Resampling Approach for Instance-based Domain Adaptation from Patent Domain to Newspaper Domain in Statistical Machine Translation

Keisuke Noguchi noguchi@ai.cs.ehime-u.ac.jp
Department of Computer Science, Faculty of Engineering, Ehime University, 3 Bunkyo-cho, Matsuyama, Ehime, 790-8577, Japan

Takashi Ninomiya ninomiya@cs.ehime-u.ac.jp
Graduate School of Science and Engineering, Ehime University, 3 Bunkyo-cho, Matsuyama, Ehime, 790-8577, Japan

Abstract

In this paper, we investigate a resampling approach for domain adaptation from a resource-rich domain (patent domain) to a resource-scarce target domain (newspaper domain) in Statistical Machine Translation (SMT). We propose two resampling methods for domain adaptation in SMT: random resampling and resampling for instance weighting. The random resampling randomly adds sentence pairs from the resource-rich parallel corpus to the target-domain parallel corpus. Instance weighting is a method which provides a weight to each sample in the resource-rich domain. The problem of instance weighting in SMT is how to provide a weight to each sentence-pair. We approximate the instance weights by resampling sentence-pairs according to the ratio of sentence-pair probabilities between the two domains. We also explore a method of selecting samples that have instance weights larger than some threshold.

1 Introduction

In the last few decades, Statistical Machine Translation (SMT) has been widely studied in the field of machine translation because SMT is mathematically well-defined in terms of the probabilistic models it uses, and it can be learned automatically from large bilingual parallel corpora by using publicly available SMT tools. One of the key issues in SMT is how to develop a large parallel corpus. Developing a large-scale parallel corpus by hand is very expensive, and in the past, SMT systems were usually trained with only tens of thousands sentence-pairs. In the last decade, large-scale parallel corpora consisting of around millions of sentence pairs were developed by using the methods for automatically acquiring parallel sentence-pairs from comparable corpora (Utiyama and Isahara, 2003; Koehn, 2005; Callison-Burch et al., 2009). For example, a large-scale Japanese-English patent parallel corpus consisting of around 3 million sentence-pairs was automatically developed from patent documents (Utiyama and Isahara, 2003). The Europarl parallel corpus consists of around from 0.4 to 2 million sentence-pairs for each language pair, which were extracted from the proceedings of the European Parliament (Koehn, 2005). The French-English Gigaword (10^9) parallel corpus consists of around 22 million sentence-pairs crawled from Canadian and European Internet pages (Callison-Burch et al., 2009). These automatically acquired parallel corpora are large enough for SMT training if their domain and the target domain are the same. However, in general, the target domains of machine

translation are often different from the domains of the automatically acquired parallel corpora. It is empirically known that translation quality drastically deteriorates when an SMT system trained on one domain is applied to other domains (Foster et al., 2010). Therefore, domain adaptation is needed from the domains of these automatically acquired parallel corpora to the target domain.

This paper investigates a resampling approach for domain adaptation from the patent domain to the newspaper domain in SMT. We propose two resampling methods for domain adaptation in SMT: random resampling and resampling for instance weighting. The random resampling method randomly adds sentence pairs from the large-scale parallel corpus to the target-domain parallel corpus. Instance weighting (Jiang and Zhai, 2007) gives each sample a weight calculated by dividing the probability of the sample in the target domain by the probability of the sample in the large-scale data domain; i.e., the weight is an estimated frequency of the sample in the target domain. The problem of instance weighting in SMT is how to provide a weight to each sentence-pair. As SMT tools have many complex components, wrapper methods which do not modify the tools themselves are preferable. We approximate the instance weights by resampling sentence-pairs according to a ratio of sentence-pair probabilities between the two domains. We also explore a method of selecting samples that have instance weights larger than some threshold.

