Producing a Test Collection for Patent Machine Translation in the Seventh NTCIR Workshop

Atsushi Fujii † **Masao Utiyama** ‡ **Mikio Yamamoto** † **Takehito Utsuro** †

† University of Tsukuba
‡ National Institute of Information and Communications Technology
fujii@slis.tsukuba.ac.jp

Abstract

In aiming at research and development on machine translation, we produced a test collection for Japanese-English machine translation in the seventh NTCIR Workshop. This paper describes details of our test collection. From patent documents published in Japan and the United States, we extracted patent families as a parallel corpus. A patent family is a set of patent documents for the same or related invention and these documents are usually filed to more than one country in different languages. In the parallel corpus, we aligned Japanese sentences with their counterpart English sentences. Our test collection, which includes approximately 2 000 000 sentence pairs, can be used to train and test machine translation systems. Our test collection also includes search topics for cross-lingual patent retrieval and the contribution of machine translation to a patent retrieval task can also be evaluated. Our test collection will be available to the public for research purposes after the NTCIR final meeting.

1. Introduction

Since the Third NTCIR Workshop in 2001[1], which is an evaluation forum for research and development on information retrieval and natural language processing, the Patent Retrieval Task had continuously been performed (Fujii et al., 2004; Fujii et al., 2006; Fujii et al., 2007b; Iwayama et al., 2006). In the Sixth NTCIR Workshop (Fujii et al., 2007b), 10 years of patent documents published by the Japanese Patent Office (JPO) and the U.S. Patent & Trademark Office (USPTO) were used as target document collections independently.

After exploring patent retrieval issues for a long time, the authors of this paper determined to address another issue in patent processing. Among a number of research issues related to patent processing (Fujii et al., 2007a), we have selected Machine Translation (MT) for patent documents, which is useful for a number of applications and services, such as Cross-Lingual Patent Retrieval (CLPR) and filing patent applications to foreign countries.

Reflecting the rapid growth in the use of multilingual corpora, a number of data-driven MT methods have recently been explored and most of these methods are termed "Statistical Machine Translation (SMT)". While large bilingual corpora for European languages, Arabic, and Chinese are available for research and development purposes, these corpora are rarely associated with Japanese and therefore it is difficult for explore SMT for Japanese.

However, we found that the patent documents used for the NTCIR Workshops can potentially alleviate this data shortage problem. Highchi et al. (2001) used "patent families" as a parallel corpus for extracting new translations. A patent family is a set of patent documents for the same or related invention and these documents are usually filed to more than one country in different languages. Following this method, we can produce a bilingual corpus for Japanese and English. In addition, there are a number of SMT engines (decoders) available to the public, such as Pharaoh and Moses[2], which are applied to bilingual corpora of any language pairs.

Motivated by the above background, we have determined to organize a machine translation task for patents ("the Patent Translation Task") in the Seventh NTCIR Workshop (NTCIR-7). Because NTCIR-7 has started in October 2007 and the final meeting will be held in December 2008, this paper describes details of the test collection for the Patent Translation Task.

2. Overview of the Patent Translation Task

The Patent Translation Task consists of the following three steps. First, the organizers, who are the authors of this paper, provide groups participating in the Patent Translation Task with a training data set, which consists of aligned sentence pairs in Japanese and English. Each participating group can use this data set to train their MT system, which can use either data-driven SMT or conventional knowledge-intensive MT methods.

Second, the organizers provide the groups with a test data set, which consists of sentences in either of Japanese and English. Each group is requested to machine translate these sentences in one language into the other language and submit their translation results to the organizers. Each group is allowed to submit more than one translation result for each test sentence.

Third, the organizers evaluate the submission of each group. We use intrinsic and extrinsic evaluation methods. In the intrinsic evaluation, we use BLEU (Papineni et al., 2002), which was proposed as an automatic evaluation measure for SMT, and human judgments. We analyze the relation between the value of BLEU and the evaluation by human judgments. In the extrinsic evaluation, we investigate the contribution of MT to CLPR. In the Patent Retrieval Task at NTCIR-5, intended to CLPR, search topics in Japanese were translated into English by human experts. We reuse these search topics for the evaluation of machine translation.

[1] http://research.nii.ac.jp/ntcir/index-en.html

[2] http://www.statmt.org/wmt07/baseline.html

We repeat the above three steps in dry run and formal run. If some problem is found in the dry run, we modify the task procedure in the formal run.
Sections 3. and 4. explain in the intrinsic and extrinsic evaluation methods, respectively.

