The Effect of Translation Quality in MT-Based Cross-Language Information Retrieval

Jiang Zhu Haifeng Wang
Toshiba (China) Research and Development Center
5/F., Tower W2, Oriental Plaza, No.1, East Chang An Ave., Dong Cheng District
Beijing, 100738, China
{zhujiang, wanghaifeng}@rdc.toshiba.com.cn

Abstract

This paper explores the relationship between the translation quality and the retrieval effectiveness in Machine Translation (MT) based Cross-Language Information Retrieval (CLIR). To obtain MT systems of different translation quality, we degrade a rule-based MT system by decreasing the size of the rule base and the size of the dictionary. We use the degraded MT systems to translate queries and submit the translated queries of varying quality to the IR system. Retrieval effectiveness is found to correlate highly with the translation quality of the queries. We further analyze the factors that affect the retrieval effectiveness. Title queries are found to be preferred in MT-based CLIR. In addition, dictionary-based degradation is shown to have stronger impact than rule-based degradation in MT-based CLIR.

1 Introduction

Cross-Language Information Retrieval (CLIR) enables users to construct queries in one language and search the documents in another language. CLIR requires that either the queries or the documents be translated from a language into another, using available translation resources. Previous studies have concentrated on query translation because it is computationally less expensive than document translation, which requires a lot of processing time and storage costs (Hull & Grefenstette, 1996).

There are three kinds of methods to perform query translation, namely Machine Translation (MT) based methods, dictionary-based methods and corpus-based methods. Corresponding to these methods, three types of translation resources are required: MT systems, bilingual wordlists and parallel or comparable corpora. CLIR effectiveness depends on both the design of the retrieval system and the quality of the translation resources that are used.

In this paper, we explore the relationship between the translation quality of the MT system and the retrieval effectiveness. The MT system involved in this research is a rule-based English-to-Chinese MT (ECMT) system. We degrade the MT system in two ways. One is to degrade the rule base of the system by progressively removing rules from it. The other is to degrade the dictionary by gradually removing word entries from it. In both methods, we observe successive changes on translation quality of the MT system. We conduct query translation with the degraded MT systems and obtain translated queries of varying quality. Then we submit the translated queries to the IR system and evaluate the performance. Retrieval effectiveness is found to be strongly influenced by the translation quality of the queries. We further analyze the factors that affect the retrieval effectiveness. Title queries are found to be preferred in MT-based query translation. In addition, the size of the dictionary is shown to have stronger impact on retrieval effectiveness than the size of the rule base in MT-based query translation.

The remainder of this paper is organized as follows. In section 2, we briefly review related work. In section 3, we introduce two systems involved in this research: the rule-based ECMT system and the KIDS IR system. In section 4, we describe our experimental method. Section 5 and section 6 reports and discusses the experimental results. Finally we present our conclusion and future work in section 7.

2 Related Work

2.1 Effect of Translation Resources

Previous studies have explored the effect of translation resources such as bilingual wordlists or parallel corpora on CLIR performance.

Xu and Weischedel (2000) measured CLIR performance as a function of bilingual dictionary size. Their English-Chinese CLIR experiments on TREC 5&6 Chinese collections showed that the initial retrieval performance increased sharply with lexicon size but the performance was not improved after the lexicon exceeded 20,000 terms. Demner-Fushman and Oard (2003) identified eight types of terms that affected retrieval effectiveness in CLIR applications through their coverage by general-purpose bilingual term lists. They reported results from an evaluation of the coverage of 35 bilingual term lists in news retrieval application. Retrieval effectiveness was found to be strongly influenced by term list size for lists that contain between 3,000 and 30,000 unique terms per language.