2 Domain Adaptation

Domain adaptation, which is also called transfer learning, is a method for adapting the model or data in a resource-rich domain to a resource-scarce target domain. The goal of domain adaptation is to increase the performance of systems in the resource-scarce domain by leveraging the model or data in the resource-rich domain. The resource-scarce domain is called the *target domain* or *in-domain*, and the resource-rich domain is called the *source domain* or *out-domain*. In what follows, we call the data/model in the resource-scarce domain the in-domain data/model and the data/model in the resource-rich domain the out-domain data/model.

2.1 Related Work

Domain adaptation can be classified into two types of adaptation: model adaptation and instance weighting (Foster and Kuhn, 2007; Jiang and Zhai, 2007). In the model adaptation, models are learned from the in-domain data and the out-domain data separately, and then a mixture model is developed by mixing the in-domain model and the out-domain model so that the mixture model performs better for the in-domain data. Foster and Kuhn (2007) proposed model adaptation for SMT. They used a log-linear model for the mixture model learned by using MERT (Och, 2003) with other log-linear parameters. They also proposed four distance metrics to measure the weight for each model; TF/IDF, LSA, perplexity, and EM . The mixture model learned by using MERT is also used as a baseline in (Foster et al., 2010). Moore and Lewis (2010) used perplexity for language model adaptation.

Instance weighting can be further classified into the metrics approach, weight optimization approach, and covariate shift approach. The metrics approach gives each sample a weight representing the distance between the sample as an in-domain sample and the sample as an out-domain sample. There have been several studies on the metrics approach in SMT, wherein the metrics used were the cross-entropy (Yasuda et al., 2008) or cross-entropy difference (Axelrod et al., 2011, 2012) for a sentence-pair. In the metrics approach, sentence-pairs in the out-domain corpus are selected if their metrics is closer than some threshold. Yasuda et al. (2008) also used model adaptation in addition to the metrics approach. In the weight optimization approach, a weight is assigned to each sentence/phrase in the training corpus, and these weights are optimized so as to maximize the objective function for the in-domain development

corpus (Matsoukas et al., 2009; Foster et al., 2010). Matsoukas et al. (2009)'s model was originally proposed for data selection, not for domain adaptation, but it can also be applied to domain adaptation. The covariate shift approach (Shimodaira, 2000; Jiang and Zhai, 2007; Xia et al., 2013, 2014) gives each sample a weight calculated by dividing the in-domain probability of the sample by the out-domain probability of the sample; i.e., the weight is an estimated frequency of the sample in the target domain. Most of the previous studies on instance weighting directly incorporate the weights into the objective function or select the samples having weights higher than some threshold. Our method is also based on instance weighting, but our method is designed for SMT and uses resampling for estimating the probability ratio.

The most similar approach to our method is that of Gascó et al. (2012). They proposed the resampling method for domain adaptation for SMT. The difference between our method and theirs is that we use the covariate shift approach using both in-domain probabilities and out-domain probabilities, whereas they use only the in-domain probabilities.

2.2 Instance Weighting for SMT

Given parallel sentence pairs $(s_i, t_i)_{i=1}^N$ as a training data set, where s_i is a sentence in the source language and t_i is a sentence in the target language, the parameter estimation for SMT in the maximum likelihood estimation framework is defined as follows.

$$\begin{aligned}\hat{\theta} &= \arg\max_\theta \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}} p(s,t)\log p(t|s;\theta) \\ &\approx \arg\max_\theta \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}} \tilde{p}(s,t)\log p(t|s;\theta) = \arg\max_\theta \sum_{i=1}^{N}\log p(t_i|s_i;\theta)\end{aligned} \tag{1}$$

where $\mathcal{S}$ is the source language, $\mathcal{T}$ is the target language, and $\tilde{p}$ is an empirical distribution. The instance weighting is derived as follows (Jiang and Zhai, 2007).

$$\begin{aligned}\hat{\theta} &= \arg\max_\theta \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}} p_{in}(s,t)\log p(t|s;\theta) \\ &= \arg\max_\theta \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}} \frac{p_{in}(s,t)}{p_{out}(s,t)} p_{out}(s,t)\log p(t|s;\theta) \\ &\approx \arg\max_\theta \sum_{i=1}^{N} \frac{p_{in}(s_i,t_i)}{p_{out}(s_i,t_i)}\log p(t_i|s_i;\theta)\end{aligned} \tag{2}$$

where p_{in} is the in-domain probabilities and p_{out} is the out-domain probabilities. Instance weighting gives each sentence pair a weight calculated by dividing the in-domain probability of the sentence pair by the out-domain probability of the sentence pair; i.e., the weight is an estimated in-domain frequency of the sentence pair.

