Analysis of Appropriate Category Level of Web Directory for Cross-Language Information Retrieval

Fuminori Kimura[*] Akira Maeda[†] Kenji Hatano[‡] Jun Miyazaki[§] Shunsuke Uemura[¶]

Abstract— **In this paper, we analyzed appropriate category level of Web directory for Cross-Language Information Retrieval (CLIR). Our proposed method for CLIR is based on estimating domains of the query using hierarchic structures of Web directories. Therefore, it is necessary for correct domain estimation to detect appropriate category level of Web directory. We conducted experiments of retrieval using four category level in order to detect appropriate category levels of Web directory. We found that 2nd lv or 3rd lv is appropriate for CLIR.**

Keywords:* *Cross-Language Information Retrieval, Web directory, appropriate category level

1 Introduction

With the worldwide popularity of the Internet, more and more languages are being used for Web documents, and it is now much easier to access documents written in foreign languages. However, existing Web search engines only support the retrieval of documents that are written in the same language as the query, so there is no efficient way for monolingual users to retrieve documents written in non-native languages. There might also be cases, depending on the user's needs, where valuable information is written in a language other than the user's native language. To satisfy these needs in a typical monolingual retrieval system, users have to manually translate queries themselves using a dictionary. This method is difficult for the user. To meet these needs, there has been intensive research in recent years on Cross-Language Information Retrieval (CLIR), a technique for retrieving documents written in one language using a query written in another language.

A variety of methods, including the use of corpus statistics to translate terms and the disambiguation of translated terms, have been investigated and some useful results have been obtained. However, corpus-based disambiguation methods are significantly affected by the domain of the training corpus, so they may be much less effective for retrieval in other domains. In addition, since the Web consists of documents in various domains or genres, methods used for CLIR of Web documents should be independent of specific domains.

2 Related Work

In Cross-Language Information Retrieval, the major approach is query translation approach [6]. In this approach, the system translates only query terms. The major problem in using an approach based on the translation and disambiguation of queries is that queries submitted by ordinary users of Web search engines tend to be very short. They consist of approximately two words on average [3], and are usually just an enumeration of keywords (i.e. there is no context). However, one advantage of this approach is that the translated queries can simply be fed into existing monolingual search engines. In this approach, a source language query is first translated into the target language using a bilingual dictionary, and the translated query is then disambiguated. Our method falls into this category.

We should point out that corpus-based disambiguation methods are significantly affected by differences between the domain of the query and the corpus. Hull suggests that these differences may adversely affect the retrieval efficiency of methods that use parallel or comparable corpora [2]. Lin et al. conducted comparative experiments between three monolingual corpora that had different domains and sizes, and concluded that a large-scale, domain-consistent corpus is needed to obtain useful co-occurrence data [5].

In relation to Web retrieval, which is the target of our research, the system has to cope with queries on many

[*] Faculty of Culture and Information Science, Doshisha University, 1–3 Miyakodani Tatara, Kyoutanabe-shi, Kyoto, Japan, Email:jt-bnk04@mail.doshisha.ac.jp

[†] Department of Media Technology, College of Information Science and Engineering, Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu, Shiga, Japan, Email:amaeda@media.ritsumei.ac.jp

[‡] Faculty of Culture and Information Science, Doshisha University, 1–3 Miyakodani Tatara, Kyoutanabe-shi, Kyoto, Japan, Email:khatano@mail.doshisha.ac.jp

[§] Graduate School of Information Science, Nara Institute of Science and Technology , 8916–5 Takayama, Ikoma, Nara, Japan, Email:miyazaki@is.naist.jp

[¶] Faculty of Informatics, Nara Sangyo University, 3–12–1 Tatsuno-kita, Sango-cho, Ikoma-gun, Nara, Japan, Email:uemurashunsuke@nara-su.ac.jp

different topics. However, it is impractical to prepare corpora that cover every possible domain. In our previous paper [4], we proposed a CLIR method that uses documents in Web directories that have several language versions (such as Yahoo!), instead of using existing corpora, to improve retrieval effectiveness.

3 Cross-Language Information Retrieval Using Web Directories

Figure 1 illustrates the outline of the proposed system. This system consists of query and target language versions of Web Directory, each language versions of feature term database, bilingual dictionary, and retrieval target document set. The part surrounded by a dotted line in right side of Figure 1 illustrates components of translation processing for query.

The processing in our system can be divided into two phases. One is the preprocessing phase, which extracts feature terms from each category of a Web directory, and stores them in the feature term database in advance. Another is the retrieval phase, which translates the given query into the target language, and retrieves documents.

