System Description of NiCT SMT for NTCIR-8

Keiji Yasuda
National Institute of Communications Technology
3–5, Hikaridai, Keihanna Science City, 619–0289, Japan
keiji.yasuda@nict.go.jp

Taro Watanabe
National Institute of Communications Technology
3–5, Hikaridai, Keihanna Science City, 619–0289, Japan
taro.watanabe@nict.go.jp

Masao Utiyama
National Institute of Communications Technology
3–5, Hikaridai, Keihanna Science City, 619–0289, Japan
mutiyama@nict.go.jp

Eiichiro Sumita
National Institute of Communications Technology
3–5, Hikaridai, Keihanna Science City, 619–0289, Japan
eiichiro.sumita@nict.go.jp

ABSTRACT
In this paper we describe the patent translation system which was submitted for the NTCIR-8 Patent Translation Task. Our phrase-based Statistical Machine Translation (SMT) system is trained on a bilingual corpus (3 million sentence pairs) and large size monolingual corpora (460 million sentences for Japanese and 350 million sentences for English). In addition to the normal SMT, we use SVM-based reranker. According to the experimental results, our baseline system gives the high BLEU score. However, the reranker gives negative effects.

Categories and Subject Descriptors
I.2.7 [Natural Language Processing]: [Machine translation]

General Terms
Experimentation

Keywords
SMT, Reranker.

1. INTRODUCTION
Current machine translation (MT) research shows the effectiveness of a corpus-based machine translation framework [10]. An MT system using a frame work such as Statistical Machine Translation (SMT) [1] is considered a useful convenient technology because of the rapid and mostly automated MT system building.

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

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

Considering all of the previously mentioned points, Japanese-English patent translation is one of the most interesting SMT research fields which satisfies these points. A large Japanese-English patent parallel corpus has just been released by the NTCIR-8 workshop [6]. Commercial demand is also very high in both directions for Japanese-English patent translation. In this paper, we describe the system overviews of our MT system which is based on Phrase-based SMT technology. Section 2 explains the system framework. Section 3 and 4 detail the experimental setting and results. Section 6 concludes the paper.

2. SYSTEM FRAMEWORK
2.1 Phrase-based SMT
We employed a log-linear model as a phrase-based statistical machine translation framework [9]. 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

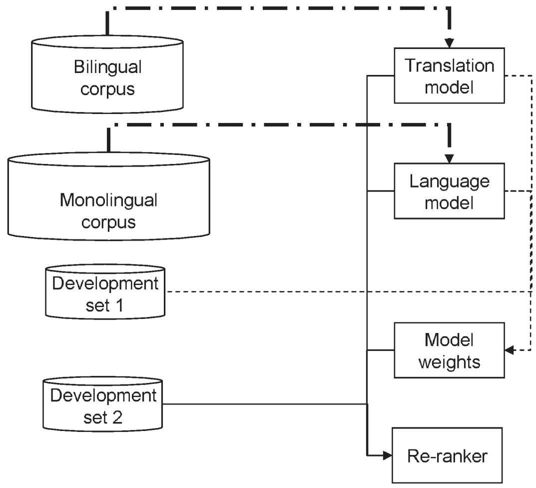

Figure 1: Framework of the submitted system (Run 1).

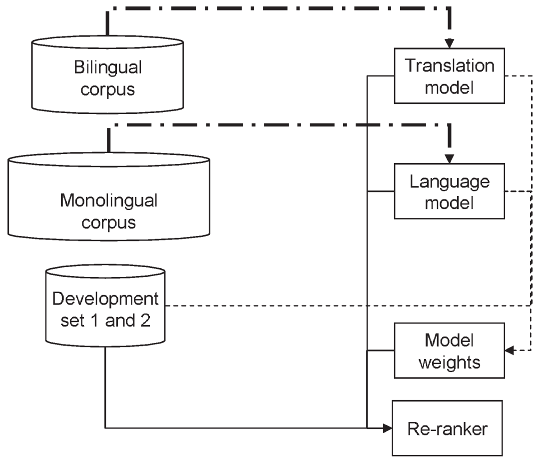

Figure 2: Framework of the submmited system (Run 2).

weight, and M is the number of features. λ_i is tuned by using the Minimum Error Rate Training (MERT) algorithm [12] on a development set.