3 Random Resampling for SMT

In domain adaptation, there are two simple and strong baseline methods: one which trains a model only on the in-domain data set and one which trains a model on the union of the in-domain and out-domain data sets (Daume III, 2007). The baseline method using both data sets is strong, but in our experiments it was worse than the baseline method using only the in-domain data set. We think that this deterioration is caused by adding too many data samples from the out-domain data set to the in-domain data set.

To solve the problem of the baseline methods mentioned above, we introduced a random resampling method (He and Garcia, 2009) for domain adaptation in SMT. Resampling was originally studied to solve the imbalanced data problem in binary classification, but it can also be applied to domain adaptation. There are two methods for random resampling: oversampling and undersampling (He and Garcia, 2009). Both methods use both an in-domain corpus and out-domain corpus as a training data set by default. Oversampling increases the size of the in-domain corpus by copying randomly selected sentences in the in-domain corpus. It may add the same sentences in the in-domain corpus many times but it increases the weights for the in-domain sentences in the objective function compared to the out-domain sentences. Undersampling decreases the weights for the out-domain corpus by removing randomly selected sentences in the out-domain corpus. In the experiments, we used undersampling as the random resampling method.

4 Resampling for Instance Weighting in SMT

We propose a resampling method for instance weighting in SMT. We approximate the probability ratio in Equation 2 with in-domain and out-domain language models as follows.

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \frac{p_{in}(s_i, t_i)}{p_{out}(s_i, t_i)} \log p(t_i|s_i;\theta) \approx \arg\max_{\theta} \sum_{i=1}^{N} \frac{p_{in}(t_i)}{p_{out}(t_i)} \log p(t_i|s_i;\theta) \quad (3)$$

where $p_{in}(t_i)$ is an in-domain language model and $p_{out}(t_i)$ is an out-domain language model. Both language models are defined as n-gram language models.

In the experiments, we used 5-gram models with Kneser-Ney smoothing for the n-gram language models. 5-gram models with Kneser-Ney smoothing was learned by using the SRILM tool kit (Stolcke et al., 2011). We calculate the probability of a sentence $w_1 w_2 \ldots w_n$ as follows.

$$p(w_1 w_2 \ldots w_n) \approx \prod_{j=1}^{n} p(w_j|w_{j-4} w_{j-3} w_{j-2} w_{j-1}) \quad (4)$$

where $p(w_j|w_{j-4} w_{j-3} w_{j-2} w_{j-1})$ represents 5-gram probabilities. Given a sentence t, $p_{in}(t)$ and $p_{out}(t)$ are calculated by using 5-gram models learned from the in-domain parallel corpus and the out-domain parallel corpus, respectively.

Given a sentence t, let $w(t)$ be the weight $p_{in}(t)/p_{out}(t)$ in Equation 3. The weight $w(t)$ for sentence t represents the in-domain frequency of t. In our resampling method, a sentence pair (s, t) in the out-domain corpus is selected with the probability of $w(t)$ if $w(t)$ is less than 1. A sentence pair (s, t) is always selected if $w(t) \geq 1$. Theoretically, a sentence pair should be resampled multiple times if $w(t)$ is greater than 1, but we resample the sentence only once because $w(t)$ can be an extremely large number. Formally, we have the modified resampling number $w'(t)$ for sentence t as follows.

$$w(t) = \frac{p_{in}(t)}{p_{out}(t)}, \qquad w'(t) = \begin{cases} w(t) & \text{if } w(t) < 1 \\ 1 & \text{otherwise} \end{cases} \quad (5)$$

In the same way as the previous methods that select sentences using thresholds, we also experimented with the sample selection approach by using thresholds and the weight $w(t)$ in Equation 5. In this approach, a sentence pair (s, t) is selected if the weight $w(t)$ is greater than a threshold. All sentence pairs that satisfy this condition in the out-domain parallel corpus are selected.