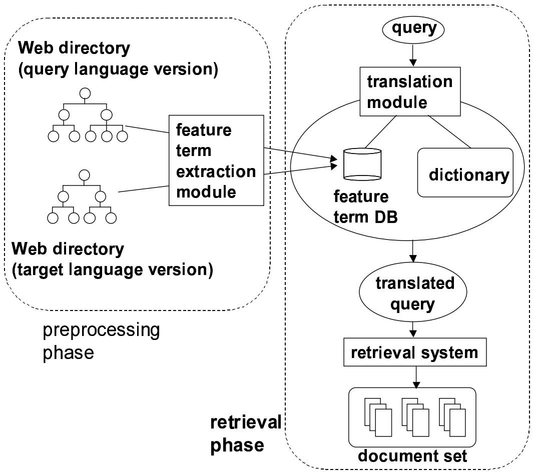

Figure 1: Outline of the proposed system.

3.1 Method of Category Merging

Each category in Web directory is useful to specify the fields of the query. However, some categories have insufficient web documents. The system cannot acquire sufficient statistical information to resolve translation disambiguation. This problem might be caused by the following reasons; one possible reason is that there are some categories which are too close in topic, and it might cause poor accuracy. Another possible reason is that some categories have insufficient amount of text in order to obtain statistically significant values for feature term extraction.

Considering the above observations, we might expect that the accuracy will be improved by merging child categories at some level in the category hierarchy in order to merge some categories similar in topic and to increase the amount of text in a category. Figure 2 illustrates the result of category merging. Each category existing some level in the category hierarchy includes all sub categories under the category.

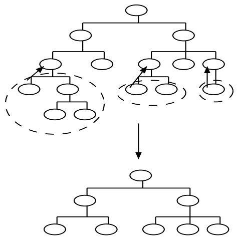

Figure 2: Category Merging.

3.2 Preprocessing Phase

In the preprocessing phase, the system conducts feature term extraction and category matching between the query and target languages in advance. This phase is illustrated in the left side of Figure 1. The following procedure is used for this phase:

1. Feature-term extraction
   For each category in all language versions of a Web directory,
   (a) extract terms from Web documents in the required category and calculate the weight of the terms.
   (b) extract the top n ranked terms as the feature terms of the category.
   (c) store the feature terms in the feature-term database.
2. Category matching between languages
   For each category in one language version, estimate the corresponding category in the other language version.

Note that the category matching method is not the focus of this paper. An arbitrary method can be used for category matching. For example, we could calculate the similarity between categories based on extracted feature terms, or we could manually match each category. The category pairs acquired by this process are used in retrieval.

3.3 Retrieval Phase

The right side of Figure 1 shows Retrieval phase. Detail of the processing flow for retrieval (the right side of Figure 1) is illustrated in Figure 3. First, the system estimates the relevant category of the query from the query

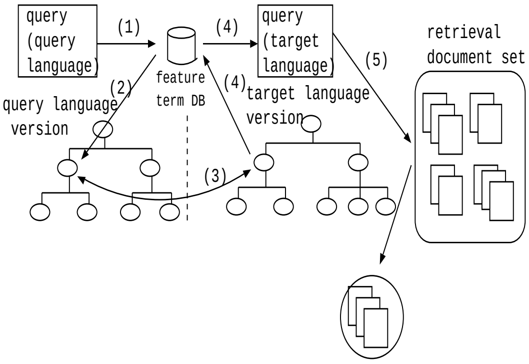

Figure 3: Flow of retrieval.

language version. Secondly, the system selects a category corresponding to the relevant category. Thirdly, the system translates the query terms into the target language using the feature-term set for the corresponding category. Finally, the system retrieves documents using the translated query. The procedure for the retrieval phase is as follows:

(1) For each category in the query language version, calculate the relevance between the query and the feature-term set for the category.

(2) Determine the category with the highest relevance as the relevant category for the query.

(3) Select the category corresponding to the most relevant category from the target language version.

(4) Translate the query terms into the target language using the feature-term set of the corresponding category.

(5) Retrieve documents using the translated query.

3.3.1 Selection of Relevant Category

The system calculates the relevance between the query and each category in the query language version (Figure 3 (1)), and determines the most relevant category to the query in the query language version (Figure 3 (2)). The relevance between the query and each category is calculated by multiplying the inner product between the query terms and the feature-term set of the target category by the angle of these two vectors. When there is more than one category whose relevance to the query exceeds a certain threshold, all are selected as relevant categories for the query.

Besides, corresponding category of the relevant category is selected from target language version (Figure 3 (3)). Feature term set of the selected corresponding category is used in order to disambiguate translation candidates.

