CLIR using Web Directory at NTCIR4

Fuminori Kimura
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, Japan
fumino-k@is.aist-nara.ac.jp

Akira Maeda
Department of Media Technology, College of Information Science and Engineering,
Ritsumeikan University
1-1-1 Noji-Higashi, Kusatsu, Shiga, Japan
amaeda@media.ritsumei.ac.jp

Shunsuke Uemura
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, Japan
uemura@is.aist-nara.ac.jp

Abstract

In this paper, we propose a CLIR method which employs a Web directory provided in multiple language versions (such as Yahoo!). In the proposed method, feature terms are first extracted from Web documents for each category in the source and the target languages. determined beforehand by comparing similarities between categories across languages. In advance, category matching is conducted in order to category pairs between categories across languages. Using these category pairs, we intend to resolve ambiguities of simple dictionary translation by narrowing the categories to be retrieved in the target language.

At NTCIR-4, we participated in the Japanese-English cross-language track. We submitted TITLE run and DESCRIPTION run. From the analysis of the experimental results, we found that the translation failure of proper nouns causes serious influence for retrieval results.

Keywords: *Japanese-English Cross-Language Information Retrieval, query translation, Web directory.*

1 Introduction

Approaches to CLIR can be classified into three categories; document translation, query translation, and the use of inter-lingual representation. The approach based on translation of target documents has the advantage of utilizing existing machine translation systems, in which more content information can be used for disambiguation. Thus, in general, it achieves better retrieval effectiveness than those based on query translation[8]. However, since it is impractical to translate a huge document collection beforehand and it is difficult to extend this method to new languages, this approach is not suitable for multilingual, large-scale, and frequently-updated collection of the Web. The second approach transfers both documents and queries into an inter-lingual representation, such as bilingual thesaurus classes or a language-independent vector space. The latter approach requires a training phase using a bilingual (parallel or comparable) corpus as a training data.

The major problem in the approach based on the translation and disambiguation of queries is that the queries submitted from ordinary users of Web search engines tend to be very short (approximately two words on average[3]) and usually consist of just an enumeration of keywords (i.e. no context). However, this approach has an advantage that the translated queries can simply be fed into existing monolingual search engines. In this approach, a source language query is first translated into target language using a bilingual dictionary, and translated query is disambiguated. Our method falls into this category.

It is pointed out that corpus-based disambiguation methods are heavily affected by the difference in domain between query and corpus. Hull[2] suggests that the difference between query and corpus may cause bad influence on retrieval effectiveness in the methods that use parallel or comparable corpora. Lin

et al.[6] conducted comparative experiments among three monolingual corpora that have different domains and sizes, and has concluded that large-scale and domain-consistent corpus is needed for obtaining useful co-occurrence data.

On the Web retrieval, which is the target of our research, the system has to cope with queries in many different kinds of topics. However, it is impractical to prepare corpora that cover any possible domains. In our previous paper[4, 5], we proposed a CLIR method which uses documents in a Web directory that has query language and target language versions (such as Yahoo!), instead of using existing corpora, in order to improve the retrieval effectiveness.

At NTCIR-4, we participated in the Japanese-English cross-language track. We submitted TITLE run and DESCRIPTION run.

2 Proposed System

Figure 1 illustrates the outline of the proposed system. Our system uses two language versions of a Web directory. One version is the query language, the others is the target languages to be retrieved . From these language versions, category correspondences between languages are estimated in advance. The proposed system has two phases of processing, preprocessing and retrieval processing.