Franz et al. (2001) investigated the CLIR performance as a function of training corpus size for three different training corpora and observed approximately logarithmically increased performance with corpus size for all the three corpora. Kraaij (2001) compared three types of translation resources for bilingual retrieval based on query translation: a bilingual machine-readable dictionary, a statistical dictionary based on a parallel web corpus and the Babelfish MT service. He drew a conclusion that the mean average precision of a run was proportional to the lexical coverage. McNamee and Mayfield (2002) examined the effectiveness of query expansion techniques by using parallel corpora and bilingual wordlists of varying quality. They confirmed that retrieval performance dropped off as the lexical coverage of translation resources decreased and the relationship was approximately linear.

Previous research mainly focused on studying the effectiveness of bilingual wordlists or parallel corpora from two aspects: size and lexical coverage. Kraaij (2001) examined the effectiveness of MT system, but also from the aspect of lexical coverage. Why lack research on analyzing effect of translation quality of MT system on CLIR performance? The possible reason might be the problem on how to control the translation quality of the MT system as what has been done to bilingual wordlists or parallel corpora. MT systems are usually used as black boxes in CLIR applications. It is not very clear how to degrade MT software because MT systems are usually optimized for grammatically correct sentences rather than word-by-word translation.

2.2 MT-Based Query Translation

MT-based query translation is perhaps the most straightforward approach to CLIR. Compared with dictionary or corpus based methods, the advantage of MT-based query translation lies in that technologies integrated in MT systems, such as syntactic and semantic analysis, could help to improve the translation accuracy (Jones et al., 1999). However, in a very long time, fewer experiments with MT-based methods have been reported than with dictionary-based methods or corpus-based methods. The main reasons include: (1) MT systems of high quality are not easy to obtain; (2) MT systems are not available for some language pairs; (3) queries are usually short or even terms, which limits the effectiveness of MT-based methods. However, recent research work on CLIR shows a trend to adopt MT-based query translation. At the fifth NTCIR workshop, almost all the groups participating in Bilingual CLIR and Multilingual CLIR tasks adopt the query translation method using MT systems or machine-readable dictionaries (Kishida et al., 2005). Recent research work also proves that MT-based query translation could achieve comparable performance to other methods (Kishida et al., 2005; Nunzio et al., 2005). Considering more and more MT systems are being used in CLIR, it is of significance to carefully analyze how the performance of MT system may influence the retrieval effectiveness.

3 System Description

3.1 The Rule-Based ECMT System

The MT system used in this research is a rule-based ECMT system. The translation quality of this ECMT system is comparable to the best commercial ECMT systems. The basis of the system is semantic transfer (Amano et al., 1989).

Translation resources comprised in this system include a large dictionary and a rule base. The rule base consists of rules of different functions such as analysis, transfer and generation.

3.2 KIDS IR System

KIDS is an information retrieval engine that is based on morphological analysis (Sakai et al., 2003). It employs the Okapi/BM25 term weighting scheme, as fully described in (Robertson & Walker, 1999; Robertson & Sparck Jones, 1997).

To focus our study on the relationship between MT performance and retrieval effectiveness, we do not use techniques such as pseudo-relevance feedback although they are available and are known to improve IR performance.

4 Experimental Method

To obtain MT systems of varying quality, we degrade the rule-based ECMT system by impairing the translation resources comprised in the system. Then we use the degraded MT systems to translate the queries and evaluate the translation quality. Next, we submit the translated queries to the KIDS system and evaluate the retrieval performance. Finally we calculate the correlation between the variation of translation quality and the variation of retrieval effectiveness to analyze the relationship between MT performance and CLIR performance.

4.1 Degradation of MT System

In this research, we degrade the MT system in two ways. One is rule-based degradation, which is to decrease the size of the rule base by randomly removing rules from the rule base. For sake of simplicity, in this research we only consider transfer rules that are used for transferring the source language to the target language and keep other kinds of rules untouched. That is, we only consider the influence of transfer rules on translation quality[1]. We first randomly divide the rules into segments of equal size. Then we remove the segments from the rule base, one at each time and obtain a group of degraded rule bases. Afterwards, we use MT systems with the degraded rule bases to translate the queries and get groups of translated queries, which are of different translation quality.

