Proceedings of the 5th WSEAS International Conference on Telecommunications and Informatics, Istanbul, Turkey, May 27-29, 2006 (pp446-451)

Using Ontological Chain to Resolve the Translation Ambiguity of Cross-Language Information Retrieval

Pei-Cheng Cheng[1,4], Been-Chian Chien [2], Hao-Ren Ke[3], and Wei-Pang Yang[1,5]
[1] Department of Computer Science, National Chiao Tung University,
1001 Ta Hsueh Rd., Hsinchu, Taiwan 30050, R.O.C.
[2]Department of Computer Science and Information Engineering,
National University of Tainan,
33, Sec. 2, Su Line St., Tainan, Taiwan 70005, R.O.C.

[3] Library and Institute of Information Management, National Chiao Tung University,
1001 Ta Hsueh Rd., Hsinchu, Taiwan 30050, R.O.C.
claven@lib.nctu.edu.tw
[4] Department of Information Management, Ching Yun University
229, Chien-Hsin Rd.,Jung-Li. Taiwan 320. R.O.C.

[5] Department of Information Management, National Dong Hwa University
1, Sec. 2, Da Hsueh Rd., Shou-Feng, Hualien, Taiwan 97401, R.O.C.

Abstract: - In this paper we proposed an ontology-based approach for reducing the ambiguity of bilingual translation. A metric based on the ontological chain for co-occurrence concept evaluation has also been proposed. Based on the ontology chain we can calculate the best translation of all possible translated terms. The experimental results show that the proposed approach can effectively remove the irrelevant results and improve the precision.

Key-words: - Cross Language Information Retrieval; Ontology; Translation Ambiguity;

1. Introduction

A large amount of information materials in various languages is available through the Internet. Thus, the issues of multilingualism arise in order to overcome the remaining technical barriers that still separate countries and cultures. Recently, a lot of digital libraries that contain large collections of information in a various languages have been established. However, it is impractical to submit a query in each language to retrieve these multilingual documents. Therefore, a multilingual information retrieval environment is essential for worldwide information resources. Most research and development activities are focused on only one language. But, this is not the objective of the digital libraries and Internet environment. Both technologies aim at establishing a global digital library containing all information resources from different areas, different countries and in different languages. The access to those materials should be possible for the worldwide community and never be restricted because of non-uniform languages.

Internet contains various types of information, video, image, voice, and text. Image retrieval is a good application for bilingual language research. People looking for an image usually do not care in which language annotations of images were written. For example, a user wants to find sunset images and these images may be annotated by different language on the Internet. A cross-language information retrieval system is especially useful to retrieve image data.

The EU funded some projects addressing the multilingual issues. For instance, in the ESPRIT project EMIR (European Multilingual Information Retrieval) a commercial information retrieval system SPIRIT has been developed that supports French, English, German, Dutch and Russian. UNESCO launched some projects in order to democratize and globalize the access to the world cultural heritage,

such as Memory of the World. Recently, UNESCO has started the MEDLIB project, which aims at creating a virtual library for the Mediterranean region.

Cross-language Information retrieval allows a user to submit a query different from the language used to write document. In 1996, ACM SIGIR Workshop for Multilingual Information Retrieval names it as Cross-Language Information Retrieval. Many international conferences focus on cross-language information retrieval issues, like as ACL Annual Meeting, ACM SIGIR99 (SIGIR00), etc. TREC, CLEF, and NTCIR are three major organizations that foster the evaluation of CLIR technologies.

In this paper, we develop a cross-language information retrieval (CLIR) system that accepts Chinese queries to retrieve English documents. While translating query language to another language will cause the ambiguous problem, a Chinese term will translate into many meaning of English. We propose an ontology-based approach to overcoming the problem of translation ambiguity. Based on the domain knowledge ontology in the ontology, we can compute the possibility of meaningful combination of query terms. The rest of this paper is organized as follows. In Section 2, we review some related work on cross-language retrieval systems. In Section 3, the ontology-based approach is proposed for solving the problem of translation ambiguity problem. We conduct an experiment to evaluate the benefit of the ontology approach in Section 4. The summary and further works are discussed in section 5.

2. Related work

Cross-language information retrieval allows a user to formulate a query in one language and retrieve documents in others. The controlled vocabulary dictionary is the first and traditional technique widely used in libraries and documentation centers. Documents are indexed manually using fixed terms which are also used for queries. These terms can be also indexed in multiple languages and maintained in a thesaurus.

