Effect of the Topic Dependent Translation Models for Patent Translation - Experiment at NTCIR-7

Takeshi ITO[†] Tomoyosi AKIBA[†] Katunobu ITOU[‡]
[†]Toyohashi University of Technology, [‡]Hosei University
[†]1-1 Tenpaku, Toyohashi, Aichi 441-8580, Japan
[†]{t-ito, akiba}@cl.ics.tut.ac.jp

Abstract

In this paper, we investigate the effect of the topic dependent translation model for patent translation. We employ clustering technique to estimate topics in the training corpus and document retrieval to identify the topic fitting to the source sentence. In our experimental evaluation, we investigate the contribution of our topic dependent models to phrase-base Statistical Machine Translation.

Keywords: *Statistical Machine Translation, Clustering, Topic Adaptation, Topic Dependent Translation*

1 Introduction

Recently, Statistical Machine Translation(SMT) becomes mainstream in machine translation research, and phrase-base SMT performs well among the methods. It is well known that using larger training data produces better performance in SMT. However, there are often more than one topics in large training data and leveraging such topics in the large training data should improve the translation performance. We focus on the topics in SMT and propose the topic dependent translation method for Patent Translation.

The idea of topic adaptation is widely used in many fields. For example, the topic adapted language model performs well in automatic speech recognition. Our method can be considered as applying the topic adaptation into the statistical machine translation.

For training the topic dependent translation models, we first divide the parallel sentences into the topic dependent clusters, then the sentences in each cluster are used to train the translation model of the topic that they belong to. At the time of translation, the topic of the source sentence is predicted by applying a document retrieval method, then the sentence is translated by using the translation model that corresponds to the predicted topic.

2 Training Method

In this section, the training procedure is explained.

A topic dependent translation model can be trained by using a subset of parallel sentences, that shares a common topic. However, it is not clear what and how many topics are in the training parallel sentences. Therefore, unsupervised topic clustering technique is applied.

Firstly, by applying the unsupervised clustering method, the documents in the PPD are divided into the clusters, each of which is considered to be shared some topic. The number of the clusters to be divided is given beforehand. Secondly, for each cluster, the the parallel sentences in the PSD that are extracted from the documents in the cluster are gathered into the topic specific training data for the topic dependent translation model.

The flow of the training process is shown on the left side of Figure 1. The detail of the process will be described below step by step.

1. The documents in the Japanese side of the PPD are classified into the fixed number of clusters by using the clustering toolkit CLUTO [7]. The bag of normalized content words are used as the feature vector of a document. Among the CLUTO's various parameters, we used the following parameter settings. Cosine is used for measuring the similarity between documents. The 2-way clustering is used for the clustering algorithm, in which a cluster is divided in two repeatedly until the given number of clusters are obtained.
2. The PSD are divided into the clusters according to the PPD cluster.
3. For each PSD cluster, the topic-dependent translation model is trained by using the parallel sentences that belongs to the cluster.

3 Translation Method

In this section, the translation procedure of the proposed method is explained.