Table 1: Specification of English-Japanese parallel corpora

domain	training set (# of sentence-pairs)	development set (# of sentence-pairs)	test set (# of sentence-pairs)
patent	3,166,284	-	-
newspaper	130,000	500	10,000

Table 2: Results

	BLEU (%)
baseline (in-domain)	13.93
baseline (in-domain + out-domain)	12.67
random resampling	14.24
instance weighting (resampling)	14.15
instance weighting (w. thresholds)	14.47

5 Experiments

We evaluated the performance of the random resampling method and the resampling method for instance weighting in English-Japanese SMT.

5.1 Settings

In the experiments, we regarded the newspaper domain as the target domain (in-domain) and the patent domain as the resource-rich domain (out-domain). We used a English-Japanese patent parallel corpus consisting of 3,166,284 sentence pairs; the same one was used for the shared task in NTCIR10 (PatentMT) (Goto et al., 2013). We also used the English-Japanese newspaper parallel corpus, JENAAD (Utiyama and Isahara, 2003). JENAAD consists of 150,000 sentence pairs extracted from the comparable corpora: the Yomiuri Shimbun and the Daily Yomiuri. Table 1 shows the details of the data sets. The language models were learned as 5-gram models with Kneser-Ney smoothing by using the SRILM tool kit (Stolcke et al., 2011). The in-domain language model was learned from the training corpus in the newspaper parallel corpus, and the out-domain language model was learned from the training corpus in the patent parallel corpus.

We used GIZA++ 1.0.7 for word alignment, SRILM 1.5.12 for learning n-gram language models, Moses for SMT (Koehn et al., 2007) and MERT for tuning. The value of distortion limit was infinite. We used Mecab 0.98 with ipadic 2.7.0 for tokenizing Japanese sentences. We used BLEU for measuring translation accuracy. We evaluated SMT only in a direction from English to Japanese.

In the experiments, we also evaluated the SMT trained with only the in-domain parallel corpus and the SMT trained with the union of the in-domain parallel corpus and the out-domain parallel corpus as baseline methods.

5.2 Results

Table 2 shows the experimental results. “baseline (in-domain)” indicates the baseline method using only the in-domain corpus. “baseline (in-domain + out-domain)” indicates the baseline method trained on the union of the in-domain corpus and the out-domain corpus. In this experiment, we used undersampling in the random resampling method. “instance weighting (resampling)” means the resampling method for instance weighting. “instance weighting (w. threshold)” means the instance weighting method that selects the sentence pairs having larger

Table 3: Change in BLEU with the size of resampled data

	# of added sentence pairs	random resampling (BLEU (%))	instance weighting (w. threshold) (BLEU (%))
baseline (in-domain)	0	13.93	-
resampling	5,525	13.74	-
	10,000	13.73	14.47
	20,000	14.10	13.50
	40,000	13.51	14.01
	80,000	14.24	13.65
	130,000	13.58	14.06
	500,000	13.22	13.44
	1,000,000	12.83	12.77
	2,000,000	12.92	12.76
baseline (in-domain + out-domain)	3,166,284	12.67	-

weights than a threshold[1][2]. The resampling method for instance weighting selected 5,525 sentence pairs from the out-domain parallel corpus. As seen in the table, the random resampling and instance weighting achieved better BLEU scores than the baseline methods. The instance weighting with thresholds achieved the best BLEU, but the resampling method for instance weighting also achieved comparative results, using fewer out-domain sentence pairs. The random resampling method was also comparable to instance weighting with thresholds.