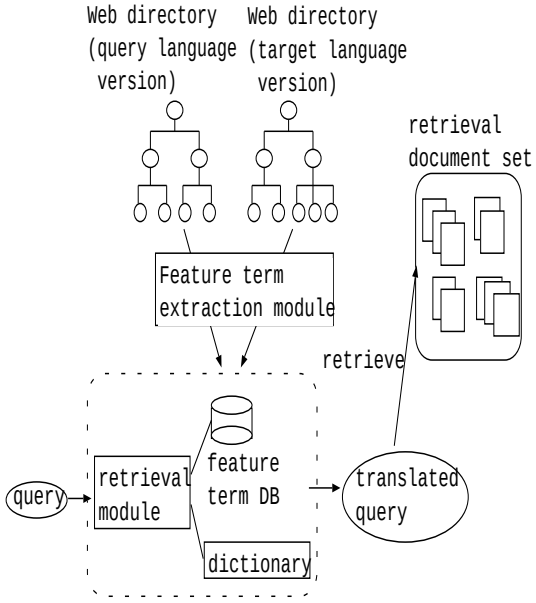

Figure 1. Outline of proposed system.

The preprocessing consists of the following four steps: 1) term extraction from Web documents in each category, 2) feature term extraction, 3) translation of feature terms, and 4) estimation of category correspondences between different languages. Figure 2 illustrates the flow of the preprocessing. This example shows a case that category a in query language corresponds to a category in target language. First, the system extracts terms from Web documents which belong to category a (1)(a). Secondly, the system calculates the weights of the extracted terms. Then higher-weighted terms are extracted as the feature term set f_a of category a (1)(b). Thirdly, the system translates the feature term set f_a into target language (1)(c). Lastly, the system estimates the corresponding category of category a from target language (2). These category pairs are used on retrieval.

At the retrieval phase, the system executes following procedures. First, the system estimates appropriate category for the query in the query language. Next, the system selects the corresponding category in the target language using the pre-estimated category pairs. Finally, the system retrieves the target document set.

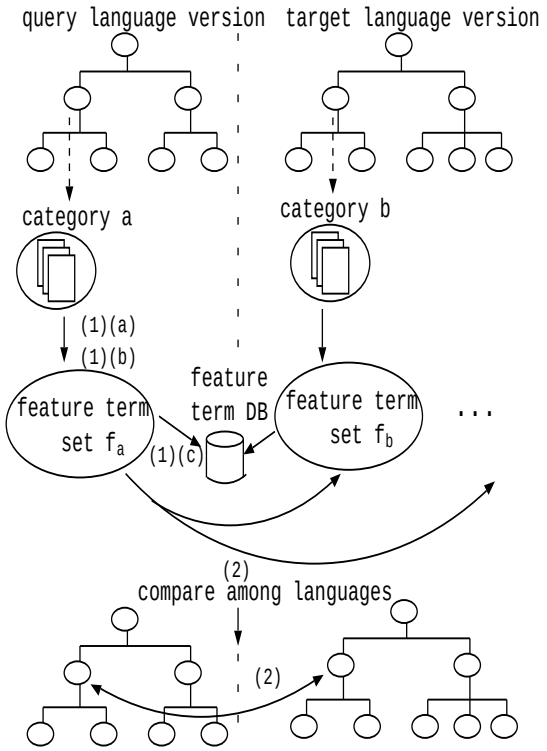

Figure 2. Preprocessing.

2.1 Preprocessing

2.1.1 Feature Term Extraction

The feature of each category is represented by its feature term set. Feature term set is a set of terms that seem to distinguish the category. The feature term set of each category is extracted in the following steps: First, the system extracts terms from Web documents that belong to a given category. Second, the system calculates the weights of the extracted terms. Lastly,

top n ranked terms are extracted as the feature term set of the category.

Weights of feature terms are calculated by TF·ICF (term frequency · inverse category frequency). TF·ICF is a variation of TF·IDF (term frequency · inverse document frequency). Instead of using a document as the unit, TF·ICF calculates weights by category.

2.1.2 Category Matching

In the proposed system, category matching between languages needs to be done in advance in order to retrieve. Arbitrary methods may be employed for category matching. For example, matching categories can be estimated by comparing the feature term sets of each category. Otherwise, categories can be matched manually if the number of categories to be matched is relatively small. In our experiments at NTCIR-4, we manually matched the 13 categories at the top levels of Japanese and English versions of Yahoo.

