of SMT adaptation methods have been proposed to select these similar-domain data, and most of them focus on sentence selection. In comparison, phrase is a smaller and more fine grained unit for data selection, therefore we propose a straightforward and efficient connecting phrase based adaptation method, which is applied to both bilingual phrase pair and monolingual n-gram adaptation. The proposed method is evaluated on IWSLT/NIST data sets, and the results show that phrase based SMT performances are significantly improved (up to +1.6 in comparison with phrase based SMT baseline system and +0.9 in comparison with existing methods).

1 Introduction

Large corpora are important for Statistical Machine Translation (SMT) training. However only the relevant additional corpora, which are also called in-domain or related-domain corpora, can enhance the performance of SMT effectively. Otherwise the irrelevant additional corpora, which are also called out-of-domain corpora, may not benefit SMT (Koehn and Schroeder, 2007).

SMT adaptation means selecting useful part from mix-domain (mixture of in-domain and out-of-domain) data, for SMT performance enhancement. The core task in adaptation is about how to select the useful data. Existing works have considered selection strategies with various granularities, though most of them only focus on sentence-level selection (Axelrod et al., 2011; Banerjee et al., 2012; Duh et al., 2013; Hoang and Sima'an, 2014a; Hoang and Sima'an, 2014b). There is a potential problem for sentence level adaptation: different parts of a sentence may belong to different domains. That is, it is possible that a sentence is overall out-of-domain, although part of it can be in-domain. Therefore a few works consider more granular level for selection. They build lexicon, Translation Models (TMs), reordering models or Language Models (LMs) to select fragment or directly adapt the models (Bellegarda, 2004; Deng et al., 2008; Moore and Lewis, 2010; Foster et al., 2010; Mansour and Ney, 2013; Carpuat et al., 2013; Chen et al., 2013a; Chen et al., 2013b; Sennrich et al., 2013; Mathur et al., 2014; Shi et al., 2015). One typical example of these methods is to train two Neural Network (NN) models (one from in-domain and the other from out-of-domain) and penalize the sentences/phrases similar to out-of-domain corpora (Duh et al., 2013; Joty et al., 2015; Durrani et al., 2015). As we know, Phrase Based SMT (PBSMT) mainly contains two models: translation model and LM, whose components are bilingual phrase pairs and monolingual n-grams. Meanwhile, most of the above methods enhance SMT performance by adapting single specific model.

*Corresponding authors. H. Zhao and B. L. Lu were partially supported by Cai Yuanpei Program (CSC No. 201304490199 and 201304490171), National Natural Science Foundation of China (No. 61672343, 61170114 and 61272248), National Basic Research Program of China (No. 2013CB329401), Major Basic Research Program of Shanghai Science and Technology Committee (No. 15JC1400103), Art and Science Interdisciplinary Funds of Shanghai Jiao Tong University (No. 14JCRZ04), and Key Project of National Society Science Foundation of China (No. 15-ZDA041).

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3135–3145, Osaka, Japan, December 11-17 2016.

Instead of focusing on sentence selection or single model adaptation, we propose a phrase adaptation method, which is applied to both bilingual phrase pair and monolingual n-gram selection. It is based on a linguistic observation that the translation hypotheses of a phrase-based SMT system are concatenations of phrases from Phrase Table (PT), which has been applied to LM growing (Wang et al., 2014a; Wang et al., 2015). As a straightforward linear method, it is much efficient in comparison with NN based non-linear methods.

The remainder of this paper is organized as follows. Section 2 will introduce the connecting phrase based adaptation method. The size of adapted connecting phrase will be tuned in Section 3. Empirical results will be shown in Section 4. We will discuss the methods and conduct extension experiments in Section 5. The last section will conclude this paper.

2 Connecting Phrase based Adaptation

Suppose that two phrases '*would like to learn*' and '*Chinese as second language*' are in the in-domain PT. In decoding, these two phrases may be connected together as '*would like to learn Chinese as second language*". The phrases '*would like to learn Chinese*' or '*learn Chinese as second language*' may be outside in-domain PT/LM, but they may possibly be in out-of-domain PT/LM. Traditionally their translation probabilities are only calculated by the combination of probabilities from in-domain PT/LM. For the proposed methods, the translation probabilities of connecting phrases from out-of-domain corpus are estimated by real corpus directly. If we can add these connecting phrases with their translation probabilities, which may be useful in decoding, into in-domain bilingual (together with source part phrases) PT or monolingual LM, they may help improve SMT.

