System Description of NiCT-ATR SMT for NTCIR-7

Keiji Yasuda[†,‡] Andrew Finch[†,‡] Hideo Okuma[†,‡]
Masao Utiyama[†] Hirofumi Yamamoto[†,‡,*] Eiichiro Sumita[†,‡]
[†]National Institute of Communications Technology
{keiji.yasuda,andrew.finch,hideo.okuma
mutiyama,eiichiro.sumita}@nict.go.jp
2–2–2, Hikaridai, Keihanna Science City, 619–0228, Japan

[‡]ATR Spoken Language Translation Research Laboratories
2–2–2, Hikaridai, Keihanna Science City, 619–0228, Japan

[*]Kinki University School of Science and Engineering Department of Information
3–4–1, Kowakae, Higashiosaka city, Osaka, 577–8502 Japan
yama@info.kindai.ac.jp

Abstract

In this paper we propose a method to improve SMT based patent translatioin. This method first employs International Patent Classification to build class based models. Then, multiple models are interpolated by weighting method employing source side language models. We carried out experiments using data from the patent translation task of NTCIR-7 workshop. According to the experimental results, the proposed method improved the most of automatic scores, which were NIST, WER and PER. Experimental results also shows BLUE score degradation in the proposed method. However, statistical tests by bootstrapping does not show significance for the degradation.
Keywords: *IPC, domain adaptation.*

1 Introduction

Current machine translation (MT) research shows the effectiveness of a copus-based machine translation framework [6]. MT system using the frame work such as Statistical Machine Translation (SMT) [1] is thought to be convenient technology because of rapid and mostly automated MT system building.

For SMT research, parallel corpora are one of the most important components. There are mainly two factors of parallel corpora contributing to system performance. The first is the quality of the parallel corpus, and the second is its quantity. A parallel corpus that has similar statistical characteristics to the target domain should yield a more efficient models. However, domain mismatched training data might reduce the model's performance. And sufficiently size of parallel corpus solves the data sparseness problem of model training.

Meanwhile from a point of view of commercial, it is more important to create consumer-demanded MTs, considering language pair and target domain of MT use than parallel corpus availability.

Considering all of the previously mentioned points, Japanse-English patent translation is one of the most interesting SMT research fields which satisfies these points. A large-sized Japanese-English patent parallel corpus has just been released [13]. Commercial demand is also very high for both directions of Japanese-English patent translation. Additinally, an SMT evaluation campaign using the parallel corpus is ongoing [3]. This boosts related technology and helps information exchange.

The research shown in this paper deals with Japanese-to-English patent translaion. The proposed method in this paper uses International Patent Classification (IPC) to improve the SMT based patent translation system. IPC is used to build class-based models. Then, multiple class based models and a general model are interpolated by weighting method employing source side language models.

Section 2 explains IPC. Section 3 describes the method of using IPC information for SMT. Section 4 details the experimental setting and results. Section 5 discusses the comparison of the propsed method to related work. Section 6 concludes the paper.

2 Internatinal Patent Classification

The International Patent Classification (IPC), established by the Strasbourg Agreement 1971, provides for