3. Intrinsic Evaluation

Figure 1 depicts the process flow of the intrinsic evaluation. We explain the entire process in terms of Figure 1.
In the Patent Retrieval Task at NTCIR-6, the following two document sets were used.

- 10 years of unexamined Japanese patent applications published by the JPO in 1993–2002. The number of documents is approximately 3 500 000.
- 10 years of patent grant data published by the USPTO in 1993–2002. The number of documents is approximately 1 300 000. Because the USPTO documents consist of only patent that have been granted, the number of these documents is smaller than that of the JPO documents.

From these document sets, we automatically extracted patent families. Among a number of ways to apply for patents in more than one country, we focused only on patents claiming priority under the Paris Convention. In a patent family applied by the Paris Convention, the member documents in a patent family are assigned with the identical priority number. Thus, we can identify patent families systematically.
Figure 2 shows an example patent family, in which the upper and lower parts are fragments (bibliographic information and abstracts) of an unexamined Japanese patent application and a USPTO patent, respectively. In Figure 2, "(31)" in the Japanese document and "[21]" in the English document denote priority numbers, both of which are "295127" in this example, respectively.
Using priority numbers, we extracted approximately 85 000 USPTO patents that originated from JPO patents. While patents are structured with a number of fields, in "Background of the Invention" and "Detailed Description of the Preferred Embodiments" fields the text is usually translated on a sentence-by-sentence basis. Thus, for these fields we used a method (Utiyama and Isahara, 2003) to align sentences in Japanese with their counterpart sentences in English.
In the real world, a reasonable scenario is that an MT system is trained using existing patent documents and is used to translate new patent documents. Thus, we produced training and test data sets based on the publication year. While we used patent documents published in 1993–2000 to produce a training data set, we used patent documents published in 2001–2002 to produce a test data set.
The training data set consists of approximately 1 800 000 Japanese-English sentence pairs, which is one of the largest collections for Japanese and English MT. We randomly selected 3000 sentence pairs from the training data, and asked human experts to judge whether each sentence pair is a translation or not. Approximately 90% of the 3000 pairs were correct translations. This training data set is used for both the dry run and the formal run.
The number of sentence pairs extracted from patent documents published in 2000-2001 was approximately 630 000. For a test data set, we select approximately 1400 sentence pairs that were judged as correct translations by human experts. In the selected pairs, the Japanese (or English) sentences are used to evaluate Japanese-English (or English-Japanese) MT. Unlike the training data set, we use different test sets for the dry run and the formal run.
To evaluate translation results submitted by participating groups, we use BLEU and human judgments independently. To calculate the value of BLEU for the test sentences, we need one or more reference translations. For each test sentence, we use their counterpart sentence as a reference translation. In addition, for randomly sampled 600 sentences, we ask more than one human expert to produce a reference translation for each test sentence independently, to enhance the objectivity of the evaluation. For human judgments, we ask human experts to evaluate each translation result based on the fluency and understandability. We analyze the relation between the evaluation by BLEU and the evaluation by human judgments.
To verify whether our task is feasible or not, we performed preliminary experiments before NTCIR-7 started (Utiyama and Isahara, 2007). We performed only the intrinsic evaluation using BLEU. For each test sentence, we used their counterpart sentence as the reference translation. We used Pharaoh as a decoder and evaluated results for different combinations of parameters. The values of BLEU ranged from 23 to 27, which are comparable with those reported for Chinese-English SMT.

4. Extrinsic Evaluation

In the extrinsic evaluation, we investigate the contribution of MT to CLPR. In brief, each group is requested to machine translate search topics in English into Japanese. Each of the translated search topics is used to search a patent document collection in Japanese for the relevant documents. The evaluation result for CLPR is compared with that for a monolingual retrieval for Japanese. Figure 3 depicts the process flow of the extrinsic evaluation. We explain the entire process in terms of Figure 3.
Processes of patent retrieval differ significantly, depending on the purpose of retrieval. One process is the "technology survey", in which patents related to a specific technology, such as "blue light-emitting diode", are searched. This process is similar to ad hoc retrieval tasks targeting nonpatent documents.
Another process is the "invalidity search", in which prior arts related to a patent application are searched. Away from academic research, invalidity searches are performed by examiners in government patent offices and searchers in the intellectual property divisions of private companies.
In the Patent Retrieval Task at NTCIR-5, we performed the invalidity search. The purpose was to search a Japanese patent collection, which is the same collection described in Section 3., for the patents that can invalidate the demand in an existing claim. Thus, a search topic is a claim in a patent