The other is dictionary-based degradation, which is to decrease the size of the dictionary by randomly removing a certain number of word entries from the dictionary iteratively. Function words are not removed from the dictionary. Using MT systems with the degraded dictionaries, we also obtain groups of translated queries of different translation quality.

4.2 Evaluation of Performance

We measure the performance of the MT system by translation quality and use NIST score as the evaluation measure (Doddington, 2002). The NIST scores reported in this paper are generated by NIST scoring toolkit[2].

For retrieval performance, we use Mean Average Precision (MAP) as the evaluation measure (Voorhees, 2003). The MAP values reported in this paper are generated by trec_eval toolkit[3], which is the standard tool used by TREC for evaluating an ad hoc retrieval run.

5 Experiments

5.1 Data

The experiments are conducted on the TREC5&6 Chinese collection. The collection consists of document set, topic set and the relevance judgment file.

The document set contains articles published in People's Daily from 1991 to 1993, and news articles released by the Xinhua News Agency in 1994 and 1995. It includes totally 164,789 documents. The topic set contains 54 topics. In the relevance judgment file, a binary indication of relevant (1) or non-relevant (0) is given.

<top>
<num> Number: CH41
<C-title> 京九铁路的桥梁隧道工程
<E-title> Bridge and Tunnel Construction for the Beijing-Kowloon Railroad
<C-desc> Description:
京九铁路，桥梁，隧道，贯通，特大桥，
<E-desc> Description:
Beijing-Kowloon Railroad, bridge, tunnel, connection, very large bridge
<C-narr> Narrative:
相关文件必须提到京九铁路的桥梁隧道工程，包括地点、施工阶段、长度.
<E-narr> Narrative:
A relevant document discusses bridge and tunnel construction for the Beijing-Kowloon Railroad, including location, construction status, span or length.
</top>

Figure 1. Example of TREC Topic

5.2 Query Formulation & Evaluation

For each TREC topic, three fields are provided: *title*, *description* and *narrative*, both in Chinese and English, as shown in figure 1. The *title* field is the statement of the topic. The *description*

[1] In the following part of this paper, rules refer to transfer rules unless explicitly stated.

[2] The toolkit could be downloaded from: http://www.nist.gov/speech/tests/mt/resources/scoring.htm

[3] The toolkit could be downloaded from: http://trec.nist.gov/trec_eval/trec_eval.7.3.tar.gz

field lists some terms that describe the topic. The *narrative* field provides a complete description of document relevance for the assessors. In our experiments, we use two kinds of queries: *title queries* (use only the *title* field) and *desc queries* (use only the *description* field). We do not use *narrative* field because it is the criteria used by the assessors to judge whether a document is relevant or not, so it usually contains quite a number of unrelated words.

Title queries are one-sentence queries. When use NIST scoring tool to evaluate the translation quality of the MT system, reference translations of source language sentences are required. NIST scoring tool supports multi references. In our experiments, we introduce two reference translations for each title query. One is the *Chinese title* (C-title) in *title* field of the original TREC topic (reference translation 1); the other is the translation of the title query given by a human translator (reference translation 2). This is to alleviate the bias on translation evaluation introduced by only one reference translation. An example of title query and its reference translations are shown in figure 2. Reference 1 is the *Chinese title* provided in original TREC topic. Reference 2 is the human translation of the query. For this query, the translation output generated by the MT system is "在中国的机器人技术研究". If only use reference 1 as reference translation, the system output will not be regarded as a good translation. But in fact, it is a good translation for the query. Introducing reference 2 helps to alleviate the unfair evaluation.