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

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

For our system, we use the following 8 features (h_i):

- Phrase translation probability from source language to target language
- Phrase translation probability from target language to source language
- Lexical weighting probability from source language to target language
- Lexical weighting probability from target language to source language
- Phrase penalty
- Word penalty
- Distortion model
- Target language model probability

2.2 Reranker

Our ranking algorithm is based on a ranking approach of Collins and Duffy [4], but differs in that we employed an online large-margin learning for structured output based on the margin infused relaxed algorithm (MIRA) [5]. Fig3 shows the outline of the procedure. We generate large N-best list e for m input sentences $\mathbf{f}_1 \cdots m$. For each iteration, we randomly choose an input sentence $\mathbf{f}_i$ and its corresponding n_i-best list $\mathbf{e}_i$. At line 5, we seek a maximum likely hypothesized translation $\mathbf{e}_{ij}$ using the current weight $\mathbf{w}$

$$\mathbf{h}(\mathbf{e}_{ij}) \cdot \mathbf{w} - \mathbf{b}(\mathbf{e}_{ij}) \quad (3)$$

where $\mathbf{h}(\mathbf{e}_{ij})$ and $\mathbf{b}(\mathbf{e}_{ij})$ are a feature vector representation and BLEU score for $\mathbf{e}_{ij}$, respectively. Then, we update $\mathbf{w}$ by the value of $\mathbf{w}'$ which minimizes

$$\frac{\lambda}{2} \parallel \mathbf{w}' - \mathbf{w} \parallel^2 + l_{ij} - \Delta\mathbf{h}(\mathbf{e}_{ij}) \cdot \mathbf{w}' \quad (4)$$

where l_{ij} is a loss incurred by selecting the $\mathbf{e}_{ij}$ as the best translation computed by the difference of BLEU from an oracle translation e_{i*}

$$l_{ij} = \mathbf{b}(\mathbf{e}_{i*}) - \mathbf{b}(\mathbf{e}_{ij}) \quad (5)$$

and $\Delta\mathbf{h}(\mathbf{e}_{ij}) = \mathbf{h}(\mathbf{e}_{i*}) - \mathbf{h}(\mathbf{e}_{ij})$. $\lambda(> 0)$ is a constant to influence the fitness to the training data. Equation 4 is solved by:

$$\mid \mathbf{w}' \mid = \mathbf{w}' + \min\left(\frac{l_{ij} - \Delta\mathbf{h}_{ij} \cdot \mathbf{w}}{\parallel \Delta\mathbf{h}_{ij} \parallel^2}, \frac{1}{\lambda}\right) \cdot \Delta\mathbf{h}_{ij} \quad (6)$$

Unlike the ranking SVM approach for training[7], our learning algorithm considers only a single pair of correct and incorrect translations in each iteration using the loss biased maximization in Equation 3 largely inspired by Chiang et al. [3]. For the loss function l_{ij} and the underlying BLEU score $\mathbf{b}(\cdot)$, we applied document scaled BLEU which computes BLEU by replacing one translation $\mathbf{e}_{i1}$ to another $\mathbf{e}_{ij}$ in a set of 1-best translations $\{\mathbf{e}_{i1}\}_{i=1\ldots m}$ [16]. Oracle translations are selected with respect to $\mathbf{b}(\cdot)$. When multiple oracle translations are found, we select the one which maximizes $\Delta\mathbf{h}(\mathbf{e}_{ij}) \cdot \mathbf{w}$ [3].

3. EXPERIMENTAL SETTING

We used the training set from NTCIR-8 workshop Patent Translation Task [6] for the experiments. As monolingual corpora, we only used the specification part of the patent document for language models training. A development set and a test set were also provided by the workshop. We divided the development set into two parts and renamed them development set 1 and 2. Details for these data are shown in Table 1.

Table 1: Data set used for the experiments

Corpus	# of sentences	Usage
Monolingual corpus (ja)	462.75 M	LM training
Monolingual corpus (en)	350.39 M	LM training
Bilingual corpus	3.19 M	TM training
Development set 1	1000	MERT, reranker training
Development set 2	1000	MERT, reranker training
Test set (je)	1251	Evaluation
Test set (ej)	1119	Evaluation

Table 2: Official evaluation results by the organizer

TASK	DIRECTION	System	Data for MERT	Data for Reranking	BLEU (Official)
Intrinsic	JE	Run 1	Dev. set 1	Dev. set 2	30.32
Intrinsic	JE	Run 2	Dev. sets 1 and 2	Dev. sets 1 and 2	30.14
Intrinsic	EJ	Run 1	Dev. set 1	Dev. Set 2	35.37
Intrinsic	EJ	Run 2	Dev. sets 1 and 2	Dev. sets 1 and 2	35.87