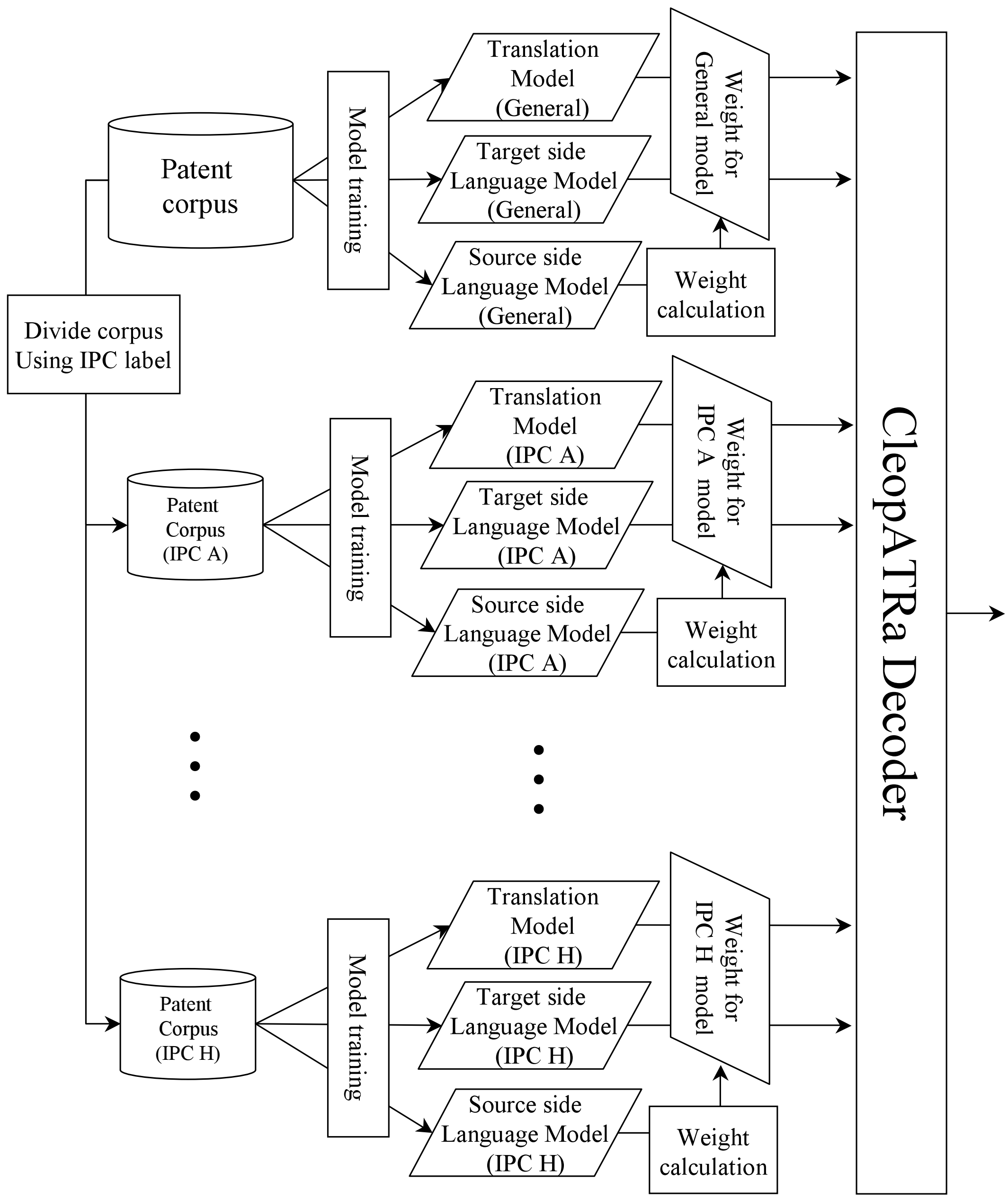

Figure 1. Framework of the proposed method.

a hierarchical system of language independent symbols for the classification of patents and utility models according to the different areas of technology to which they pertain. Each section has a title and a symbol. The title consists of one or more words and the symbol consists of a capital letter of the Roman alphabet. They are as follows:

A Human Necessities
B Performing Operations; Transporting
C Chemistry; Metallurgy
D Textiles; Paper
E Fixed Constructions
F Mechanical Engineering; Lighting; Heating; Weapons; Blasting
G Physics
H Electricity

In this research, we only use above information from IPC, i.e., the top layer of the IPC hierarchy.

3 Proposed method

Figure 1 shows the flow of the proposed method. In the proposed method, first we train general models by using all available parallel corpus. Three kinds of models are trained here: a translation model, a target-side language model, and a source-side language model.

Secondly, we divide the original patent parallel corpus into eight subcorpora using IPC label. Then, three models (IPC based models) are trained for each subcorpus.

In the proposed method, source side language models are used to calculate weights for general and IPC based models. Detailed weight calculation formulas are shown in Section 4.2.

4 Experiments

4.1 Phrase-based SMT

We employed a log-linear model as a phrase-based statistical machine translation framework [5]. This model expresses the probability of a target-language word sequence (e) of a given source language word sequence (f) given by:

$$P(e|f) = \frac{\exp\left(\sum_{i=1}^{M} \lambda_i h_i(e, f)\right)}{\sum_{e'} \exp\left(\sum_{i=1}^{M} \lambda_i h_i(e', f)\right)} \quad (1)$$

where $h_i(e, f)$ is the feature function, such as the translation model or the language model, λ_i is the feature function's weight, and M is the number of features. λ_i is tuned by using the Minimum Error Rate Training (MERT) algorithm [9] on a development set.

We can approximate Eq. 1 by regarding its denominator as constant. The translation results ($\hat{e}$) are then obtained by

$$\hat{e}(f) = \arg\max_{e} \exp\left\{\sum_{i=1}^{M} \lambda_i h_i(e, f)\right\} \quad (2)$$

4.2 Weight Calculation

As mentioned in Section 3, features (models) are trained for each IPC. Additionally, λ_i is also tuned for each IPC.

Addition to the feature's weight (λ_i) shown in Eq. 2, we also need to compute weight for each IPC-based model (μ) for each given input sentence. By using μ, Eq. 2 is reformulated as follows

$$\begin{aligned} &\hat{e}(f) \\ &= \arg\max_{e} \exp\left\{\sum_{i=1}^{M} \sum_{j=A}^{IPC} \lambda_{i,j} \cdot \mu_j \cdot h_{i,j}(e, f)\right\} \end{aligned} \quad (3)$$

where

$$IPC \in \{A, B, C, D, E, F, G, H, General\} \quad (4)$$

and μ is calculated by the following formula

$$\mu_j = \frac{P(j|S_{input})}{\sum_{k=A}^{IPC} P(k|S_{input})} \quad (5)$$

Here, $P(j|S_{input})$ is the probability of input sentence (S_{input}) belonging to the IPC j. Using the source-side language model of IPC i, $P(j|S_{input})$ is calculated by the following formula:

$$\begin{aligned} &P(j|S_{input}) \\ &= P(S_{input}|j) \times \frac{P(j)}{P(S_{input})} \end{aligned} \quad (6)$$

where $P(S_{input}|j)$ is the sentence probability of the input sentence on the source side language model of IPC j. $P(j)$ is the category probability which is a sentence ratio of the training subcorpus of IPC j to the full-sized corpus.