Title Query: CH27
<query> Robotics Research in China
<reference 1> 中国在机器人方面的研制
<reference 2> 中国的机器人技术

Figure 2. Example of Title Query

A desc query is not a sentence but a string of terms that describes the topic. The term in the desc query is either a word, a phrase or a string of words. A desc query is not a proper input for the MT system. But the MT system still works. It translates the desc query term by term. When the term is a word or a phrase that exists in the dictionary, the MT system looks up the dictionary and takes the first translation in the entry as the translation of the term without any further analysis. When the term is a string of words such as "number(数量) of(的) infections(感染)", the system translates the term into "感染数量". Besides using the *Chinese description* (C-desc) in the *description* field of the original TREC topic as the reference translation of each desc query, we also have the human translator give another reference translation for each desc query. Comparison on the two references shows that they are very similar to each other. So in our final experiments, we use only one reference for each desc query, which is the *Chinese description* (C-desc) provided in the original TREC topic. An example of desc query and its reference translation is shown in figure 3.

Desc Query: CH22
<query> malaria, number of deaths, number of infections
<reference> 疟疾，死亡人数，感染病例

Figure 3. Example of Desc Query

5.3 Runs

Previous studies (Kwok, 1997; Nie et al., 2000) proved that using words and n-grams indexes leads to comparable performance for Chinese IR. So in our experiments, we use bi-grams as index units.

We conduct following runs to analyze the relationship between MT performance and CLIR performance:

- *rule-title*: MT-based title query translation with degraded rule base
- *rule-desc*: MT-based desc query translation with degraded rule base
- *dic-title*: MT-based title query translation with degraded dictionary
- *dic-desc*: MT-based desc query translation with degraded dictionary

For baseline comparison, we conduct Chinese monolingual runs with title queries and desc queries.

5.4 Monolingual Performance

The results of Chinese monolingual runs are shown in Table 1.

Run	MAP
title-cn1	0.3143
title-cn2	0.3001
desc-cn	0.3514

Table 1. Monolingual Results

title-cn1: use reference translation 1 of each title query as Chinese query

title-cn2: use reference translation 2 of each title query as Chinese query

desc-cn: use reference translation of each desc query as Chinese query

Among all the three monolingual runs, *desc-cn* achieves the best performance. *Title-cn1* achieves better performance than *title-cn2*, which indicates directly using *Chinese title* as Chinese query performs better than using human translation of title query as Chinese query.

5.5 Results on Rule-Based Degradation

There are totally 27,000 transfer rules in the rule base. We use all these transfer rules in the experiment of rule-based degradation. The 27,000 rules are randomly divided into 36 segments, each of which contains 750 rules. To degrade the rule base, we start with no degradation, then we remove one segment at each time, up to a complete degradation with all segments removed. With each of the segment removed from the rule base, the MT system based on the degraded rule base produces a group of translations for the input queries. The completely degraded system with all segments removed could produce a group of rough translations for the input queries.

Figure 4 and figure 5 show the experimental results on title queries (*rule-title*) and desc queries (*rule-desc*) respectively.

Figure 4(a) shows the changes of translation quality of the degraded MT systems on title queries. From the result, we observe a successive change on MT performance. The fewer rules, the worse translation quality achieves. The NIST score varies from 7.3548 at no degradation to 5.9155 at complete degradation. Figure 4(b) shows the changes of retrieval performance by using the translations generated by the degraded MT systems as queries. The MAP varies from 0.3126 at no degradation to 0.2810 at complete degradation. Comparison on figure 4(a) and 4(b) indicates similar variations between translation quality and retrieval performance. The better the translation quality, the better the retrieval performance is.