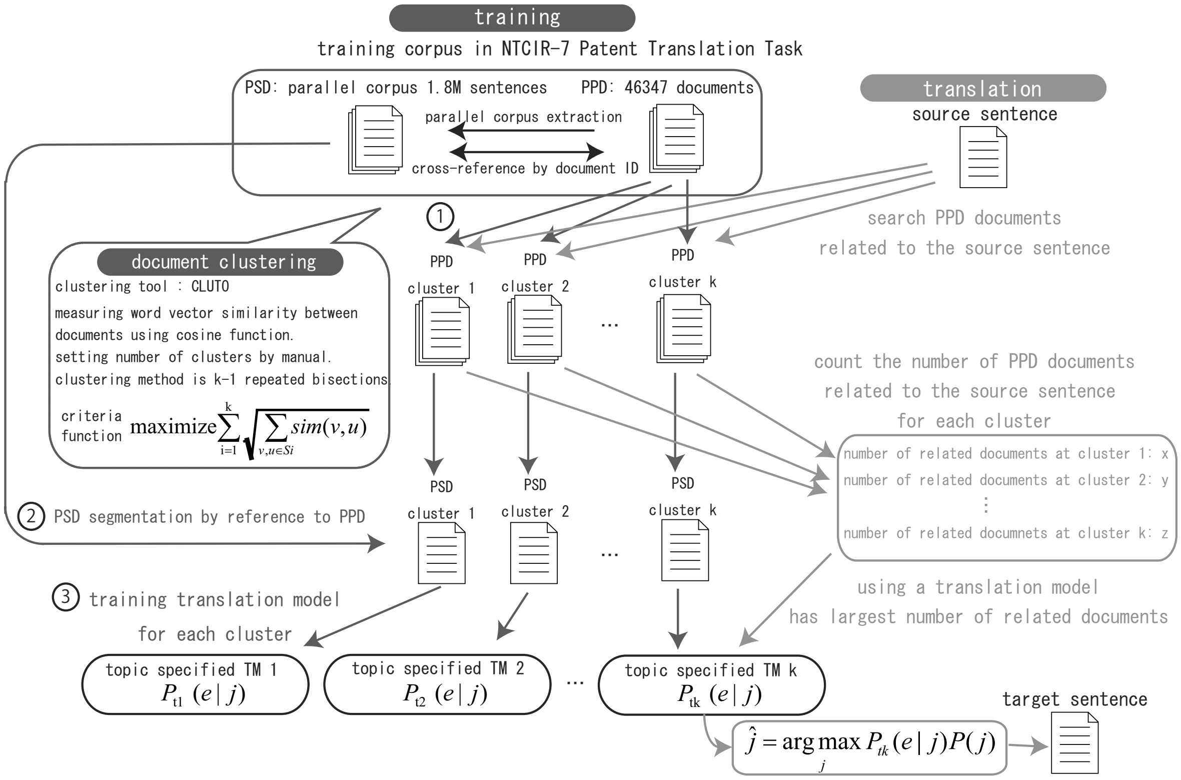
training
training corpus in NTCIR-7 Patent Translation Task
PSD: parallel corpus 1.8M sentences
PPD: 46347 documents
parallel corpus extraction
cross-reference by document ID
translation
source sentence
search PPD documents
related to the source sentence
document clustering
clustering tool : CLUTO
measuring word vector similarity between documents using cosine function.
setting number of clusters by manual.
clustering method is k-1 repeated bisections
criteria function $\text{maximize}\sum_{i=1}^{k}\sqrt{\sum_{v,u\in Si} sim(v,u)}$
①
PPD cluster 1
PPD cluster 2
PPD cluster k
count the number of PPD documents
related to the source sentence
for each cluster
number of related documents at cluster 1: x
number of related documents at cluster 2: y
number of related documnets at cluster k: z
② PSD segmentation by reference to PPD
PSD cluster 1
PSD cluster 2
PSD cluster k
③ training translation model
for each cluster
using a translation model
has largest number of related documents
topic specified TM 1 $P_{t1}(e|j)$
topic specified TM 2 $P_{t2}(e|j)$
topic specified TM k $P_{tk}(e|j)$
target sentence
$\hat{j} = \arg\max_{j} P_{tk}(e|j)P(j)$

Figure 1. Training and translation process flow.

To apply the topic dependent translation model to the input sentence, its topic must be predicted at first. We predict the topic by finding the most similar cluster to the input sentence among the clusters constructed at the training phase. It is done by applying a document retrieval method, so that the input sentence are used as the query for the document retrieval targeting the documents in PPD. For each document in the n-best results, the corresponding cluster that it belongs to is checked and counted. The majority among the clusters is selected as the predicted topic of the input sentence.

Then, the input sentence is translated by using the translation model that corresponds to the predicted topic.

We used GETA for the document retrieval. The content words are used for indexing the documents. The TF-IDF with pivoted normalization is used for the term weighting.

4 Evaluation

In the experimental evaluation, we used Moses[4] for translation model training and decoding, SRILM[6] for language model training, CLUTO for clustering, GETA for retrieval. We did not apply the minimum error rate model tuning.