Note that connecting phrases are generated from in-domain PT, it is necessary to check if these in-domain connecting phrases actually occur in out-of-domain PT/LM. Connecting phrases can occur in decoding by combining two phrases from in-domain PT.

Let w_a^b be a phrase starting from the a-th word and ending with the b-th word, and $\gamma w_a^b \beta$ be a phrase including w_a^b as a part of it, where γ and β represent any word sequence or none. An i-gram phrase $w_1^k w_{k+1}^i$ $(1 \leq k \leq i-1)$ is a connecting phrase[1] (Wang et al., 2014a), if

1) w_1^k is right (rear) part of one phrase γw_1^k in the in-domain PT, and

2) w_{k+1}^i is left (front) part of one phrase $w_{k+1}^i \beta$ in the in-domain PT.

For example, let '*a b c d*' be a 4-gram phrase, it is a connecting phrase if at least one of the following conditions holds:

1) 'γ *a*' and '*b c d* β' are in phrase table, or

2) 'γ *a b*' and '*c d* β' are in phrase table, or

3) 'γ *a b c*' and '*d* β' are in phrase table.

For a phrase pair (F, E) in out-of-domain PT, there are four cases: a) Both F and E, b) either F or E, c) only F, d) only E are/is connecting phrase(s). We empirically evaluate the performance of these four cases and the results show that a) gains the highest BLEU, so it is adopted at last. For an n-gram LM, we only consider target side information.

3 Adapted Phrase Size Tuning

A lot of connecting phrases are generated in the above way. We propose two methods to rank these phrases and only the top ranked ones are added into in-domain PT/LM.

[1]We are aware that connecting phrases can be applied to three or more phrases. Experimental results show that using more than two connecting phrases cannot further improve the performance, so only two connecting phrases are applied.

3.1 Occurring Probability based Tuning

The potential Occurring Probability (OP) of a source phrase $P_{op}(F)$ and $P_{op}(E)$ are defined as,

$$P_{op}(F) = \sum_{k=1}^{p-1}(\sum_{\beta} P_s(\beta f_1^k) \times \sum_{\gamma} P_s(f_{k+1}^p \gamma)),$$

$$P_{op}(E) = \sum_{k=1}^{q-1}(\sum_{\beta} P_t(\beta e_1^k) \times \sum_{\gamma} P_t(e_{k+1}^q \gamma)),$$

respectively, where P_s (for source phrase f_1^p) or P_t (for target phrase e_1^q) is calculated using source or target monolingual LM trained from in-domain corpus.

The $P_{op}(F, E)$ of a connecting phrase pair (F, E) in SMT decoding is defined as $P_{op}(F) \times P_{op}(E)$. $P_{op}(F, E)$ is used to rank connecting phrase pairs. For target LM, only $P_{op}(E)$ is used to rank connecting n-gram (Wang et al., 2014a).

3.2 NN based Tuning

The basic hypothesis of NN based adaptation is: two NN models (translation model as NNTM or LM as NNLM), one from in-domain and one from out-of-domain are trained. Taking NNTM as example, for a phrase pair (F, E) relevant with in-domain ones, the translation probabilities $P_{in}(E|F)$ by $NNTM_{in}$ should be larger and $P_{out}(E|F)$ by $NNTM_{out}$ should be lower. This hypothesis is partially motivated by (Axelrod et al., 2011), which use bilingual cross-entropy difference to distinguish in-domain and out-of-domain data.

The translation probability of a phrase-pair is estimated as,

$$P(E|F) = P(e_1, ..., e_q|f_1, ..., f_p), \tag{1}$$

where f_s ($s \in [1, p]$) and e_t ($t \in [1, q]$) are source and target words, respectively. Originally,

$$P(e_1, ..., e_q|f_1, ..., f_p) = \prod_{k=1}^{q} P(e_k|e_1, ..., e_{k-1}, f_1, ...f_p). \tag{2}$$