Figure 5(a) shows the changes of translation quality of the degraded MT systems on desc queries. Figure 5(b) shows the corresponding changes of retrieval performance. We observe a similar relationship between MT performance and retrieval performance as to the results based

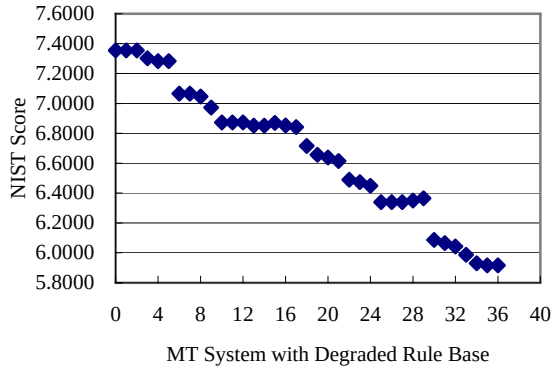

Figure 4(a). MT Performance on Rule-based Degradation with Title Query

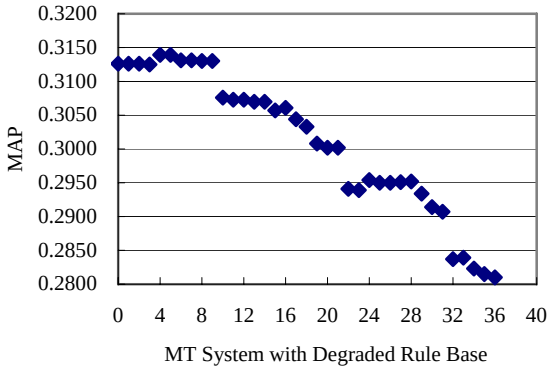

Figure 4(b). Retrieval Effectiveness on Rule-based Degradation with Title Query

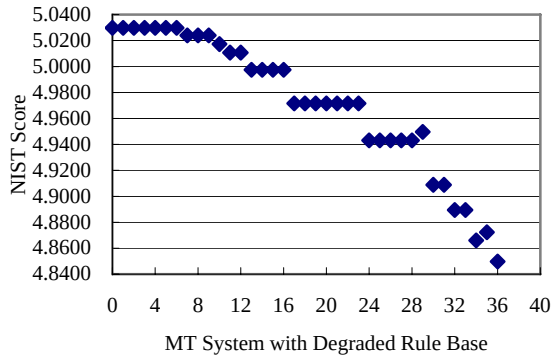

Figure 5(a). MT Performance on Rule-based Degradation with Desc Query

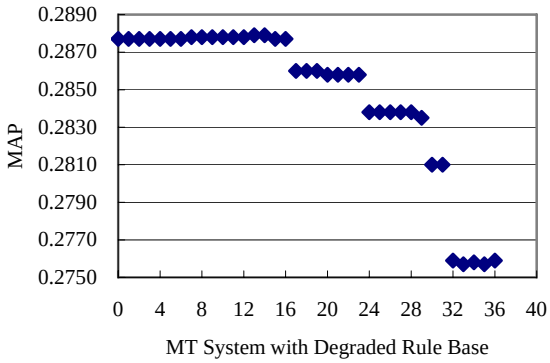

Figure 5(b). Retrieval Effectiveness on Rule-based Degradation with Desc Query

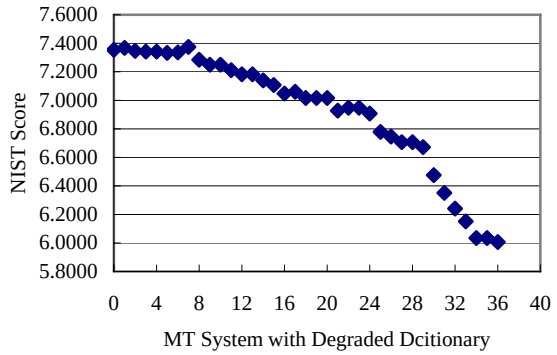

Figure 6(a). MT Performance on Dictionary-based Degradation with Title Query

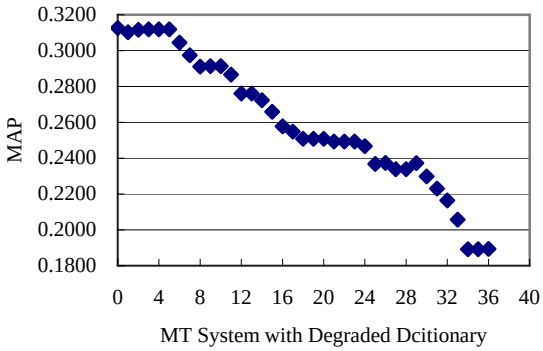

Figure 6(b). Retrieval Effectiveness on Dictionary-based Degradation with Title Query

