UQAM's System Description for the NTCIR-10 Japanese and English PatentMT Evaluation Tasks

Fatiha Sadat
University du Quebec à Montréal
201 President Kennedy
Montréal, QC,
Canada, H2X 3Y7
sadat.fatiha@uqam.ca

Fu Zhe
University du Quebec à Montréal
201 President Kennedy
Montréal, QC,
Canada, H2X 3Y7
fu.zhe@courrier.uqam.ca

ABSTRACT

This paper describes the development of a Japanese-English and English-Japanese translation system for the NTCIR-10 Patent MT tasks. The MT system is based on the provided training data and Moses decoder. We report our first attempt on statistical machine translation for these pairs of languages and the Patent domain.

Categories and Subject Descriptors

D.3.3 [**Artificial Intelligence**]: Natural Language Processing – *Machine Translation*

General Terms

Natural Language Processing.

Keywords

Statistical machine translation, segmentation, patent translation.

Team Name

UQAM

Subtasks/Languages

Japanese-to-English and English-to-Japanese Patent MT.

External Resources Used

MeCab, Giza++, SRILM, Moses.

1. INTRODUCTION

This paper describes the Statistical Machine Translation (SMT) system developed by the team of Natural Language Processing at the computer science department of the University of Quebec in Montreal (UQAM), QC, Canada. Our participation is considered as a first attempt to develop a SMT system for the Japanese-English and English-Japanese pairs of languages. Thus, being a first participation at NTCIR-10 PatentMT task, we relied on existing NLP tools to develop the SMT system for Japanese-English and English-Japanese language pairs. We used the classical tools of the state-of-the-art in SMT, such as Moses decoder [2] [1], GIZA++ [4], SRILM [1], etc.

Japanese language is written without delimiters between words; the analogous situation in English would be if words were written without spaces between those words. Also, Japanese is written in three different scripts (*kanji*, *hiragana*, *katakana*), which makes word segmentation a less difficult problem compared to Chinese language.

Given the lack of word delimiters in written Japanese, word segmentation is generally considered a crucial first step in processing Japanese texts. For instance, the sequence **ここでは、第２のコンタクトホール６１内に、** (i.e., here, in the contact hall 61 of the second,) will have a proper segmentation as follows: **|ここ|で|は|、|第２|の|コンタクト|ホール|６１|内|に|.**

We used MeCab [5], a part-of-speech tagger and a morphological analyzer for the segmentation of Japanese sentences and words.

Although, we consider the use of linguistic information as crucial in the translation process between Japanese and English; we did not have enough time to implement a rule-based system and combine it to the statistical machine translation.

This paper is organized as follows. In the next section, we describe the approach used to develop the SMT system. Section 3 will give an overview of the experiments and results. Section 4 concludes the present paper and discusses the possible future directions.

[1] Available on http://www.statmt.org/moses/

2. SYSTEM DESCRIPTION

Our approach on statistical machine translation for Japanese and English pairs of languages are described as follows. First, a pre-processing step is performed on the source language, in order to convert raw texts into a format suitable for both training and decoding models. For the supplied English-Japanese and Japanese-English parallel corpora, we relied on a simple tokenisation of English phrases and a segmentation of Japanese texts using Mecab tool [5].

A the common practice of extracting bilingual phrases from the parallel data usually consists of three steps: first, words in bilingual sentence pairs are aligned using state-of-the-art automatic word alignment tools, such as GIZA++ [4], in both directions; second, word alignment links are refined using heuristics, such as Grow-Diagonal-Final (GDF) method; third, bilingual phrases are extracted from the parallel data based on the refined word alignments with predefined constraints [4]. The trigram language models are implemented using the SRILM toolkit [1].

Decoding is the central phase in SMT, involving a search for the hypotheses *t* that have highest probabilities of being translations of the current source sentence *s* according to a model for *P(t|s)*. Moses [2], an open source toolkit for phrase-based SMT system, was used as a decoder.

These steps of building a translation system are considered as a common practice in the state-of-the-art of phrase-based SMT systems.

Once this is accomplished, a variant of Powell's algorithm is used to find weights that optimize BLEU score [3] over these hypotheses, compared to reference translations.