3.3.2 Query Translation

Figure 4 illustrates the processing flow for query translation. This figure is detail of the Figure 3 (4). First, for each query term q, 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 is included in the feature-term set of the corresponding category. If it is, the system checks the weight of the candidate in the feature-term set. Lastly, the highest-weighted translation candidate in the feature-term set of the corresponding category is selected as the translation of the feature term.

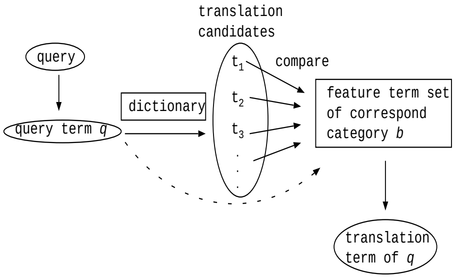

Figure 4: Translation of a query.

3.3.3 Retrieval of Documents

The system retrieves documents using queries translated by the method described in Section 3.3.2 (Figure 3 (5)). The documents to be retrieved need not be those registered in the Web directory. Instead, the system may use an existing retrieval system.

4 Experiments

We conducted experiments on the proposed method using English and Japanese versions of Yahoo! category. In these experiments, we used Japanese queries and retrieved English documents. The purpose of the experiments was to investigate what level of category merging for Web directory is most effective to improve the precision of CLIR. We conducted experiments in the four cases, used the category of Web directory is merged into top level from the top of Web directory(hereafter called the “1-lv” for short) or second level of it(“2-lv”) or third level of it(“3-lv”) or forth level of it(“4-lv”).

We also conducted experiments on the case of no disambiguation of translations for comparison (hereafter called the “baseline” for short). In the baseline, we used all the translation candidates in a dictionary as query terms, except for multi-word terms. Processing after the translation of a query was done using our proposed method. Besides, we conducted experiments on the case of using machine translation service on the web in order to translate query terms(hereafter called the "web translation" for short). Web translation cannot be directly compared with another cases because the dictionary of web translation is different from another cases. Then we experimented web translation for reference. In this experiment, we used "excite translation[1]".

4.1 Method of Experiments

In these experiments, we used document sets and queries presented in the CLIR task at The 3rd NTCIR Workshop[2] (hereafter called the "NTCIR3 test collection" for short). NTCIR is an evaluation workshop of retrieval. CLIR task is one of the tasks in NTCIR. We used two document sets from the NTCIR3 test collection: EIRB010, which consists of several English newspapers published in Taiwan from 1998 to 1999, and Mainichi Daily 1998-1999, which consists of English newspaper articles published in Japan from 1998 to 1999. The Japanese query set for the NTCIR3 test collection, which consists of 50 queries, was used in these experiments.

To resolve ambiguities, we used English and Japanese versions of Yahoo! category as the Web directory. We excluded all sub-categories in the category "Regional" in each version. We eliminated this category because it is unsuitable for translation since it consists of documents written about regions all over the world.

Table 1 shows the number of categories in each level. In English, 1-lv has 13 categories, 2-lv has 397 categories, 3-lv has 4066 categories and 4-lv has 8672 categories. In Japanese, 1-lv has 13 categories, 2-lv has 391 categories, 3-lv has 2953 categories and 4-lv has 3259 categories. We merged categories in order to resolve shortage of statistical information. However, some of the merged categories cannot acquire sufficient statistical information. We eliminate such categories that have less than 10,000 feature terms. Table 1 also shows the number of categories in each level after eliminating categories that have less than 10,000 feature terms. In English, 1-lv has 13 categories, 2-lv has 255 categories, 3-lv has 644 categories and 4-lv has 292 categories. In Japanese, 1-lv has 13 categories, 2-lv has 154 categories, 3-lv has 153 categories and 4-lv has 42 categories. In addition, category matching between languages was done manually.

In extracting terms from English Web documents, the terms were transformed into the original form, and stop words were eliminated. We used the stop word list published by Frakes and Baeza-Yates [1] and we used the Japanese morphological analyzer, "Chasen"[3]. In these experiments, to extract terms from Japanese Web documents, sentences were separated by Chasen, and the system extracted nouns, verbs, adjectives, and unknown terms.

Table 1: The number of categories in each level.

		1-lv	2-lv	3-lv	4-lv
English	all	13	397	4066	8672
	eliminated	13	255	644	292
Japanese	all	13	391	2953	3259
	eliminated	13	154	153	42

Table 2: The number of categories in 4 level for feature terms.

feature term	3,000	5,000	10,000	all
English	1185	674	292	8672
Japanese	233	115	42	3259

For translation, we used the "EDR Electronic Dictionary: Jpn.-Eng. Bilingual Dictionary"[4]. The average number of translation candidates for translating the Japanese queries in the NTCIR3 test collection was 5.17.