Algorithm 1 Online Learning Algorithm

procedure ONLINELEARNING($\mathbf{e}$)
  $\mathbf{w} = 0$ ▷ initialize
  for $t = 1, ..., T$ **do**
    for $i = \text{RANDOM}(1, ..., m)$ **do**
      $j = \text{argmax}_{j'=1...n_i} \; \mathbf{h}(\mathbf{e}_{ij'}) \cdot \mathbf{w} - \mathbf{b}(\mathbf{e}_{ij'})$
      $\mathbf{w} \leftarrow \text{argmin}_{\mathbf{w}'} \; \frac{\lambda}{2}||\mathbf{w}' - \mathbf{w}||^2 + l_{ij} - \Delta\mathbf{h}(\mathbf{e}_{ij}) \cdot \mathbf{w}'$
    end for
  end for
  return $\mathbf{w}$
end procedure

Figure 3: Outline of the learning algorithm

For the statistical machine-translation experiments, we segmented Japanese words using the Japanese morphological analyzer ChaSen [11]. Then, we used the preprocessed data to train the phrase-based translation model by using GIZA++ [13] and the Moses tool kit [8]. Target-side language models are trained by the NiCT in-house toolkit. The language model configuration is a modified Kneser-Ney [2] and 7-gram.

For the reranking experiments, the 200 best translations are used for both reranker training and actual test set reranking ($N = 200$) . As features for the reranking, we used decoder scores, source side input and target side output.

We primarily submitted 2 systems (Run 1 and Run 2) for the task. Fig. 1 and 2 shows the respective framework for each systems. The differences between these systems are data usage for MERT and the reranker. As shown in Fig. 1, Run 1 uses development set 1 for MERT and development set 2 for the reranker training. Meanwhile, Run 2 uses the whole development set for both MERT and reranker training.

4. EXPERIMENTAL RESULTS

Table 2 shows the experimental results. The BLEU[14] scores shown in the table are computed by the workshop organizer. Comparing the BLEU score of Run 1 and Run 2, Run 1 gives a better BLEU score than Run 2 for Japanese to English translation direction. However, opposite results are obtained in the other translation direction. Considering these results, it difficult to conclude better data usage between Run 1 and Run2.

5. DISCUSSION

To evaluate the effects of the reranker, we calculate the BLEU score of a baseline system using the provided reference translation of the test set. The baseline system uses exactly the same models as Run 2 except the reranker.

Table 3 shows the evaluation results. In the table, the BLEU scores for Run 1 and 2 are also shown. The difference in BLUE scores between table 2 and 3 may have been caused by a preprocessing difference between us and the organizer.

According to the table, the baseline system gives a better BLEU score than Runs 1 and 2. Degradation in English to Japanese translation direction is especially large. Since the previous research about reranking[15] gives positive results, we will consider further work to improve the reranker by tuning SVM parameters, or adding more features.

6. CONCLUSION

We describe system submitted for the NTCIR-8 patent translation track. Our system was trained on a bilingual corpus (about 3 million sentences pairs) and monolingual corpora (460 million sentences for Japanese and 350 million sentences for English).

In addition to the normal phrase-based SMT, we built the reranker which chooses 1 best out of 200 best outputs. According to the post evaluation results, however, the reranker gives a negative effect.

7. REFERENCES

Table 3: Results of our evaluation experiments

TASK	DIRECTION	System	Data for MERT	Data for Reranking	BLEU (unoffical)
Intrinsic	JE	Run 1	Dev. set 1	Dev. set 2	30.28
Intrinsic	JE	Run 2	Dev. sets 1 and 2	Dev. sets 1 and 2	30.15
Intrinsic	JE	Baseline	Dev. sets 1 and 2	N/A	30.58
Intrinsic	EJ	Run 1	Dev. set 1	Dev. set 2	35.11
Intrinsic	EJ	Run 2	Dev. sets 1 and 2	Dev. sets 1 and 2	35.40
Intrinsic	EJ	Baseline	Dev. sets 1 and 2	N/A	36.62

[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] D. Chiang, Y. Marton, and P. Resnik. Online Large-Margin Training of Syntactic and Structural Translation Features. In In Proceedings of EMNLP, pages 224–233, 2008.

[4] M. Collins and D. Nigel. New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron. In In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), pages 263–270, 2002.

[5] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. Journal of Machine Learning Research, pages 551–585, 2006.

[6] A. Fujii, M. Utiyama, M. Yamamoto, T. Utsuro, T. Ehara, H. Echizen-ya, and S. Shimohata. Overview of the Patent Translation Task at the NTCIR-8 Workshop. In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2010, 2010.

[7] T. Joachims. Optimizing search engines using clickthrough data. In In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 133–142, 2002.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

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

[14] 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.

[15] K. Sudoh, T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki. NTT statistical machine translation system for IWSLT 2008. In In Proceedings of IWSLT, pages 92–97, 2008.

[16] T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki. Online Large-Margin Training for Statistical Machine Translation. In In Proceedings of EMNLP-CoNLL, pages 764–773, 2007.