Training data in NTCIR-7 Patent Translation Task is used for training. The number of sentences in PSD is 1798571. The number of documents in PPD is 46347.

In USPTO patents which English PPDs are extracted from, topic information is annotated according to international patent classification(IPC). However, we did not use such information.

Our evaluation results on NTCIR-7 Patent Translation Task formal-run are shown in Table 1. For each direction, i.e. E-J or J-E, we submitted two runs; the parallel sentences in PSD was divided into 5 (TH-je1, TH-ej5) or 10 (TH-je2, TH-ej6) clusters. 10-best retrieval results are used to select the cluster at the translation time.

Unfortunately, we found that the system that had been used to obtain the formal-run results did not work correctly because of several mistakes in the programming. The size of the parallel sentences actually used for training the model was happened to be only 20K for each cluster. Moreover, we wrongly used the inconsistent case representation for words between training and testing for the J-E task.

After all, we corrected the errors in the system, then evaluated our methods again. The results are shown in Table 2.

For the baseline methods, **Baseline1** trains the model by using all the PSD training data and translates the input sentences using it, while **Baseline2** trains the model by using 640,000 sentences, which size is decided to be matched with the maximum number of the

Table 1. Evaluation result for formal-run test data of NTCIR-7 Patent Translation Task. ("intrinsic" result in the task, group ID is "TH")

Japanese-English Translation			
run ID	cluster number	average sentences number per cluster	BLEU
TH-je1	5	339843	15.9
TH-je2	10	179823	14.86

English-Japanese Translation			
run ID	cluster number	average sentences number per cluster	BLEU
TH-ej5	5	339843	2.23
TH-ej6	10	179823	2.32

Table 2. Corrected evaluation result for formal-run test data of NTCIR-7 Patent Translation Task.

Japanese to English Translation	average sentences number per cluster	BLEU
Baseline1(1 cluster)	1798571	23.96
Baseline2(1 cluster)	640000	23.27
Cluster-10	179823	23.29
Cluster-10-oracle	197823	28.85
Cluster-5	339843	23.52
Cluster-5-oracle	339843	27.50

English to Japanese Translation	average sentences number per cluster	BLEU
Baseline1(1 cluster)	1798571	29.80
Baseline2(1 cluster)	640000	28.66
Cluster-10	179823	29.29
Cluster-10-oracle	179823	35.35
Random-10-oracle	179857	35.44
Cluster-5	339843	29.71
Cluster-5-oracle	339843	34.28
Random-5-oracle	359714	34.68

sentences in our **Cluster-5** models.

For our proposed methods, **Cluster-5** divides PSD into 5 clusters, while **Cluster-10** divides PSD into 10 clusters. 10-best retrieval results are used to select the cluster at the translation time.

For the reference methods, **Cluster-5-oracle** and **Cluster-10-oracle** use the same clusters as **Cluster-5** and **Cluster-10**, respectively, but we manually select the optimal cluster that draws the best BLEU score at the translation time.

Comparing with the baseline, our methods (**Cluster-5** and **Cluster-10** for both directions) give better BLEU score than **Baseline2**, but lower than **Baseline1**. However, oracle methods (**Cluster-5-oracle** and **Cluster-10-oracle** for both directions) consistently give better score than the baseline methods.

In order to see if the results should indicate that some appropriate cluster selection method can improve the performance or the good results would be obtained simply by the oracle method, we conducted another experiment.

Random-5-oracle and **Random-10-oracle** *randomly* divide PSD into 5 and 10 clusters, respectively, then the optimal cluster is selected manually at the translation time. The results shows that the randomized oracle methods performs as well as the oracle methods, indicating that the oracle methods improve the performances by themselves.

To improve the results, we further applied the following modifications into our methods. Firstly, we increased the n-best retrieval results from 10 to 50 used to select the cluster at the translation time. Secondly, we changed the similarity measure used to select the similar sentence in the clusters. However, both methods did not improve the results.

5 Discussion

The evaluation showed that the our topic adapted translation model was not so effective. To see the reason why it did not work well, we investigated the phrase translation model in detail.