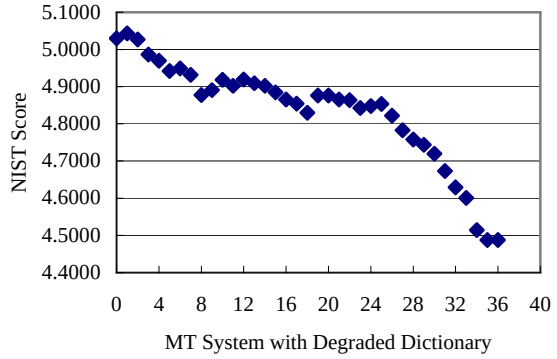

Figure 7(a). MT Performance on Dictionary-based Degradation with Desc Query

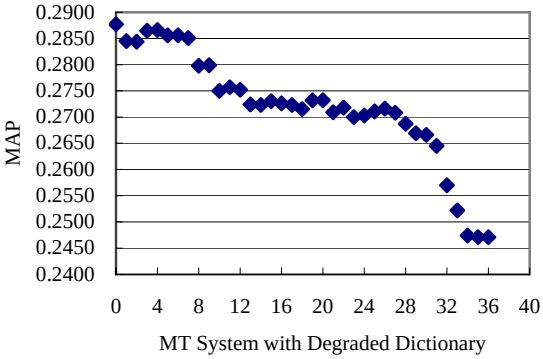

Figure 7(b). Retrieval Effectiveness on Dictionary-based Degradation with Desc Query

on title queries. The NIST score varies from 5.0297 at no degradation to 4.8497 at complete degradation. The MAP varies from 0.2877 at no degradation to 0.2759 at complete degradation.

5.6 Results on Dictionary-Based Degradation

The dictionary contains 169,000 word entries. To make the results on dictionary-based degradation comparable to the results on rule-based degradation, we degrade the dictionary so that the variation interval on translation quality is similar to that of the rule-based degradation. We randomly select 43,200 word entries for degradation. These word entries do not include function words. We equally split these word entries into 36 segments. Then we remove one segment from the dictionary at each time until all the segments are removed and obtain 36 degraded dictionaries. We use the MT systems with the degraded dictionaries to translate the queries and observe the changes on translation quality and retrieval performance. The experimental results on title queries (*dic-title*) and desc queries (*dic-desc*) are shown in figure 6 and figure 7 respectively. From the results, we also observe a similar relationship between translation quality and retrieval performance as what we have observed in the rule-based degradation. For both title queries and desc queries, the larger the dictionary size, the better the NIST score and MAP is. For title queries, the NIST score varies from 7.3548 at no degradation to 6.0067 at complete degradation. The MAP varies from 0.3126 at no degradation to 0.1894 at complete degradation. For desc queries, the NIST score varies from 5.0297 at no degradation to 4.4879 at complete degradation. The MAP varies from 0.2877 at no degradation to 0.2471 at complete degradation.

5.7 Summary of the Results

Here we summarize the results of the four runs in Table 2.

Run	NIST Score	MAP
title queries		
No degradation	7.3548	0.3126
Complete: *rule-title*	5.9155	0.2810
Complete: *dic-title*	6.0067	0.1894
desc queries		
No degradation	5.0297	0.2877
Complete: *rule-desc*	4.8497	0.2759
Complete: *dic-desc*	4.4879	0.2471

Table 2. Summary of Runs

6 Discussion

Based on our observations, we analyze the correlations between NIST scores and MAPs, as listed in Table 3. In general, there is a strong correlation between translation quality and retrieval effectiveness. The correlations are above 95% for all of the four runs, which means in general, a better performance on MT will lead to a better performance on retrieval.