The structure of NN based translation model is similar to Continuous Space Translation Model (CSTM) (Schwenk, 2012). For the purpose of adaptation, the dependence between target words is dropped[2] and the probabilities of different length target phrase are normalized. For an incomplete source phrase, i.e. with less than seven words, we set the projections of the missing words to zero. The normalized translation probability $Q(E|F)$ can be approximately computed by the following equation,

$$Q(E|F) \approx \sqrt[q]{\prod_{k=1}^{q} P(e_k|f_1, ...f_p)}. \tag{3}$$

Finally, the minus $D_{minus}(E|F)$ is used to rank connecting phrase pairs from mix-domain PT,

$$D_{minus}(E|F) = Q_{in}(E|F) - Q_{in}(E|F). \tag{4}$$

where $Q_{in}(E|F)$ and $Q_{in}(E|F)$ are corresponding probabilities from in-domain and out-of-domain NNTMs.

For monolingual n-gram tuning, two NNLMs (in and out) are trained, and

$$D_{minus}(E) = Q_{in}(E) - Q_{out}(E), \tag{5}$$

[2]We have also empirically compared the performance of using NN with target word dependence and the results are not that positive.

where $Q_{in}(E)$ and $Q_{out}(E)$ are corresponding probabilities from in-domain and out-of-domain NNLMs. $D_{minus}(E)$ is used for n-gram ranking.

Beside for connecting phrases size tuning, this NN[3] based method can also be applied to phrase adaptation directly, which is similar as other NN based adaptation methods, such as (Duh et al., 2013) for sentence selection and (Joty et al., 2015) for joint model adaptation. In addition, the translation probabilities of connecting phrases calculated by NN can also be used to enhance SMT, and the experimental results will be shown in Section 5.4.

3.3 Integration into SMT

The thresholds of P_{op} and D_{minus} are tuned using development data. Selected phrase pairs are added into the in-domain PT. Because they are not so useful as the in-domain ones, a penalty score is added. For in-domain phrase pairs, the penalty is set as 1; for the out-of-domain ones the penalty is set as e (= 2.71828...). Other phrase scores (lexical weights et. al.) are used as they are. This penalty setting is similar to (Bisazza et al., 2011). Penalty weights, together with all of existing score weights, will be further tuned by MERT (Och, 2003). The phrase pairs in re-ordering model are selected using the same way as PT. The selected monolingual n-grams are added to the original LM, and the probabilities are re-normalized by SRILM (Stolcke, 2002; Stolcke et al., 2011).

4 Experiments

4.1 Data sets

The proposed methods are evaluated on two data sets. 1) IWSLT 2014 French (FR) to English (EN) corpus[4] is used as in-domain data and dev2010 and test2010/2011 (Niehues and Waibel, 2012), are selected as development (dev) and test data, respectively. Out-of-domain corpora contain Common Crawl, Europarl v7, News Commentary v10 and United Nation (UN) FR-EN parallel corpora[5]. 2) NIST 2006 Chinese (CN) to English corpus[6] is used as in-domain corpus, which follows the setting of (Wang et al., 2014b) and mainly consists of news and blog texts. Chinese to English UN data set (LDC2013T06) and NTCIR-9 (Goto et al., 2011) patent data are used as out-of-domain bilingual (Bil) parallel corpora. The English patent data in NTCIR-8 (Fujii et al., 2010) is also used as additional out-of-domain monolingual (Mono) corpus. NIST Eval 2002-2005 and NIST Eval 2006 are used as dev and test data, respectively.

IWSLT FR-EN	Sentences	Tokens
in-domain	178.1K	3.5M
out-of-domain	17.8M	450.0M
dev	0.9K	20.1K
test2010	1.6K	31.9K
test2011	1.1K	21.4K
NIST CN-EN	**Sentences**	**Tokens**
in-domain	430.8K	12.6M
out-of-domain (Bil)	8.8M	249.4M
out-of-domain (Mono)	33.7M	1.0B
dev (average of four)	4.4K	145.8K
test (average of four)	1.6K	46.7K

Table 1: Statistics on data sets ('B' for billions).

[3]NN based methods have been applied to a series of NLP tasks, such as Chinese word segmentation and parsing (Cai and Zhao, 2016; Zhang et al., 2016).