Table 3 shows the change in BLEU with the number of added sentence pairs in the random resampling method and the instance weighting method with thresholds. In the results of the random resampling method, the best result was achieved with 80,000 sentence pairs, and BLEU increased by 0.31 from the baseline. In the results of the instance weighting method with thresholds, the best result was achieved with 10,000 sentence pairs, and BLEU increased by 0.54 from the baseline. The number of sentence pairs selected by the resampling method for instance weighting, 5,525 sentence pairs, was close to the number of sentence pairs selected by the instance weighting method with thresholds. This means that the resampling method for instance weighting provided a good estimation for the weights without brute-force searching for the threshold.

6 Conclusion

We investigated a resampling approach for domain adaptation from the patent domain to the newspaper domain in SMT. The random resampling method randomly selects sentence pairs from the out-domain parallel corpus. The resampling method for instance weighting selects sentence pairs according to the ratio of sentence-pair probabilities between the two domains. Instance weighting selects the sentence pairs that are likely to be the in-domain sentence pairs from the out-domain parallel corpus. We also explored instance weighting with thresholds, which selects all the sentence pairs having higher weights than a threshold. In this study, n-gram language models were used for calculating the in-domain and out-domain probabilities

[1] As an implementation issue, we first sorted the sentence pairs in the out-domain corpus in descending order of their weights. Then we selected the sentence pairs which had high weights.

[2] However, in the experiments, the number of added sentences in the random resampling method and the thresholds in instance weighting with thresholds were wrongly determined by using the test set. So, the true BLEU scores of the random resampling method and instance weighting with thresholds might be lower than the BLEU scores in Table 2.

for sentence pairs. In the experiments, instance weighting with thresholds achieved the best results, but both random resampling and resampling for instance weighting also achieved comparable results. Though the BLEU score of the resampling method for instance weighting was worse than other two methods, the number of resampled sentence pairs was very close to that of the instance weighting method with thresholds. The advantage of the resampling method for instance weighting is that it does not require tuning for finding the thresholds while other methods need a tuning process.

Acknowledgement

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (B) Grant Number 25280084.

References

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Axelrod, A., Li, Q., and Lewis, W. (2012). Applications of data selection via cross-entropy difference for real-world statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2012)*, pages 201–208. Internaltional Workshop on Spoken Language Translation (IWSLT).

Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT 2009)*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Daume III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Foster, G., Goutte, C., and Kuhn, R. (2010). Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 451–459, Cambridge, MA. Association for Computational Linguistics.

Foster, G. and Kuhn, R. (2007). Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT 2007)*, pages 128–135, Prague, Czech Republic. Association for Computational Linguistics.

Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J., and Casacuberta, F. (2012). Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 152–161, Avignon, France. Association for Computational Linguistics.

Goto, I., Lu, B., Chow, K. P., Sumita, E., and Tsou, B. K. (2013). Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In *Proceedings of the 10th NTCIR Conference*, pages 260–286.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, 21(9):1263–1284.

Jiang, J. and Zhai, C. (2007). Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 264–271, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Machine Translation Summit (MT Summit 2005)*, pages 79–86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matsoukas, S., Rosti, A.-V. I., and Zhang, B. (2009). Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 708–717, Singapore. Association for Computational Linguistics.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

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

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244.

Stolcke, A., Zheng, J., Wang, W., and Abrash, V. (2011). SRILM at sixteen: Update and outlook. In *Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*.

Utiyama, M. and Isahara, H. (2003). Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 72–79, Sapporo, Japan. Association for Computational Linguistics.

Xia, R., Hu, X., Lu, J., Yang, J., and Zong, C. (2013). Instance selection and instance weighting for cross-domain sentiment classification via pu learning. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI '13)*, pages 2176–2182. AAAI Press.

Xia, R., Yu, J., Xu, F., and Wang, S. (2014). Instance-based domain adaptation in NLP via in-target-domain logistic approximation. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1600–1606, Québec City, Québec, Canada.

Yasuda, K., Zhang, R., Yamamoto, H., and Sumita, E. (2008). Method of selecting training data to build a compact and efficient translation model. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 655–660. Association for Computational Linguistics.