Using dictionary-based techniques, queries will be translated into the target language in which a document may be found. The dictionary-based method is easy to implement; however, because the vocabulary is too general, the problem of ambiguity will be caused. The corpus-based technique analyzes large collections of existing texts and automatically extracts the information needed on which the translation will be based. However, this technique tends to require the integration of linguistic constraints, because the use of only statistical techniques by extracting information can introduce errors and achieve bad performance [1]. The corpus-based approach can assist the dictionary-based approach in application. Latent Semantic Indexing (LSI) is a new approach and a new experiment in multilingual information retrieval which allows a user to retrieve documents by concepts and meanings and not only by pattern matching. The drawback of Latent Semantic Indexing is that it is more complex in computer time and the additional documents is hard to direct appending.

The size of public domain and commercial dictionaries in multiple languages on the Internet is increasing steadily. Electronic monolingual and bilingual dictionaries build a solid platform for developing multilingual applications. Using dictionary-based technique queries will be translated into another language in which a document may be found. However, this technique sometimes achieves unsatisfactory results because of ambiguities. Many words do not have only one translation and the alternate translations have very different meanings. The dictionary based approach cannot decide the priority of translated terms. Moreover, the scope of a dictionary is limited. It lacks in particular a technical and topical terminology which is very crucial for a correct translation. Nevertheless, this technique can be used for implementing simple dictionary-based application or can be combined with other approaches.

Using the dictionary-based approach for query translation has achieved an effectiveness of 40-60% in comparison with monolingual retrieval [2], [3]. In [4] a multilingual search engine, called TITAN, has been developed. Based on a bilingual dictionary it allows to translate queries from Japanese to English and English to Japanese. TITAN helps Japanese users to search in the Web using their own language. However, this system suffers again from polysemy.

The main problems associated with the dictionary-based CLIR are (1) untranslatable search terms due to the limitations of general dictionaries, (2) the processing of inflected words, (3) phrase identification and translation, and (4) lexical ambiguity in source and target languages. The categories of untranslatable terms involve new compound words, special terms, and cross-lingual spelling variants, i.e., equivalent words in different

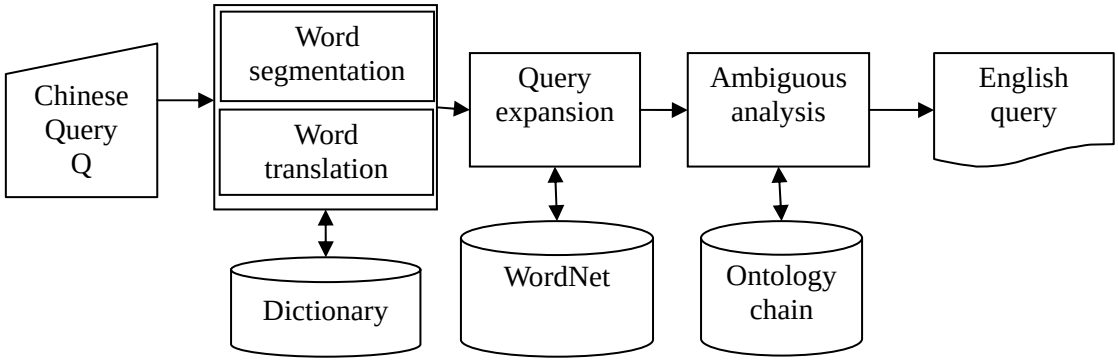

Figure 1: Proposed query translation flowchart

languages which differ slightly in spelling, particularly proper names and loanwords. *Translation ambiguity* refers to the increase of irrelevant word sentences in translation due to lexical ambiguity in the source and target languages. We proposed an ontology approach to reduce the translation ambiguity.

In this paper, we discuss the ambiguity problem of the dictionary-based approach. The translated words need expanding to improve the recall. The expanding process also will cause the ambiguity problem and reduce the precision. The meaning of a word is related to the context. The co-occurrence terms will help to distinguish the sense of a word. We propose an ontology approach to reduce the ambiguous translation. Based on the ontology we can inference the most possible combination of translated words. Each combination of translated words has a probability, and we can eliminate un-meaningful translated words.

3. Proposed Ontology based Cross language retrieval approach

The overall interactive search process is shown in Fig. 1. Given an initial query, Q, in which Q denotes a Chinese text query, the system performs the cross-language retrieval, and returns a set of relevant documents to the user. The Chinese query sentence first was decomposed in to significant phases from left to right, and every significant phase translates into candidate English terms by looking up the dictionary. In the second step, we expand the candidate terms to get more words of similar concept by WordNet. At the Ambiguous analysis process, we calculate the possibility of which combination of candidate English words will be the translation of Chinese query. After translate into English query terms, we use the vector space model to retrieve the related documents.