2.2 Retrieval Processing

2.2.1 Retrieval

Figure 3 illustrates the processing flow of a retrieval. When the user submits a query, the following four steps are processed.

First, the system calculates the relevance between the query and each category in the query language (1), and determines the relevant category of the query in the query language (2). The relevance between the query and each category is calculated by the inner product between query terms and the feature term set of the target category.

Second, the corresponding category in the target language is selected by using category correspondences between languages (3). Third, the query is translated into the target language by using a dictionary and disambiguate translations using the feature term set of the corresponding category (4). Finally, the system retrieves documents in the retrieval document set (5).

2.2.2 Query Term Translation

In order to translate query terms, the system executes the following procedures. Figure 4 illustlates these procedures. First, for each feature term, the system looks up the term in a bilingual dictionary and extracts all translation candidates for the feature term. Next, the system checks whether each translation candidate exists in the feature term set of the corresponding category mentioned in 2.1.2. Lastly, the highest-weighted translation candidate in the feature term set of the target category is selected as the translation of the feature term.

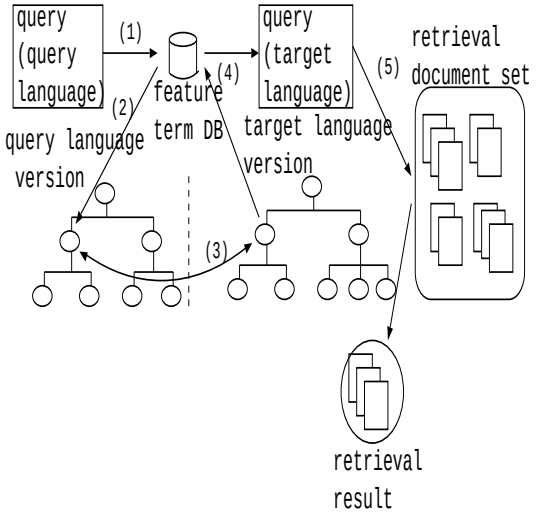

Figure 3. Processing on Retrieval.

If no translation candidate for a feature term exists in the feature term set of the target category, that term is ignored in the comparison. However, there are some cases that the source language term itself is useful as a feature term in the target language. For example, some English terms (mostly abbreviations) are commonly used in documents written in other languages (e.g. “WWW”, “HTM”, etc.). Therefore, in case that no translation candidate for a feature term exists in the feature term set of the target category, the feature term itself is checked whether it exists in the feature term set of the target category. If it exists, the feature term itself is treated as the translation of the feature term in the target category.

As an example, we consider that an English term “system” is translated into Japanese for the category “コンピュータとインターネット＞ソフトウェア ＞セキュリティ (Computers and Internet ＞Software ＞Security)” (hereafter called “セキュリティ” for short). The English term “system” has the following translation candidates in a dictionary; “宇宙 (universe/space)”, “方法 (method)”, “組織 (organization)”, “器官 (organ)”, “システム (system)”, etc. We check each of these translation candidates in the feature term set of the category “セキュリティ.” Then the highest-weighted term of these translation candidates in the category “セキュリティ” is determined as the translation of the English term “system” in this category. If no translation candidate exists in the feature term set of the category “セキュリティ,” the English term “system” itself is treated as the translation.

3 Experiments

We have conducted experiments of the proposed method at the Japanese-English cross-language track of NTCIR-4. We use TITLE field and DESCRIPTION

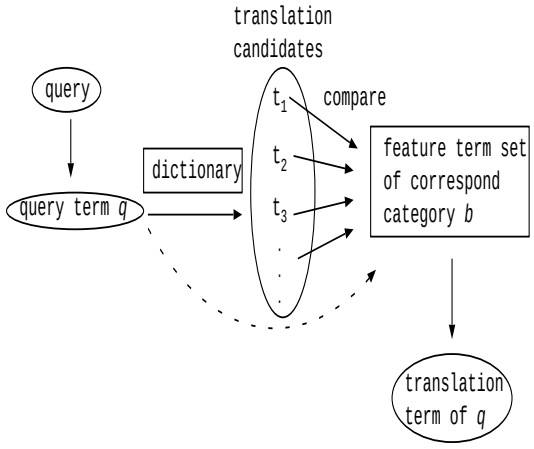

Figure 4. Query term translation.

field in the Japanese query set.