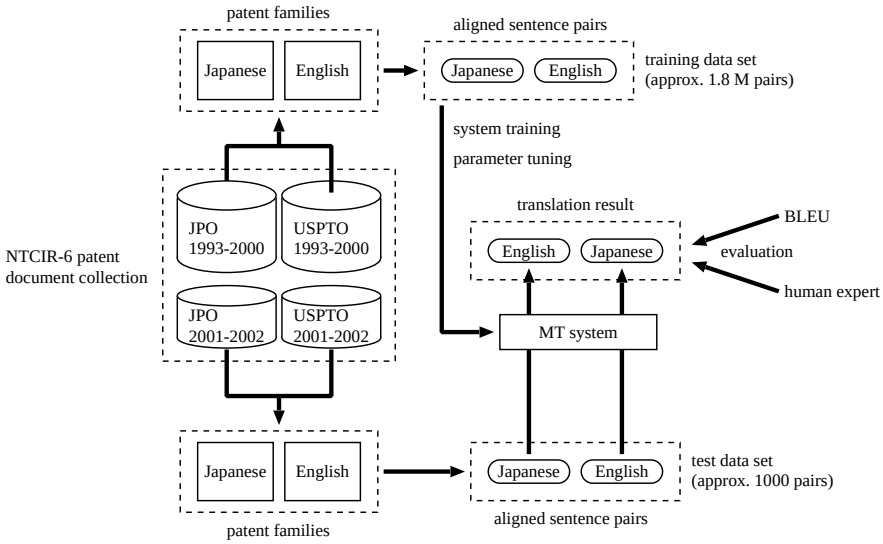

Figure 1: Overview of the intrinsic evaluation.

(11)【公開番号】特開平８－１１４２７８
(43)【公開日】平成８年（１９９６）５月７日
(54)【発明の名称】マイクロアクチュエータ
(21)【出願番号】特願平７－２３９２３０
(22)【出願日】平成７年（１９９５）８月２４日
(31)【優先権主張番号】２９５，１２７
(32)【優先日】１９９４年８月２４日
(33)【優先権主張国】米国（ＵＳ）
(57)【要約】
【課題】断熱構造を備えるマイクロアクチュエータ。
【解決手段】フローチャネルを介して運搬される流体流を制御する超小型バルブの形態をなすマイクロアクチュエータであり、サーマルアクチュエータによって選択的に駆動される熱駆動部材を有し、これが駆動されることによって熱エネルギを生成する第１基板と、対向する第１、第２主要面を有する第２基板よりなる。第２基板が第１主要面で第１基板に取付けられる。第２の主要面は第２基板が支持体に取付られると絶縁セルを画定し、これによってマイクロアクチュエータの熱容量を減少させ、第１基板を支持体から熱遮断する。

[11] Patent Number 5,529,279
[45] Date of Patent June 25, 1996
[54] Thermal isolation structures for microactuators
[57] Abstract
A microactuator preferably in the form of a microminiature valve for controlling the flow of a fluid carried by a flow channel includes a first substrate having a thermally-actuated member selectively operated by a thermal actuator such that the first substrate thereby develops thermal energy, and a second substrate having opposed first and second major surfaces. The second substrate is attached to the first substrate at the first major surface. The second major surface defines an isolation cell for enclosing a volume when the second substrate is attached to the support to thereby reduce the thermal mass of the microactuator and to thermally isolate the first substrate from the support.
[21] Appl. No.: 295127
[22] Filed: August 24, 1994

Figure 2: Example of JP-US patent family.

application. We selected search topics from patent applications that had been rejected by the JPO. For each search topic, we used one or more citation (i.e., prior art) that were used for the rejection, as the relevant documents. We produced 1189 search topics. Additionally, in aiming to CLPR, these search topics were translated by human experts into English. However, because the use of English search topics was optional, no participating group in NTCIR-5 performed CLPR.

In the extrinsic evaluation at NTCIR-7, we reuse these search topics. Each search topic file includes a number of additional SGML-style tags. The claim used as the target of invalidation is specified by <CLAIM>. The date of filing is specified by <FDATE> and only the patents published before this date can potentially be relevant. Figure 4 shows an example topic claim translated into English.