We create a query representation by expressing a query as a vector in the vector space model [8]. In our proposed approach, a textual vector representation is modeled in terms information. The Textual Vector Representation is defined as follows. Let W ($|W| = n$) be the set of significant keywords in the corpus, for an document D, its textual vector representation (i.e., D_T) is defined as Eq. (1),

$$D_T =< w_{t_1}(D_T),...,w_{t_n}(D_T) > \quad (1)$$

where the first n dimensions indicate the weighting of a keyword t_i in D_T, which is measured by TF-IDF [8], as computed in Eq. (2);

$$w_{t_i}(P_T) = \frac{tf_{t_i,P_T}}{\max tf} \times \log \frac{N}{n_{t_i}} \quad (2)$$

In Eq. (2), $\frac{tf_{t_i,P_T}}{\max tf}$ stands for the normalized frequency of t_i in P_T, maxtf is the maximum number of occurrences of any keyword in D_T, N indicates the number of documents in the corpus, and n_{t_i} denotes the number of documents in which caption t_i appears.

In the above, we introduce how to create a textual vector representation for D_T. As for a query Q, one problem is that since Q_T is given in Chinese, it is necessary to translate Q_T into English, which is the language used in the document collection. We first perform the word segmentation process to obtain a set of Chinese words. For each Chinese word, it is then translated into one or several corresponding English words by looking it up in a dictionary. The dictionary that we use is pyDict[1]. Up to now, it is hard to determine which of these English translations are correct. We tend to keep all English translations in order not to lose the consideration of any correct words.

[1] An English/Chinese dictionary written by D. Gau, which is available at http://sourceforge.net/projects/pydict/.

Another problem is the so-called short query problem. A short query usually cannot cover as many useful search terms as possible because of the lack of sufficient words. We address this problem by performing the query expansion process to add new terms to the original query. The additional search terms is taken from a thesaurus – WordNet [9]. For each English translation, we include its synonyms, hypernyms, and hyponyms into the query.

Now, it comes out a new problem. Assume $Expansion(Q_T) = \{e_1,...,e_h\}$ is the set of all English words obtained after query translation and query expansion, it is obvious that $Expansion(Q_T)$ may contain a lot of words which are not correct translations or useful search terms. To resolve the translation ambiguity problem, we define *word co-occurrence relationships* by an *ontology chain* to determine final query terms. If the co-occurrence frequency of e_i and e_j in the corpus is greater than a predefined threshold, both e_i and e_j are regarded as useful search terms for document retrieval. So far, we have a set of search terms, $Disambiguity(Q_T)$, which is presented as Eq. (3),

$$Disambiguity(Q_T) = \{e_i, e_j \mid e_i, e_j \in Expansion(Q_T) \\ \& \, e_i, e_j \text{ have a significant cooccurrence}\} \quad (3)$$

After giving the definition of $Disambiguity(Q_T)$, for a query Q, its textual vector representation (i.e., Q_T) is defined in Eq. (4),

$$Q_T = < w_{t_1}(Q_T),...,w_{t_n}(Q_T) > \quad (4)$$

where $w_{t_i}(Q_T)$ is the weighting of a keyword t_i in Q_T, which is measured as Eq. (5).

In Eq. (5), W is the set of significant keywords as defined before, $\frac{tf_{t_i,Q_T}}{\max tf}$ stands for the normalized frequency of t_i in $Disambiguity(Q_T)$, maxtf is the maximum number of occurrences of any keyword in $Disambiguity(Q_T)$, N indicates the number of images in the corpus, and n_{t_i} denotes the number of images in whose caption t_i appears.

$$w_{t_i}(Q_T) = \begin{cases} \frac{tf_{t_i,Q_T}}{\max tf} \times \log \frac{N}{n_{t_i}} \end{cases} \quad (5)$$

Ontology is a formal explicit specification. It constructs a common conceptualization for users. Ontology contains principal concepts and relationships among concepts, especially useful in any specific domain. Figure 2 is an example of ontology. In our work, we use the St Andrews data set from ImageClef2004 to evaluate our proposed methods. The St Andrews dataset consists of 28,133 photographs from St Andrews University Library photographic collection, which holds one of the largest and most important collections of historic photography in Scotland.