[4]https://wit3.fbk.eu/mt.php?release=2014-01

[5]http://statmt.org/wmt15/translation-task.html

[6]http://www.itl.nist.gov/iad/mig/tests/mt/2006/

4.2 Common Setting

The basic settings of IWSLT-2014 FR to EN and NIST-06 CN to EN phrase based translation baseline systems are followed. 5-gram interpolated KN (Kneser and Ney, 1995) LMs are trained. Translation performances are measured by case-insensitive BLEU (Papineni et al., 2002) with significance test (Koehn, 2004) and METEOR (Lavie and Agarwal, 2007). MERT (Och, 2003) (BLEU based) is run three times for each system and the average BLEU/METEOR scores are recorded. 4-layer CSTM (Schwenk, 2012) are applied to NN translation models: phrase length limit is set as seven, shared projection layer of dimension 320 for each word (that is 2240 for seven words), projection layer of dimension 768, hidden layer of dimension 512. The dimensions of input/output layers for both in/out-of-domain CSTMs follows the size of vocabularies of source/target words from in-domain corpora. That is 72K/57K for IWSLT and 149/112K for NIST. Since out-of-domain corpora are huge, part of them are resampled (resample coefficient 0.01 for IWSLT and NIST).

Several related existing methods are selected as baselines[7]: Koehn and Schroeder (2007)'s method for using two (in and out-of-domain) TMs and LMs together, entropy based method for TM (Ling et al., 2012) and LM (Stolcke, 1998) adaptation (pruning), (Duh et al., 2013) for NNLM based sentence adaptation, (Sennrich, 2012) for TM weights combination, and (Bisazza et al., 2011) for TM fill-up. In Table Tables 2 and 3, 'in-domain', 'out-of-domain' and 'mix-domain' indicate training all models using corresponding corpora, 'in+NN' indicates applying NN based adaptation directly for all phrases, and 'in+connect' indicates adding all connecting phrases and n-grams to in-domain PT and LM, respectively. For tuning methods, 'in+connect+OP/NN' indicates tuning connecting phrase pairs and n-grams using Occurring Probability (OP) and NN, respectively. Only the best preforming systems (for both the baselines and proposed methods) on development data are chosen to be evaluated on test data.

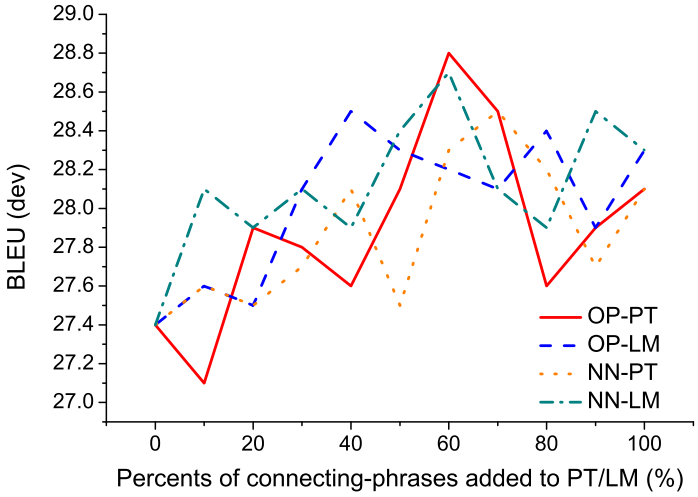

Figure 1: Connecting phrases size tuning on IWSLT.

4.3 Results and Analysis

For all ranked connecting phrase pairs and n-grams, we empirically add different sized (top) parts of them into PT/LM for size tuning. Figure 1 shows performances of the proposed tuning methods on IWSLT development data set. The results show that adding connecting phrases can enhance SMT performance in most of cases. Meanwhile, the tuned connecting phrases, which are parts of the whole, gain more BLEU improvement. They are considered as the most useful connecting phrases and evaluated on the test data sets.

[7]We are aware that there are various SMT adaptation works such as (Deng et al., 2008; Hoang and Sima'an, 2014a; Joty et al., 2015). However, there does not exist a commonly used evaluation corpus for this task, and either detailed implementations or experimental settings are absent for most published works.