3. EXPERIMENTAL RESULTS

We used the described tools to develop a basic SMT system for the translation of Patents from Japanese to English and English to Japanese.

Table 1 shows the formal run evaluation results at the NTCIR-10 PatentMT tasks for Japanese-to-English and English-to-Japanese languages, in terms of BLEU and NIST scores.

Some examples on a sentence of the test file are shown in Table 2. We can see and compare with the baseline in terms of adequacy and fluency. These examples show that our translation is not very fluent but comprehensible and even we can consider it as very close to the reference or baseline. Also, our results in terms of BLEU score (or NIST) are very close to the rest of participants who used the statistical model for Japanese to English translation. However, we have a wide room for improvement on the English to Japanese translation as our basic system performed very poorly. We believe that we need to improve the segmentation and post-processing for the English-to-Japanese translation as the target language (here Japanese) is not a straightforward language but needs more work. Also, using a hybrid rule-based and statistical model will help improve the performance of our SMT system.

Table 1. Results on the Japanese-to-English and English-to-Japanese PatentMT Tasks

	J->E (BLEU/NIST)	**E->J (BLEU/NIST)**
UQAM (our results)	21.8/7.072	14.97/5.66
BASELINE(1)	28.56/7.97	32.98/8.08
BASELINE(2)	28.86/7.99	33.61/8.18

Table 2. Examples of translations from Japanese to English and English to Japanese with the references

Japanese to English:

図３のフローチャートでステップＳ６に到達すると、図５の「行程判別の可否判定」が起動される。

Our Translation:

In the flowchart of fig. 3 and arrives at the step S6 in fig. 5, and the stroke of the determination execution-prohibition decision" is started.

Baseline1:

Step S6 in the flowchart of FIG. 3, when the judgement of the determination "process" shown in FIG. 5 is started.

Baseline2:

In the flowchart of FIG. 3 reaches the step S6 of FIG. 5, the judgement "determined" stroke is started.

English to Japanese:

It is to be noted that an interface using a slip ring or optical communication is inserted between the X-ray detector 12 and the data collecting section 16.

Our Translation:

なお、インタフェーススリップリングを用いる又は光学通信が間に挿入Ｘ線検出器１２及びデータ収集部１６。

Baseline:

なお、Ｘ線検出器１２とデータ収集部１６との間には、スリップリングや光通信などを用いたインタフェースが介挿される。

4. CONCLUSION

In this paper, we have reported the results of our participation at NTCIR-10 PatentMT task for Japanese-English and English-Japanese pairs of languages, using the provided training data only.

Using a basic statistical translation system, our results on the Japanese-to-English showed a comparative results with the rest of participants and could generate adequate and quite fluent translated sentences. However, the translation from English to Japanese did not perform well on a basic MT approach and did not achieve satisfactory results compared to the baseline and results of the rest of participants. The rate of OOVs is high and the segmentation on the target language (Japanese) was not done well. In the future, we would like to investigate the translation from English-to-Japanese and combine the statistical machine translation to a rule-based one. We believe this combination will improve our translation system, boost its performance and help produce more fluent translated sentences. We also need to work more on the post-processing step, especially for the English-to-Japanese translation.

5. REFERENCES

[1] A. Stolcke. Srilm-An Extensible Language Modeling Toolkit. In *Proc. Of the International Conference on Spoken Language Processing* (2002).

[2] P. Koehn, W. Shen, M. Federico, N. Bertoldi, C. Callison-Burch, B. Cowan, C. Dyer, H. Hoang, O. Bojar, R. Zens, A. Constantin, E. Herbst, C. Moran, and A. Birch, “Moses: Open source toolkit for statistical machine translation,” in Proceedings of the ACL 2007 Interactive Presentation Sessions, Prague, (2007).

[3] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176(W0109-022), IBM Research Division, Yorktown Heights, NY (2001).

[4] Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. Computational linguistics 29 (1), 19-51 (2003).

[5] Y. M. Taku Kudo. Japanese dependency analysis using cascaded chunking. In CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops), pages 63–69, 2002.