We used the query that were extracted from the "TITLE" fields of the Japanese query set in the NTCIR3 test collection. We used these fields, which contain comparatively fewer terms, because ordinary users generally use about two terms for a single query [3]. Each query was subjected to morphological analysis by Chasen, and we used nouns, verbs, adjectives, and unknown terms as query terms.

4.2 Lower Bound of Feature Term in the Category

In this experiment, categories that have less than 10,000 feature terms are eliminated. Essentially, it is better not to eliminate these categories because information in these categories is lost. However, these categories have a possibility of causing bad influence because these categories have insufficient statistical information. Then, it is necessary to consider which influence is more serious.

Table 2 shows the number of 4-lv categories that have more feature terms than each under bounce of feature terms. We conducted retrieval experiment mentioned in 4.1 when the under bound of feature terms are 3,000, 5,000 and 10,000 terms. Its result is shown in table 3 with non-interpolated average precision. The case of 10,000 terms marks 0.0361 and it is the best average of all. This result indicates that it is more important to decide under bounce of feature term in order to acquire sufficient statistical information from categories than to keep the number of categories. Therefore, the under bound of the feature term is set for 10,000 terms in after experiments.

[1] http://www.excite.co.jp/world/
[2] http://research.nii.ac.jp/ntcir/ntcir-ws3/work-en.html
[3] http://chasen-legacy.sourceforge.jp/
[4] http://www2.nict.go.jp/r/r312/EDR/index.html

Table 3: Average precision for the under bounce of feature terms in each 4 level categories.

feature term	3,000	5,000	10,000
average precision	0.0301	0.0278	0.0361

Table 4: Average precision about each query.

query number	1-lv	2-lv	3-lv	4-lv	base line	Web trans
2	0.0971	0.0870	0.1042	0.1048	0.0270	0.1195
13	0.0222	0.0222	0.0102	0.0222	0.0067	0.0070
20	0.1627	0.2048	0.2390	0.2390	0.1313	0.2321
21	0.0245	0.0245	0.0281	0.0002	0.0189	0.0495
23	0.1962	0.2059	0.2059	0.0001	0.0000	0.0003
39	0.0121	0.0141	0.0040	0.0040	0.0035	0.0000
50	0.0151	0.0212	0.0151	0.0151	0.0137	0.0011
ave	0.0400	0.0429	0.0462	0.0361	0.0203	0.0377

4.3 Result of Experiments

Table 4 shows result of the experiments mentioned in 4.1. This result shows non-interpolated average precision about each query in the case of 1-lv, 2-lv, 3-lv, 4-lv, baseline and web translation.

In the case of 1-lv, the system used 13 categories linked from the top page of Yahoo! category. In the case of 2-lv, the system used child categories of 1-lv. In the case of 3-lv, the system used child categories of 2-lv. 4-lv is child categories of 3-lv. In each case, each category in top three levels contains all sub categories.

Table 5 shows results of T-test between proposed method and baseline. We tested if there are significant difference between each three levels and baseline. We assumed no difference between each three levels and baseline, and tested by two-tailed paired T-test.

Besides, table 6 shows translation list about each query that has difference in average precision among three levels.

4.4 Discussion

4.4.1 Effectiveness of Proposed Method

In average of all queries, the average precision of all our proposed method exceed the average precision of baseline.

Table 5: Probability value of T-test between Proposed Method and Baseline.

merged level	1-lv	2-lv	3-lv	4-lv
probability	0.0480	0.0462	0.0296	0.0703

Table 6: Translation list about each query.

query number	lv	translations
2	1-lv	WTO subscription affiliation entry admission joint business cooperation
	2-lv	WTO joining subscription affiliation entry admission joint business cooperation
	3-lv	WTO subscription entry adherence
20	1-lv	Nissan Renault funds capital fund investment money joint business cooperation
	2-lv	Nissan Renault capital fund investment money cooperation
	3-lv	Nissan Renault capital fund investment money cooperation
50	1-lv	fashion mode style
	2-lv	fashionable clothes vogue fashion mode style
	3-lv	fashion mode style

This result verified that our proposed method is effective for Cross-Language Information Retrieval. Table 5 shows probabilities all of three levels are below 0.05. This means that assumption of non-difference between each three levels and baseline is rejected, and there are significant difference. The probability of 4-lv is not below 0.05, but below 0.10. If this probability tested with level of significance as 0.10, assumption of non-difference between 4-lv and baseline is also rejected, and there is significant difference. These results also verified effectiveness of our proposed method.