Run	Correlation
rule-title	0.9728
rule-desc	0.9500
dic-title	0.9521
dic-desc	0.9582

Table 3. Correlation Between Translation Quality & Retrieval Effectiveness

6.1 Impacts of Query Format

For Chinese monolingual runs, retrieval based on desc queries achieves better performance than the runs based on title queries. This is because a desc query consists of terms that relate to the topic, i.e., all the terms in a desc query are precise query terms. But a title query is a sentence, which usually introduces words that are unrelated to the topic.

Results on bilingual retrieval are just contrary to monolingual ones. Title queries perform better than desc queries. Moreover, MAP at no degradation for title queries is 0.3126, which is about 99.46% of the performance of monolingual run *title-cn1*, and outperforms the performance of *title-cn2* run. But MAP at no degradation for desc queries is 0.2877, which is just 81.87% of the performance of the monolingual run *desc-cn*. Comparison on the results shows that the MT system performs better on title queries than on desc queries. This is reasonable because desc queries are strings of terms, however the MT system is optimized for grammatically correct sentences rather than word-by-word translation. Considering the correlation between translation quality and retrieval effectiveness, it is rational that title queries achieve better results on retrieval than desc queries.

6.2 Impacts of Rules and Dictionary

Table 4 shows the fall of NIST score and MAP at complete degradation compared with NIST score and MAP achieved at no degradation.

Comparison on the results of title queries shows that similar variation of translation quality leads to quite different variation on retrieval effectiveness. For *rule-title* run, 19.57% reduction in translation quality results in 10.11% reduction in retrieval effectiveness. But for *dic-title* run, 18.33% reduction in translation quality results in 39.41% reduction in retrieval effectiveness. This indicates that retrieval effectiveness is more sensitive to the size of the dictionary than to the size of the rule base for title queries. Why dictionary-based degradation has stronger impact on retrieval effectiveness than rule-based degradation? This is because retrieval systems are typically more tolerant of syntactic than semantic translation errors (Fluhr, 1997). Therefore although syntactic errors caused by the degradation of the rule base result in a decrease of translation quality, they have smaller impacts on retrieval effectiveness than the word translation errors caused by the degradation of dictionary.

For desc queries, there is no big difference between dictionary-based degradation and rule-based degradation. This is because the MT system translates the desc queries term by term, so degradation of rule base mainly results in word translation errors instead of syntactic errors. Thus, degradation of dictionary and rule base has similar effect on retrieval effectiveness.

Run	NIST Score Fall	MAP Fall
title queries		
rule-title	19.57%	10.11%
dic-title	18.33%	39.41%
desc queries		
rule-desc	3.58%	4.10%
dic-desc	10.77%	14.11%

Table 4. Fall on Translation Quality & Retrieval Effectiveness

7 Conclusion and Future Work

In this paper, we investigated the effect of translation quality in MT-based CLIR. Our study showed that the performance of MT system and IR system correlates highly with each other. We further analyzed two main factors in MT-based CLIR. One factor is the query format. We concluded that title queries are preferred for MT-based CLIR because MT system is usually optimized for translating sentences rather than words. The other factor is the translation resources comprised in the MT system. Our observation showed that the size of the dictionary has a stronger effect on retrieval effectiveness than the size of the rule base in MT-based CLIR. Therefore in order to improve the retrieval effectiveness of a MT-based CLIR application, it is more

effective to develop a larger dictionary than to develop more rules. This introduces another interesting question relating to MT-based CLIR. That is how CLIR can benefit further from MT. Directly using the translations generated by the MT system may not be the best choice for the IR system. There are rich features generated during the translation procedure. Will such features be helpful to CLIR? This question is what we would like to answer in our future work.