We think that the topic adapted phrase translation model $P_a(t|s)$ works well when it can reduce the possible translation candidates t given a source phrase s by the topic adaptation. To investigate the assumption, we calculate the perplexity of the phrase translation model $P(t|s)$ defined as follows.

$$Perplexity(s) = 2^{-\sum_t P(t|s) log_2 P(t|s)}$$

The perplexity represents the average number of the target candidates given a source phrase s.

For the phrase translation table trained by using all the PSD, we calculated the average of the perplexities for each length of the source phrases. The results are shown in Table 3. It shows that the average perplexity decreases as the length of the source phrase increases, and that the phrases longer than four words have less than two candidates in average. This indicates that, for longer source phrases, the phrase translation model works almost deterministic to select the candidate target phrase. It seems to give a reason why the topic adaptation for the phrase translation model does not work well.

Table 3. The average of the perplexity for each length of the source phrases.

English to Japanese translation		
number of words in a phrase	number of phrases	average number of branch
1	122127	6.02
2	1594698	4.11
3	5146112	2.61
4	7302523	1.90
5	6845443	1.58
6	5096549	1.42
7	3352540	1.33

Japanese to English translation		
number of words in a phrase	number of phrases	average number of branch
1	59382	7.29
2	996557	4.35
3	3795944	2.7
4	6292354	2.01
5	7155370	1.69
6	6666304	1.52
7	5567892	1.41

On the other hand, for the word translation model, the topic adaptation seems to work well, as the word candidates vary with the topic adapted word translation model. For example, Table 4 shows the most probable translation candidates of the source word “blade” predicted by the cluster specific word translation models used in our **Cluster-5** English-Japanese setting. However, as the phrase translation model is a dominant component in the current phrase-based SMT framework, the topic adapted translation model seems not to work well in total.

6 Related Work

Utiyama and Isahara [2] trained the topic dependent translation models from the PSD by using the international patent classification (IPC) associated with the PPD. They reported that their topic dependent models are less effective than the model trained by using all the PSD. Their result is consistent with ours.

Yamamoto and Sumita [3] also applied a domain adaptation method for both translation and language models on the corpus for travel arrangements task. Their approach is similar to ours, but their clustering and domain selection methods are quite different from ours. They reported that their method was effective, though they used the different phrase-based SMT decoder, namely Pharaoh [5].

Table 4. Difference of word translation probability between clusters at English to Japanese translation.

	English	Japanese	word probability
cluster 0	blade	羽根	0.24
cluster 1	blade	ブレード	0.26
cluster 2	blade	刃	0.33
cluster 3	blade	羽根	0.22
cluster 4	blade	ブレード	0.34

7 Conclusion

In this paper, we proposed the topic adaptation method for the translation model used in the SMT system. We investigated the effect of the proposed adaptation method for the NTCIR-7 Patent translation task. In the future work, we will try other adaptation techniques, including model interpolation method, in order to improve the translation accuracy.

References

[1] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, ”Overview of the Patent Translation Task at the NTCIR-7 Workshop”, Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2008.

[2] Masao Utiyama, Hitoshi Isahara, ”A Japanese-English Patent Parallel Corpus”, MT summit XI, pp. 475-482, 2007.

[3] Hirofumi Yamamoto, Eiichiro Sumita, ”Bilingual Cluster Based Models for Statistical Machine Translation”, PRoc. of Conference on Empirical Methods in Natural Language Processing(EMNLP), pp.514-523, June 2007.

[4] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst, ”Moses:Open source toolkit for statistical machine translation”, ACL 2007 demonstration session, pp.177-180, 2007.

[5] P. Koehn, ”PHARAOH: A beam search decoder for phrase-based statistical machine translation models”, http://www.isi.edu/publications/licensed-sw/pharaoh/

[6] A. Stolcke, ”SRILM - an extensible language modeling toolkit”, ICSLP, pp.901-904, 2002.

[7] George Karypis, ”CLUTO - A Clustering Toolkit”, Technical Report 02-017, Dept. of Computer Science, University of Minnesota, 2002. Available at http://www.cs.umn.edu/cluto.