Methods	PT Size	LM Size	BLEU test10	METEOR test10	BLEU test11	METEOR test11
in-domain	9.8M	7.9M	31.94	34.07	29.16	32.34
out-of-domain	759.0M	497.4M	27.34	32.22	23.80	30.48
mix-domain	765.4M	503.1M	30.07	33.19	26.42	31.06
Koehn's method	N/A	N/A	32.42	**34.32**	29.41	32.41
entropy method	247.8M	146.1M	32.54	34.12	29.23	32.17
Duh's method	765.4M	271.0M	**32.65**	34.31	29.18	32.57
Sennrich's method	765.4M	503.1M	32.41	34.32	**29.67**	**32.71**
Bisazza's method	765.4M	503.1M	32.24	34.28	29.35	32.53
in+NN	296.8M	156.2M	32.54	34.25	29.67	32.68
in+connect	184.5M	133.8M	33.26+	34.60	30.07	32.89
in+connect+OP	122.0M	53.5M	**33.53**++	**34.77**	30.25	32.91
in+connect+NN	141.3M	80.3M	32.91	34.56	**30.32**+	**33.17**

Table 2: IWSLT FR-EN Results. "++": BLEU significantly better than corresponding the best performed baseline (in **bold**) at level $\alpha = 0.01$, "+": $\alpha = 0.05$. Koehn's method uses two TMs and LMs, so their sizes are hard to tell.

Methods	PT Size	LM Size	BLEU	METEOR
in-domain	27.2M	23.9M	32.10	29.29
out-of-domain	365.8M	1.2B	27.85	22.48
mix-domain	370.9M	1.2B	31.37	28.80
Koehn's method	N/A	N/A	31.93	29.32
entropy method	165.3M	279.5M	32.29	29.17
Duh's method	160.5M	519.3M	**32.51**	29.36
Sennrich's method	370.9M	1.2B	32.36	**29.88**
Bisazza's method	370.9M	1.2B	32.15	29.72
in+NN	187.6M	394.1M	32.82+	30.23
in+connect	142.6M	298.1M	32.63	29.97
in+connect+OP	92.6M	208.7M	32.76	**30.63**
in+connect+NN	113.6M	142.1M	**33.23**++	30.54

Table 3: NIST-06 CN-EN Results.

Tables 2 and 3 shows that directly using 'out-of-domain' or 'mix-domain' data will cause SMT performances decrease in comparison with 'in-domain' data. Adding connecting phrases will enhance SMT performances and the proposed tuning method can further increase SMT performances significantly (up to +1.6 BLEU in IWSLT task and +1.1 in NIST task) and outperform the existing methods (up to +0.9 BLEU in IWSLT task and +0.7 in NIST task). The NN method performs better as a tuning method than as a direct adaptation method.

5 Discussions

5.1 Individual Model Analysis

Most of the existing methods focus on single model adaptation, however the proposed connecting phrase method can be applied to both TM and LM. So it seems a little unfair to compare the existing methods with our methods. To compare with them in a more fair way, we show the performance of individual model in Tables 4 and 5 for IWSLT tasks. Similar as the previous experiments, only the best performing system on development data of each method is evaluated on the test data.

Methods	LM Size	BLEU test10	BLEU test11
in-domain	7.9M	31.94	29.16
out-of-domain	497.4M	31.01	27.42
mix-domain	503.1M	32.23	28.42
Koehn's method	N/A	32.34	29.10
entropy method	146.1M	32.31	29.24
Duh's method	271.0M	32.65	29.18
in+NN	156.2M	32.66	29.38
in+connect	133.8M	32.78	29.32
in+connect+OP	53.5M	**32.95**	29.45
in+connect+NN	80.3M	32.56	**29.78**

Table 4: IWSLT FR-EN results on LM adaptation.

Methods	PT Size	BLEU test10	BLEU test11
in-domain	9.8M	31.94	29.16
out-of-domain	759.0M	28.62	24.56
mix-domain	765.4M	29.56	26.78
Koehn's method	N/A	31.97	29.21
entropy method	247.8M	32.43	28.73
Sennrich's method	765.4M	32.41	29.67
Bisazza's method	765.4M	32.24	29.35
in+NN	296.8M	32.31	29.63
in+connect	184.5M	32.87	29.48
in+connect+OP	122.0M	**33.05**	29.77
in+connect+NN	141.3M	32.73	**29.89**

Table 5: IWSLT FR-EN results on TM adaptation.

As shown in Tables 4 and 5, the proposed methods outperform existing methods in individual model performance (up to +0.3 BLEU in LM task and +0.6 BLEU in TM task for test10 and +0.5 BLEU in LM task and +0.2 BLEU in TM task for test11). Another observation is that adding out-of-domain data into

TM hurt SMT system more seriously than LM (-0.9 BLEU in LM task versus -3.4 BLEU in TM task for test10 and -1.7 BLEU in LM task versus -4.6 BLEU in TM task for test11).

