Evaluating Effects of Machine Translation Accuracy on Cross-Lingual Patent Retrieval

Atsushi Fujii
University of Tsukuba
1-2 Kasuga Tsukuba
305-8550 Japan

Masao Utiyama
NICT
2-2-2 Hikaridai
619-0288 Japan

Mikio Yamamoto Takehito Utsuro
University of Tsukuba
Tennodai Tsukuba
305-8577 Japan

ABSTRACT

We organized a machine translation (MT) task at the Seventh NTCIR Workshop. Participating groups were requested to machine translate sentences in patent documents and also search topics for retrieving patent documents across languages. We analyzed the relationship between the accuracy of MT and its effects on the retrieval accuracy.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Measurement, Performance, Experimentation

Keywords

Cross-lingual information retrieval, Machine translation, Patent information, Evaluation measures, NTCIR

1. INTRODUCTION

To aid research and development in cross-lingual information access, we produced a test collection for machine translation (MT) targeting Japanese and English, and organized the Patent Translation Task at the Seventh NTCIR Workshop (NTCIR-7). To evaluate submissions from participating groups, which are MT results for a test data set, we used intrinsic and extrinsic evaluation methods. In the intrinsic evaluation, we used both the Bilingual Evaluation Understudy (BLEU) [2], which had been proposed as an automatic evaluation measure for MT, and human rating. In the extrinsic evaluation, we investigated the contribution of the MT to Cross-Lingual Patent Retrieval (CLPR). This paper focuses mainly on the extrinsic evaluation and explores the relationship between the accuracy of MT, evaluated by BLEU and human rating, and its effects on CLPR.

2. TRAINING DATA FOR MT

Research groups participated in NTCIR-7 were allowed to use any types of MT. However, compared with a knowledge-intensive rule-based MT, statistical MT (SMT), which has recently been explored, can easily be implemented if a large parallel corpus is available for training purposes. To obtain a parallel corpus, we extracted patent documents for the same or related inventions published in Japan and the United States during 1993–2002. We extracted approximately 85 000 USPTO patents that originated from Japanese patent applications. While patents are structured in terms of several fields, in the "Background of the Invention" and the "Detailed Description of the Preferred Embodiments" fields, text is often translated on a sentence-by-sentence basis. For these fields, we used a method to automatically align sentences in Japanese with their counterpart sentences in English. While we used patent documents published during 1993–2000 to produce a training data set, we used patent documents published during 2001–2002 to produce a test data set for the intrinsic evaluation. The training data set has approximately 1 800 000 J/E sentence pairs, which can be used for both the intrinsic and extrinsic evaluations.

3. TEST COLLECTION FOR CLPR

In the Patent Retrieval Task at NTCIR-5 [1], the purpose was to search Japanese patent applications published during 1993–2002 for the applications that can invalidate the demand in an existing claim. Each search topic is a claim in a Japanese patent application. Search topics were selected from patent applications that had been rejected by the Japanese Patent Office. For each search topic, one or more citations (i.e., prior arts) that were used for the rejection were used as relevant or partially relevant documents. With the aim of CLPR, these search topics were translated by human experts into English during NTCIR-5. We reused these search topics in NTCIR-7.

Although each group was requested to machine translate the search topics, the retrieval was performed by the organizers. As a result, we were able to standardize the retrieval system and the contribution of each group was evaluated in terms of the translation accuracy alone. The standard system performs word indexing, uses Okapi BM25 as the retrieval model, and retrieves up to the top 1000 documents for each topic. This system also uses the International Patent Classification to restrict the retrieved documents.

As evaluation measures for CLPR, we used the Mean Average Precision (MAP) and Recall for the top N documents (Recall@N). In the real world, an expert in patent retrieval usually investigates hundreds of documents. Therefore, we set $N = 100$, 200, 500, and 1000. We also used BLEU as an evaluation measure, for which we used the source search topics in Japanese as the reference translations.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

In principle, for the extrinsic evaluation we were able to use all of the 1189 search topics produced in NTCIR-5. However, because the length of a single claim is usually much longer than that of an ordinary sentence, the computation time for the translation can be prohibitive. Therefore, in practice we independently selected a subset of the search topics for the dry run and the formal run. If we use search topics for which the average precision of the monolingual retrieval is small, the average precision of CLPR methods can be so small that it is difficult to distinguish the contributions of participating groups to CLPR. Therefore, we sorted the 1189 search topics according to the Average Precision (AP) of monolingual retrieval using the standard retrieval system and selected 100 topics (AP $\geq$ 0.9) and 124 topics (0.9 $>$ AP $\geq$ 0.3) for the dry run and the formal run, respectively.