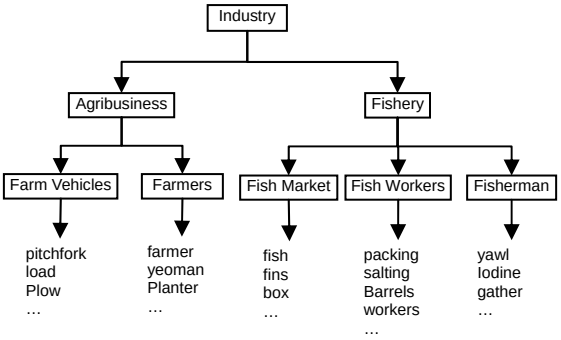

Figure 2: An example of ontology.

The St Andrews data set from CLEF2004, each document belong to more than one category. It contains 946 categories. We arrange the categories to construct an ontology map that define the relationship between each category. Creating the ontology by category is easier than construct ontology by terms in time complexity.

If the translations of query terms contain multiple meaning, it will generate ambiguity result. We use the ontological chain to evaluate the most appropriate concepts. The related similarity between an ontology node L_i and a query express Q is defined as Eq. (6).

$$Sim(L_i, Q) = \sqrt{\frac{\sum_{j=1}^{N} t_{ij}^{\,2}}{N}} \quad (6)$$

Where t_{ij} is the frequencies of translated English terms, N is the number of Chinese query terms, a Chinese query term will expand into several English terms. Q presents one of the translated combinations. Our objective is to estimate the probability of each candidate term's combination. Two terms falling into the same node, they have the same concept and those terms will cause a meaningful concept. Otherwise, if two terms belong to different concept they must be excessive expanded, it will have low probability to be a correct query terms.

In the ontology map connect node may have some degree of relationship. A concept may consist

of several nodes. We let arbitrary two node of ontology have a distance. For example as shown in Figure 2 the "herring" and "fish processing" have a path of distance three. Based on the similarity of query terms and the distance we can define a relationship between two nodes as Eq.(7).

$$\text{Rel}(L_i, L_j) = \frac{Sim(L_i, Q) \times Sim(L_j, Q)}{distance(L_i, L_j)} \qquad (7)$$

Figure 3 is a relationship map between nodes. We set a threshold 1.5 to eliminate less important relationships. The connected nodes will construct many semantic concepts. We can eliminate the translated English term that do no fall in the connected nodes. The sum of relationships of each island connected nodes can be used as important degree for ranking candidate English terms.

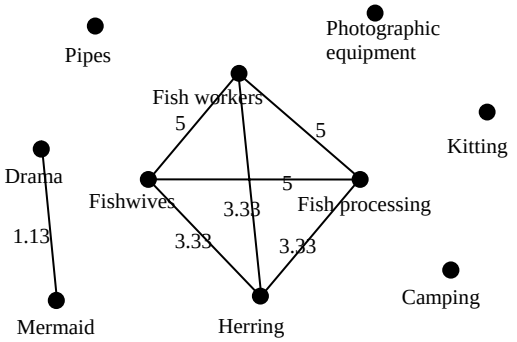

Figure 3: semantic relationship map.

4. Experiment and result

We test our method using the CLEF English-Chinese ImageCLEF test collection. The collection includes about 30,000 historic photographs with British English semi-structured captions. ImageCLEF is the cross-language image retrieval track that is run as part of the Cross Language Evaluation Forum (CLEF) campaign. The campaign is run in a similar manner as the TREC and NTCIR information retrieval evaluations. This is a bilingual query translation task from a source language to English. The goal of this task lies in finding as many relevant images as possible from the St. Andrews image collection.

Each image has an accompanying textual description consisting of 8 distinct fields. These fields can be used individually or collectively to facilitate image retrieval. The 28,133 captions consist of 44,085 terms and 1,348,474 word occurrences; the maximum caption length is 316 words, but on average 48 words in length. All captions are written in British English, although the language also contains colloquial expressions. Approximately 81% of captions contain text in all fields, the rest generally without the description field. In most cases the image description is a grammatical sentence of around 15 words.

Each image in the St Andrews collection has been assigned to one or more descriptive categories, e.g. "airports", "airships", "flowers", "beach scenes" and "breweries". On average, each image is assigned to 4 categories. We use the categories to manually construct the ontology. Related categories will be gathered into a class from bottom to up by experts to construct a hierarchic structure, which is the ontology used in this paper. Totally the ontology has 946 categories. The language used to express the associated texts or textual queries should not affect retrieval, i.e. an image with a caption written in English should be searchable in languages other than English. In the experiment we will show that ontology approach will improve the precision by 13%, which is better than the result without dis-ambiguity process.