5.2 Manual Example

A few adapted phrase examples of IWSLT FR-EN task are in Table 6. For NN based method (direct apply NN in adaptation), some phrases with similar meaning are adapted, such as *third world countries* and *developing countries*. For connecting phrase method, phrases which are combination of phrases are adapted, such as *the reason* and *why I like* form *the reason why I like*.

Methods	Source Phrases	Original Target Phrases	Adapted Phrases
NN	*les pays en voie de développement*	*1. developing countries* *2. the developing countries* *3. all developing countries*	*1. developing countries* *2. third world countries* *3. countries in the developing world*
Connect	*la raison pour laquelle je tiens*	*1. the reason I want* *2. why I like* *3. I therefore wish*	*1. the reason why I like* *2. the reason I want* *3. the reason I would like*

Table 6: Some examples of adapted phrases, which are ranked by translation probabilities.

5.3 Efficiency Comparison

Table 7 shows the adaptation time of each method[8] on IWSLT task. The proposed methods show significant advantage over others, and NN based methods are very time consuming.

Methods	Adaptation Time
entropy method	12 hours
Duh's method	7 days
Bisazza's method	6 hours
in+NN	10 days
in+connect	**2 hours**
in+connect+OP	3 hours
in+connect+NN	3 days

Table 7: Efficiency comparison (CPU time) on IWSLT.

5.4 Adding NN Probabilities

As mentioned in Section 3.2, NN can be used to predict the translation probabilities of bilingual phrase pairs and the occurring probabilities of monolingual n-grams. The minus D_{minus} between in-domain NN probabilities Q_{in} and out-of-domain NN probabilities Q_{out} are used to judge whether a phrase (pair) is similar to the in-domain ones. Meanwhile, these in-domain NN probabilities Q_{in} themselves are also useful information. In the previous sections, the adapted phrase pairs are added into original PT or LM with their own probabilities. In this subsection, Q_{in} of adapted and original phrases are also adopted in SMT decoding. That is, $Q_{in}(E|F)$ is added as a feature for adapted and original phrase pairs in PT and $Q_{in}(E)$ of adapted and original n-grams are interpolated with n-gram LM probabilities.

The results in Table 8 show that the NN feature can enhance SMT performance slightly. Although this is not our main contribution, it shows the NN method cannot only be applied to phrase pair and n-gram adaptation, but also to probability estimation.

[8] Koehn and Schroeder (2007) is only for model combination, so we do not compare with it.

Methods	PT Size	LM Size	BLEU without Q_{in}	BLEU with Q_{in}
in-domain	9.8M	7.9M	31.94	32.34
in+NN	296.8M	156.2M	32.54	32.48
in+connect	184.5M	133.8M	33.26	33.45
in+connect+OP	122.0M	53.5M	33.53	**33.67**
in+connect+NN	141.3M	80.3M	32.91	33.12

Table 8: IWSLT FR-EN Results.

6 Conclusion

In this paper, we propose a straightforward connecting phrase based SMT adaptation method. Two model size tuning methods, NN and occurring probability are proposed to discard less reliable connecting phrases. The empirical results in IWSLT French to English and NIST Chinese to English translation tasks show that the proposed methods can significantly outperform a number of the existing SMT adaptation methods in both performance and efficiency. We also show some empirical results to discuss where the SMT improvements come from by individual model and manual example analysis.

Acknowledgements

Thanks Zhisong Zhang for helpful discussions and the three anonymous reviewers for helpful comments.

References

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, U.K.

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier, Andy Way, and Josef van Genabith. 2012. Translation quality-based supplementary data selection by incremental update of translation models. In *Proceedings of 24th International Conference on Computational Linguistics*, pages 149–166, Mumbai, India.

Jerome R Bellegarda. 2004. Statistical language model adaptation: review and perspectives. *Speech Communication*, 42(1):93–108.

Arianna Bisazza, Nick Ruiz, Marcello Federico, and FBK-Fondazione Bruno Kessler. 2011. Fill-up versus interpolation methods for phrase-based smt adaptation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 136–143, San Francisco.