3.1 System description

We used English and Japanese versions of Yahoo! as corpora for disambiguation. The English subset consists of 84,835 categories and 800,000 documents except for the category "Regional." The Japanese subset consists of 3,175 and 34,443 documents except for the category "地域情報 (Regional)." The reason we exclude these two categories is that documents in these categories are unsuitable for Japanese-English translation because these documents are written about regions all over the world. In these experiments, we merged all sub-categories two levels below the top page of each version into 13 categories linked from the top page.

On the extraction of terms from English Web documents, we exchanged conjugations the original form, and eliminated stop words. We used the stop word list in chapter 7 of "Information Retrieval: Data Structures and Algorithms"[1]. On the extraction from Japanese Web documents, we used "ChaSen"[7][1] morphological analyzer. We extracted noun, verb, and adjective from documents.

On the calculation of the weights of feature terms, we calculated used the revised version of TF·ICF mentioned in Section 2.1.1. In the normal TF·ICF, the number of terms appearing is counted after category merging. On the other hand, in the revised version, it is counted before category merging. The revised TF·ICF is calculated as follows:

$$tf \cdot icf_{rev}(t_i, c) = \frac{f(t_i)}{N_c} \cdot \log \frac{N_b}{n_{b_i}} + 1$$

where n_{b_i} is the number of categories that contain the term t_i before category merging, and N_b is the number of all categories in the directory before category merging.

In this experiment, we fixed the number of feature term in each category to 10,000 terms. Category matching is done manually as mentioned in Section 2.1.2.

At formulating the queries, we used "ChaSen" for extract query terms from TITLE and DESCRIPTION fields. At the term translation, we used "EDR Electronic Dictionary: Jpn.-Eng. Bilingual Dictionary."[2] At retrieval, we used "SMART"[3] retrieval system.

Table 1. Average precision and R-precision of each run.

	Ave.precision (relax)	R-precision (relax)
TITLE	0.0255	0.0531
DESC	0.0063	0.0199
TITLE-revise	0.0289	0.0594
DESC-revise	0.0100	0.0288
TITLE-proper	0.0559	0.0979
DESC-proper	0.0204	0.0431

3.2 Evaluation and discussion

Table 1 shows the result of our experiments. "TITLE" and "DESC" are our submitted runs. Other 4 runs are additional runs. The result of the submitted runs is not satisfactory. We analyze failure as follows:

1. Proper nouns in the queries were not translated.
2. Xinhua News Service mistakenly excluded from the test collections used in the experiments.
3. Term extraction from queries was insufficient.

Proper nouns are usually essential terms for retrieval, and the failure to translate proper nouns might cause heavy decrease in retrieval effectiveness. However, in our submitted runs, we failed to deal with terms including capital letters, which is usually the case for proper nouns. Therefore, we conducted additional runs which solve this problem. The results of these additional runs are shown at "TITLE-revise" and "DESC-revise" rows in Table 1. In these runs, some of the proper nouns which were not translated in the submitted runs were translated, thus the average precisions and R-precisions are improved in each run.

Another cause is failure of translation by bilingual dictionary. In general, most of proper nouns, for example person's name or place-name and so on,

[1] http://chasen.aist-nara.ac.jp/

[2] http://www.jsa.co.jp/EDR/

[3] ftp://ftp.cs.cornell.edu/pub/smart/

Table 2. Average precision and R-precision of each run whithout Xinhua News Service document set.