The cross language evaluation forum includes data set, 25 query topics and query answer to evaluate related systems. We evaluate three models; they are Mono-lingual IR (mono-lingual Information retrieval), CLIR (cross language information retrieval), and Ontology-based CLIR (ontology-based cross language information retrieval). The Mono-lingual IR uses the original English queries with expansion by WordNet to retrieve related documents; it is the baseline in evaluation. The CLIR model uses Chinese queries to retrieve related documents; it first translates the Chinese terms to all possible English terms by a dictionary. After translation, the CLIR model also expands synonym by WordNet and uses the possible English terms as query without dis-ambiguous analysis. The CLIR model does not consider the ambiguity problem that is used to compare with Ontology-based CLIR model. The Ontology-based CLIR analysis the candidate terms based on Ontology chain trying to resolve the ambiguity translation problem.

The performance is normally evaluated based on precision and recall as defined in Eq.(8)and (9)

$$\text{precision} = \frac{number \text{ of correct answer retrieved}}{\text{total number of retrieved}}, \qquad (8)$$

$$recall = \frac{number \text{ of correct answer retrieved}}{total\ number\ of \text{ actual answer}}. \qquad (9)$$

We used the UMASS and Lemur versions of the standard trec_eval tool to compute the precision, recall and mean average precision scores. This provides the "standard" information retrieval evaluation measures, e.g. precision at a given rank cut-off, average precision across 11 recall points, and single-valued summaries for each measure.

Table 1 is the experiment result, it show that Ontology-based CLIR improve the performance from 49% to 55 %, and reach to the 92% precision of mono lingual IR.

	Mono IR	CLIR	Ontology CLIR
Mean Average Precision (MAP)	60.63%	49.18%	55.18%

Table 1: The Mean Average Precision.

The mono-lingual IR using the original English query avoids the translator ambiguity, so that it has the best performance in the experiment result. The normal CLIR translated Chinese to English may have miss translation or ambiguity translated will cause find more irrelevant documents lowering the precision. The experiment result shows that the Ontology chain is useful to recognize the acceptation of query. The proposed Ontology based CLIR effective to locate the correct translation increasing a 10% of monolingual IR.

5. Conclusion

A novel ontology-based approach for reducing the ambiguity of bilingual translation has been presented in this paper. Using a dictionary to translate a Chinese term into English may cause the problem of translation ambiguity. We successfully find the best translation by analyzing the all possible translated terms. A metric based on the ontology chain for co-occurrence concept evaluation has also been proposed. The experimental results show that the proposed approach can effectivly remove the irrelevant results and improve the precision.

The Ontology is very useful in the synonmy and polysemy analysis. A word in different domain may have different lexical meaning. The ontology is domain dependent. In this paper we define the ontology by manual for St. Andrews corpus; we now try to cluster the words into concept capable of automatic constructing the ontology.

References:

[1] H. Haddouti. Multilinguality Issues in Digital Libraries. *Proceedings of the EuroMed Net'98 Conference Nicosia*, March 3-7, 1998.

[2] L. Ballesteros, W.B. Croft . Dictionary-based methods for cross- lingual information retrieval, *Proc. of the 7th Int. DEXA Conference on Database and Expert Systems Applications*, 1996.

[3] D. A. Hull, G. Grefenstette. Querying across languages. A dictionary-based approach to multilingual information retrieval, *In Proceedings of the 19th ACM SIGIR Conference*, 1996

[4] Y. Hayashi, et al. TITAN: A Cross-linguistic Search Engine for the WWW, *in Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, Stanford, CA, 1997.

[5] P. Sheridan, J. P. Ballerini. Experiments in Multilingual Retrieval Using the Spider System. *In Proceeding of the 19th Annual International ACM SIGIR*, 1996

[6] M. Wechsler, P. Schäuble. Multilingual Information Retrieval Based on Document Alignment Techniques. *Lecture Notes in Computer Science. Ed. Ch. Nikolau, C. Stephanidis. Second European Conference on Research and Advanced Technology for Digital Libraries ECDL´98*, Crete 1998

[7] S. Deerwester, S. T. Dumais, R. Harshman. Indexing by Latent Semantic analysis. *Journal of the American Society for Information Science*, vol. 41, 6, Sept. 1990.

[8] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1983.

[9] G. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM*, 1995, pp. 39-45.

[10] P. C. Cheng, J. Y. Yeh., B. C. Chien, H. R. Ke and W. P. Yang, Comparison and Combination of Textual and Visual Features for Interactive Cross-Language Image Retrieval, *In Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign, Lecture Notes in Computer Science (LNCS)*, Springer, Heidelberg, Germany,2005.