In the proposed method, values of μ are calculated for each input. Our in-house decoder, which is CleopATRa can handle multiple models with changing μ sentence by sentence.

4.3 Experimental Setting

We used training set from NTCIR-7 workshop patent translation task [3] for the experiments. A development set and a test set were also provided data by the workshop. Details of this data is shown in Table 1.

Table 1. Details of data for the experiments (# of sentences)

	IPC								ALL (General)
	A	B	C	D	E	F	G	H	
Training set	58.3 K	271.4 K	41.0 K	10.6 K	6.8 K	161.0 K	1122.1 K	751.9 K	2423.2 K
Dev. set	14	75	16	1	1	54	489	265	915
Test set	19	64	11	4	8	58	469	266	899

For the statistical machine-translation experiments, we segmented Japanese words using Japanese morphological analyzer ChaSen [7]. Then, we used the preprocessed data to train the phrase-based translation model by using GIZA++ [10] and the Moses tool kit [4]. Source-side and target-side language model are trained by SRI language model toolkit [12]. The language model configuration is a modified Kneser-Ney [2] 4-gram.

4.4 Experimental Results

Table 2 shows the experimental results. This table shows the evaluation results of the baseline and the proposed method. The evaluation is done by several automatic metrics, BLEU[11], NIST[8], WER and PER. In this table, the better score is underlined. As shown in the table, the BLEU score shows degradation of the MT performance using the proposed method. Meanwhile, all of the other metrics shows improvement using the proposed method.

To test the significance of the scores improvement and degradation, we carried out MT evaluation bootstrapping [14] with a 1000 times sampling. In the table, if there is significant difference between the baseline and the proposed method, boldface numbers are used to express the significantly better score. Looking at the table, BLEU score degradation by the proposed method is not significant. Meanwhile, all of the other scores' improvement by the proposed method is significant.

5 Discussions

A Japanese-English patent parallel corpus was just released in 2007 [13]. Thus, there is not much SMT-related research on English-to-Japanese patent translation. However, [13] carried out experiments using the aforementioned corpus and IPC information. The experiments simply used the training corpus in the same IPC as an input sentence.

According to their results, that method causes degradation of the BELU score compared to a baseline method which is similar to our baseline method. They only used the BLEU score for the evaluation, and the score degradation is around 0.78% to 2.02%, which is larger than our results (0.33%). The baseline performance of the experiments is lower than our experiments, thus the relative score degradation of their method is even higher than ours. Considering this point, our method is thought to work better than their method.

6 Conclusions

We proposed a method of using IPC information for SMT based patent translation. This method uses IPC section information to train class based translation and language models. Then, multiple class based models and a general model are interpolated by weighting method employing source side language models.

We carried out experiments using data from the NTCIR-7 workshop patent translation task. The experimental results indicated that our method improved most of automatic scores, which are NIST, WER and PER. Although the proposed method caused BLUE score degradation, the statistical test using a bootstrapping method does not show significant difference.

References

[1] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. In *Computational Linguistics*, pages 19(2):263–311, 1993.

[2] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Technical report TR-10-98, Center for Research in Computing Technology (Harvard University)*, 1998.

[3] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Overview of the Patent Translation Task at the NTCIR-7Workshop. In *Proc. of NTCIR-7*, 2008.

[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. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, 2007.

[5] P. Koehn, F. J. Och, and D. Marcu. Statistical Phrase-Based Translation. *Proc. of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 127–133, 2003.

Table 2. Evaluation results by automatic metrics

		GROUP-ID	TASK	RUN	Evaluation Metrics			
					BLEU	NIST	WER	PER
Method	Proposed method	NICT-ATR	JE	1	0.3179	**6.9745**	**0.7963**	**0.4458**
	Baseline	NICT-ATR	JE	2	0.3212	6.8908	0.8054	0.4513

[6] M. Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In *In the International NATO Symposium on Artificial and Human Intelligence*, 1981.

[7] NAIST. ChaSen, 2008. `http://chasen-legacy.sourceforge.jp/`.

[8] NIST. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurence Statistics, 2002. `http://www.nist.gov/speech/tests/mt/mt2001/resource/`.

[9] F. J. Och. Minimum Error Rate Training for Statistical Machine Translation. *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, 2003.

[10] F. J. Och and H. Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

[12] A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of International Conference on Spoken Language Processing*, 2002.

[13] M. Utiyama and H. Isahara. A Japanese-English Patent Parallel Corpus. In *MT Summit XI*, 2007.

[14] Y. Zhang and S. Vogel. Measuring confidence intervals for the machine translation evaluation metrics. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, 2004.