4. EVALUATION IN THE FORMAL RUN

The number of groups participated in the extrinsic evaluation was 12. All of these groups also participated in the E–J intrinsic evaluation, in which the purpose was to machine translate sentences in patent documents from English to Japanese. As a baseline system, the organizers submitted a result produced by Moses[1]. Table 1 shows the results for the E–J intrinsic evaluation and the extrinsic evaluation, which are denoted as "Intrinsic" and "Extrinsic", respectively. The rows in Table 1, each of which corresponds to the result of a single group, are sorted according to the values for BLEU in "Intrinsic". For human rating, experts evaluated each translation result based on fluency and adequacy, using a five-point rating. The value for "Human" is the average of adequacy and fluency. However, mainly because of time and budget constraints, human rating was performed only for five systems. To calculate MAP values, we used both relevant and partially relevant documents as the correct answers for the top 1000 documents. In Table 1, the row "Mono" shows the results for monolingual retrieval. The best MAP for CLPR obtained by HCRL is 0.3536, which is 74% of that for Mono.

We also used Recall@N as an evaluation measure for CLPR. We calculated the correlation coefficient ("R") between BLEU in the extrinsic evaluation and each CLPR evaluation measure. We found that the value of R for MAP was 0.936 whereas the values of R for Recall@N were below 0.9, irrespective of the value of N. In other words, we can potentially use BLEU to predict the contribution of MT to CLPR with respect to MAP, without performing retrieval experiments. This is a significant step toward the automatic evaluation of CLPR by means of the evaluation of MT. However, human rating did not correlate with MAP because as in Table 1 tsbmt outperformed the other groups with respect to human rating, but achieved the lowest MAP.

We used the two-sided paired t-test for statistical testing with respect to MAP. We also analyzed the extent to which the BLEU value should be improved to achieve a statistically significant improvement in MAP value. Figure 1 shows the relationship between the difference in BLEU value and the level of statistical significance of the MAP value. In Figure 1, each bullet point corresponds to a comparison of two groups. The bullet points are classified into three clusters according to the level of statistical significance for MAP. The y-axis denotes the difference between the two groups' BLEU values.

Table 1: Results of E–J int/ext evaluations.

		Intrinsic		Extrinsic	
Group	Method	BLEU	Human	BLEU	MAP
Moses	SMT	30.58	3.30	20.70	.3140
HCRL	SMT	29.97	—	21.10	.3536
NiCT-ATR	SMT	29.15	2.89	19.40	.3494
NTT	SMT	28.07	3.14	18.69	.3456
NAIST-NTT	SMT	27.19	—	20.46	.3248
KLE	SMT	26.93	—	19.07	.2925
tori	SMT	25.33	—	17.54	.3187
MIBEL	SMT	23.72	—	18.67	.2873
HIT2	SMT	22.84	—	17.71	.2777
Kyoto-U	EBMT	22.65	2.48	13.75	.2817
tsbmt	RBMT	17.46	3.60	12.39	.2264
FDU-MCandWI	SMT	10.52	—	11.10	.2562
TH	SMT	2.23	—	1.39	.1000
Mono	—	—	—	—	.4797

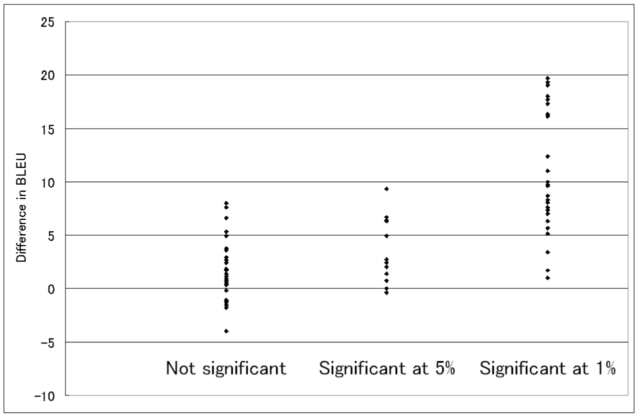

Figure 1: Relationship between difference in BLEU and statistical significance of MAP.

The y-coordinate of each bullet point was calculated from the values for "Extrinsic BLEU" in Table 1.

By comparing the three clusters in Figure 1, we deduce the difference in BLEU value should be more than 10 to safely achieve the 1% level of significance for MAP values. In the dry run, this threshold was 9 and thus the result was almost the same as the formal run. At the same time, because the values for BLEU and MAP can depend on the data set used, further investigation is needed to clarify the relationship between improvements in BLEU and MAP.

5. REFERENCES

[1] A. Fujii, M. Iwayama, and N. Kando. 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, 2006.

[2] 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*, pages 311–318, 2002.

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