References

Shin-ya Amano, Hideki Hirakawa, Hirosysu Nogami, and Akira Kumano. 1989. The Toshiba Machine Translation System. *Future Computing System*, 2(3):227-246.

Dina Demner-Fushman, and Douglas W. Oard. 2003. The Effect of Bilingual Term List Size on Dictionary-Based Cross-Language Information Retrieval. In *Proc. of the 36th Hawaii International Conference on System Sciences (HICSS-36)*, pages 108-117.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proc. of the Second International Conference on Human Language Technology (HLT-2002)*, pages 138-145.

Christian Fluhr. 1997. Multilingual Information Retrieval. In Ronald A Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue (Eds.), *Survey of the State of the Art in Human Language Technology*, pages 261-266, Cambridge University Press, New York.

Martin Franz, J. Scott McCarley, Todd Ward, and Wei-Jing Zhu. 2001. Quantifying the Utility of Parallel Corpora. In *Proc. of the 24th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-2001)*, pages 398-399.

David A. Hull and Gregory Grefenstette. 1996. Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval. In *Proc. of the 19th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-1996)*, pages 49-57.

Gareth Jones, Tetsuya Sakai, Nigel Collier, Akira Kumano and Kazuo Sumita. 1999. Exploring the Use of Machine Translation Resources for English-Japanese Cross-Language Infromation Retrieval. In *Proc. of MT Summit VII Workshop on Machine Translation for Cross Language Information Retrieval*, pages 15-22.

Kazuaki Kishida, Kuang-hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, Hsin-Hsi Chen, and Sung Hyon Myaeng. 2005. Overview of CLIR Task at the Fifth NTCIR Workshop. In *Proc. of the NTCIR-5 Workshop Meeting*, pages 1-38.

Wessel Kraaij. 2001. TNO at CLEF-2001: Comparing Translation Resources. In *Proc. of the CLEF-2001 Workshop*, pages 78-93.

Kui-Lam Kwok. 1997. Comparing Representation in Chinese Information Retrieval. In *Proc. of the 20th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-1997)*, pages 34-41.

Paul McNamee and James Mayfield. 2002. Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. In *Proc. of the 25th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-2002)*, pages 159-166.

Jian-Yun Nie, Jianfeng Gao, Jian Zhang, and Ming Zhou. 2000. On the Use of Words and N-grams for Chinese Information Retrieval. In *Proc. of the Fifth International Workshop on Information Retrieval with Asian Languages (IRAL-2000)*, pages 141-148.

Giorgio M. Di Nunzio, Nicola Ferro, Gareth J. F. Jones, and Carol Peters. 2005. CLEF 2005: Ad Hoc Track Overview. In C. Peters (Ed.), *Working Notes for the CLEF 2005 Workshop*.

Stephen E. Robertson and Stephen Walker. 1999. Okapi/Keenbow at TREC-8. In *Proc. of the Eighth Text Retrieval Conference (TREC-8)*, pages 151-162.

Stephen E. Robertson and Karen Sparck Jones. 1997. Simple, Proven Approaches to Text Retrieval. Technical Report 356, Computer Laboratory, University of Cambridge, United Kingdom.

Tetsuya Sakai, Makoto Koyama, Masaru Suzuki, and Toshihiko Manabe. 2003. Toshiba KIDS at NTCIR-3: Japanese and English-Japanese IR. In *Proc. of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering (NTCIR-3)*, pages 51-58.

Ellen M. Voorhees. 2003. Overview of TREC 2003. In *Proc. of the Twelfth Text Retrieval Conference (TREC 2003)*, pages 1-13.

Jinxi Xu and Ralph Weischedel. 2000. Cross-lingual Information Retrieval Using Hidden Markov Models. In *Proc. of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pages 95-103.