We can use all the 1189 search topics for the dry run and the formal run. However, because the length of a single claim is usually much longer than that of a single sentence, the computation time for the translation can potentially be prohibitive. Thus, we select a set of 100 search topics for the dry run and the formal run independently.

While each group is requested to machine translate the 100 search topics, the retrieval is performed by the organizers. As a result, we can standardize the retrieval system and the contribution of each group can be compared only in terms of the translation accuracy. In addition, for most of the participating groups, which are research groups for natural language processing, the retrieval for 10 years of patent documents is not an easy task. We use a system that participated in the NTCIR-5 Patent Retrieval Task (Fujii and Ishikawa, 2005) for retrieval purposes.

As evaluation measures, we use Mean Average Precision (MAP) and recall at the top N documents. In the real world, an expert of patent retrieval usually investigates hundreds of documents. Thus, we set $N = 100, 200, 500, 1000$. We also analyze the relation between the results of the intrinsic evaluation and the extrinsic evaluation.

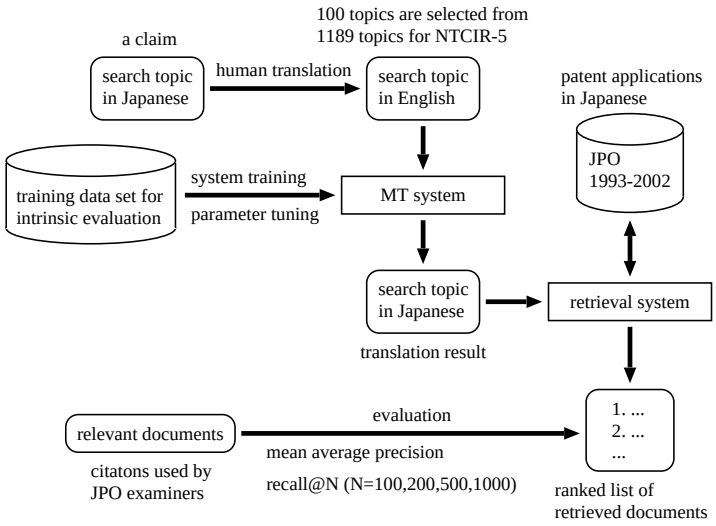

Figure 3: Overview of the extrinsic evaluation.

```
<TOPIC>
<NUM>1048</NUM>
<LANG>EN</LANG>
<FDATE>19950629</FDATE>
<CLAIM>A milk-derived calcium-containing composition comprising an inorganic salt mainly composed of calcium obtained by baking a milk-derived prepared matter containing milk casein-bonding calcium and/or colloidal calcium. </CLAIM>
</TOPIC>
```

Figure 4: Example search topic produced at NTCIR-5.

5. Conclusion

In aiming at research and development on machine translation, we produced a test collection for Japanese-English machine translation in the seventh NTCIR Workshop. This paper describes details of our test collection. Our test collection, which includes approximately 2 000 000 Japanese-English aligned sentence pairs, can be used to train and test machine translation systems. Our test collection also includes search topics for cross-lingual patent retrieval and the contribution of machine translation to a patent retrieval task can also be evaluated. Our test collection is useful from scientific and commercial points of view and will be available to the public for research purposes after the final meeting of NTCIR-7.

6. References

Atsushi Fujii and Tetsuya Ishikawa. 2005. Document structure analysis for the NTCIR-5 patent retrieval task. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 292–296.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2004. The patent retrieval task in the fourth NTCIR workshop. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 560–561.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2006. Test collections for patent retrieval and patent classification in the fifth NTCIR workshop. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 671–674.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2007a. Introduction to the special issue on patent processing. *Information Processing & Management*, 43(5):1149–1153.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2007b. Overview of the patent retrieval task at the NTCIR-6 workshop. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 359–365.

Shigeto Higuchi, Masatoshi Fukui, Atsushi Fujii, and Tetsuya Ishikawa. 2001. PRIME: A system for multilingual patent retrieval. In *Proceedings of MT Summit VIII*, pages 163–167.

Makoto Iwayama, Atsushi Fujii, Noriko Kando, and Yuzo Marukawa. 2006. Evaluating patent retrieval in the third NTCIR workshop. *Information Processing & Management*, 42(1):207–221.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79.

Masao Utiyama and Hitoshi Isahara. 2007. A Japanese-English patent parallel corpus. In *Proceedings of MT Summit XI*, pages 475–482.