	Ave.precision (relax)	R-precision (relax)
TITLE-revise	0.0535	0.0830
DESC-revise	0.0213	0.0357
TITLE-proper	0.0932	0.1292
DESC-proper	0.0402	0.0531

are not contained in bilingual dictionary. Therefore, our method using bilingual dictionary cannot translate such proper nouns. In order to clarify the effect of this failure, we also experimented in the case of translating proper nouns manually. The results are shown at "TITLE-proper" and "DESC-proper" in rows in Table 1. In terms of average precision, "TITLE-proper" is improved 3.04 point, and "DESC-proper" is improved 1.04 point as compared with submitted runs. These results show that the translation of proper nouns has much influence for retrieval.

The English test collection for NTCIR-4 CLIR task consists of 5 document sets. However, we failed to obtain Xinhua News Service document set. Consequently, our method cannot retrieve any relevant documents contained in Xinhua News Service document set, although these documents are contained in relevant document list. Therefore, the retrieval effectiveness, especially recall factor, is indisputably decreased by this mistake. Table 2 shows evaluation result in the case of evaluation by relevant document list without Xinhua News Service document set. In terms of average precision, "TITLE-proper" without Xinhua News Service document set is improved 3.73 point, and "DESC-proper" without Xinhua News Service document set is improved 1.98 point than the case with Xinhua News Service document set.

When formulating a query, insufficient term extraction from topics is serious factor of decline in average precision and R-precision. One of the failures in term extraction from topics is term segmentation. Our method extracted terms from topics by "ChaSen", then some terms were separated although the term is one term. For example, person's name "フローレンス (Florence)" is separated into three terms ;" フロー", "レン" and "ス." Most of these terms cannot be translated. Besides, it is the factor of decline in retrieval effectiveness that our method did not deal with compound terms. This factor causes mistranslation. For example, Japanese term "電子商取引 (e-commerce)" is separated into two terms, "電子 (electron)" and "商取引 (commercial transaction)." Then, our method did not acquire translated term "e-commerce."

3.3 Conclusions

We proposed a method using a Web directory for CLIR. The proposed method is independent of a particular domain because it uses documents in a Web directory as the corpus. Our method is particularly effective for the case that the document collection covers wide range of domains such as the Web. Besides, our method does not require expensive linguistic resources except for a dictionary. Therefore, our method can easily be extended to other languages as long as the language versions of a Web directory exist and the dictionary can be obtained.

This paper described our submitted runs and additional runs at NTCIR-4 Japanese-English cross-language track. Though the results of evaluation were insufficient, we find some problems for our method. Our future work is to solve these problems; translate proper nouns and extract query terms.

References

[1] W. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms, chapter 7*. Prentice-Hall, 1992.

[2] D. A. Hull. Using structured queries for disambiguation in cross-language information retrieval. *Electronic Working Notes of the AAAI Symposium on Cross-Language Text and Speech Retrieval*, March 1997.

[3] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real user queries on the web. *Information Processing & Management*, 36(2):207–227, March 2000.

[4] F. Kimura, A. Maeda, M. Yoshikawa, and S. Uemura. Cross-language information retrieval based on category matching between language versions of a web directory. *The 6th International Workshop on Information Retrieval with Asian Languages (IRAL2003) in conjunction with 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 153–159, July 2003.

[5] F. Kimura, A. Maeda, M. Yoshikawa, and S. Uemura. Cross-language information retrieval using web directories. *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM '03)*, pages 911–914, August 2003.

[6] C.-J. Lin, W.-C. Lin, G.-W. Bian, and H.-H. Chen. Description of the ntu japanese-english cross-lingual information retrieval system used for ntcir workshop. *First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 145–148, August 1999.

[7] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara. *Japanese morphological analysis system ChaSen version 2.3.3 manual*. 2003.

[8] T. Sakai. Mt-based japanese-enlish cross-language ir experiments using the trec test collections. *Proceedings of The Fifth International Workshop on Information Retrieval with Asian Languages (IRAL2000)*, pages 181–188, September 2000.