Deng Cai and Hai Zhao. 2016. Neural word segmentation learning for chinese. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 409–420, Berlin, Germany.

Marine Carpuat, Hal Daume III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi, and Rachel Rudinger. 2013. Sensespotting: Never let your parallel data tie you to an old domain. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1435–1445, Sofia, Bulgaria.

Boxing Chen, George Foster, and Roland Kuhn. 2013a. Adaptation of reordering models for statistical machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 938–946, Atlanta, Georgia.

Boxing Chen, Roland Kuhn, and George Foster. 2013b. Vector space model for adaptation in statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1285–1293, Sofia, Bulgaria.

Yonggang Deng, Jia Xu, and Yuqing Gao. 2008. Phrase table training for precision and recall: What makes a good phrase and a good phrase pair? In *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 81–88, Columbus, Ohio.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683, Sofia, Bulgaria, August.

Nadir Durrani, Hassan Sajjad, S Joty, A Abdelali, and S Vogel. 2015. Using joint models for domain adaptation in statistical machine translation. *Proceedings of the Fifteenth Machine Translation Summit*.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Cambridge, MA, October.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2010. Overview of the patent translation task at the ntcir-8 workshop. In *In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 293–302, Tokyo, Japan.

Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. 2011. Overview of the patent machine translation task at the NTCIR-9 workshop. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 559–578, Tokyo, Japan.

Cuong Hoang and Khalil Sima'an. 2014a. Latent domain phrase-based models for adaptation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 566–576, Doha, Qatar.

Cuong Hoang and Khalil Sima'an. 2014b. Latent domain translation models in mix-of-domains haystack. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1928–1939, Dublin, Ireland.

Shafiq Joty, Hassan Sajjad, Nadir Durrani, Kamla Al-Mannai, Ahmed Abdelali, and Stephan Vogel. 2015. How to avoid unwanted pregnancies: Domain adaptation using neural network models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1259–1270, Lisbon, Portugal.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. *Proceedings of 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing,*, 1:181–184.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231.

Wang Ling, João Graça, Isabel Trancoso, and Alan Black. 2012. Entropy-based pruning for phrase-based machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 962–971, Jeju Island, Korea.

Saab Mansour and Hermann Ney. 2013. Phrase training based adaptation for statistical machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 649–654, Atlanta, Georgia.

Prashant Mathur, Sriram Venkatapathy, and Nicola Cancedda. 2014. Fast domain adaptation of smt models without in-domain parallel data. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1114–1123, Dublin, Ireland.

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 220–224, Uppsala, Sweden.

Jan Niehues and Alex Waibel. 2012. Continuous space language models using restricted boltzmann machines. In *Proceedings of the International Workshop for Spoken Language Translation*, IWSLT 2012, pages 311–318, Hong Kong, China.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Holger Schwenk. 2012. Continuous space translation models for phrase-based statistical machine translation. In *Proceedings of 24th International Conference on Computational Linguistics: Posters*, pages 1071–1080, Mumbai, India.

Rico Sennrich, Holger Schwenk, and Walid Aransa. 2013. A multi-domain translation model framework for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 832–840, Sofia, Bulgaria.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549, Avignon, France.

Yangyang Shi, Martha Larson, and Catholijn M. Jonker. 2015. Recurrent neural network language model adaptation with curriculum learning. *Computer Speech and Language*, 33(1):136 – 154.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at sixteen: Update and outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, HI.

Andreas Stolcke. 1998. Entropy-based pruning of backoff language models. In *Proceeding of DARPA Broadcast News Transcription and Understanding Workshop*, pages 270–274, Lansdowne, VA.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286, Seattle.

Rui Wang, Hai Zhao, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita. 2014a. Neural network based bilingual language model growing for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 189–195, Doha, Qatar.

Xiaolin Wang, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2014b. Empirical study of unsupervised chinese word segmentation methods for smt on large-scale corpora. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 752–758, Baltimore, Maryland.

Rui Wang, Hai Zhao, Bao-Liang Lu, M. Utiyama, and E. Sumita. 2015. Bilingual continuous-space language model growing for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7):1209–1220.

Zhisong Zhang, Hai Zhao, and Lianhui Qin. 2016. Probabilistic graph-based dependency parsing with convolutional neural network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1382–1392, Berlin, Germany.