4.4.2 Translation for Using Category Level

In 2-lv, there are 16 queries that changed its average precision comparing with 1-lv queries as table 4 indicates. 11 queries improved, 5 queries got worse. In these queries, some increase in the number of its translations, others decrease. The query no. 50 is one of increasing case. In this query, translation of the Japanese term “fasshon (fashion)” increase two translations (fashion → fashion, fashinable closes, vogue). In increasing case, queries tend to acquire derivations and synonyms. On the other hand, the query no. 20 is one of decreasing case. In this query, translation of the Japanese term “teikei (cooperation)” decrease one translation (joint business, cooperation → cooperation). This tendency indicates that restricting to target fields of the query is effective to acquire suited translation of the query.

In 3-lv, there are 16 queries that changed its average pre-

Table 7: Probability value of T-test among Proposed Method.

merged level	1, 2-lv	2, 3-lv	3, 1-lv	3, 4-lv
probability	0.2987	0.4103	0.1069	0.2156

cision comparing with 2-lv queries as table 4 indicates. 14 queries improved, 11 queries got worse. In the query no. 2, 2-lv has six translations for the Japanese term “kanyu (adherence)”, which are “joining”, “subscription”, “affiliation”, “entry”, “admission” and “joint business cooperation”. On the other hand, 3-lv has only two translations (“subscription”, “adherence”). These two translations are proper terms in diplomacy field. This result shows that restricting to narrower fields is more effective to acquire suited translation of the query.

However, excessive restriction also has risk of causing a bad influence. In the query no. 50, translations of 3-lv are decreased two terms from the translations of 2-lv. These terms improved average precision comparing with 1-lv. This result indicates that if the specified fields of the query are too narrow, there is a possibility of omitting important translations from the field.

4.4.3 Appropriate Level of Using Category

Table 4 indicates that the average precision increases by using lower level categories in 3-lv or above levels. Significant differences between each level are tested by T-test in Table 7. There are no differences between neighbor level categories. However, probability value there is significant difference between 1-lv and 3-lv. In the case of 4-lv, average precision become worse than upper levels.

These results indicate that precision increases by using lower level category, however using too lower level category causes decrease of precision. Lower level categories can restrict target fields. Then these categories are effective in narrowing proper translations. There are two factors about this bad influence.

First, excessive restriction causes probability of failing to acquire translations. Excessive restriction increases the probability that the restricted field suits for some query term, but does not suit for other query terms. This case increases the probability that some query terms cannot acquire proper translation.

Second, selected category is omitted from proper fields by excessive restriction. There are cases that unsuitable subcategory is selected by restriction even if its upper category suits proper field. In this case, it is difficult to acquire proper translations.

In conclusion from above discussion, restricting the target fields of the query is effective to acquire suited translation of the query. However, excessive restriction causes decline in retrieval effectiveness. Thus, it is needed to find appropriate level of the merged category in Web directory in order to resolve translation disambiguity. However, bad influence becomes more serious than effect of restriction in too lower categories.

5 Conclusion

In this paper, we proposed a query disambiguation method for Cross-Language Information Retrieval using Web directories. In addition, we conducted experiments of retrieval using NTCIR3 test collection and verified that the proposed method is effective for Cross-Language Information Retrieval. We found that it is effective to restrict to target fields of the query using lower level merged categories in order to acquire suited translation of the query. However, excessive restriction has possibility of causing decline in retrieval effectiveness. In this experiment, 3-lv marked the best precision of four levels. This means that 3-lv most balances merit of restriction with demerit of excessive restriction.

In future work, we need to detect most suited level of using merged categories in order to acquire more proper translations of query term. Besides, we consider to use Yahoo! category as linguistic resource. There is possibility of improving to retrieval precision. However, lower category has difficulty of category matching among different language category. Lower category also has a problem that it has insufficient Web documents. Thus, we have to consider using suitable linguistic resource for Yahoo! category (e.g. Wikipedia).

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*, 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, 2000.

[4] F. Kimura, A. Maeda, M. Yoshikawa, and S. Uemura. Cross-Language Information Retrieval using Web Directories. *Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM ’03)*, pages 911–914, 2003.

[5] 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, 1999.

[6] H.-C. Seo, S.-B. Kim, H.-C. Rim, and S.-H. Myaeng. Improving Query Translation in English-Korean Cross-Language Information Retrieval. *Information Processing and Management*, 41(3):507–522, May 2005.