Asian Journal of Information Technology 6 (11): 1187-1191, 2007
ISSN: 1682-3915

Applications of Word Sense Disambiguation in Mixed Language Information Retrieval

[1]M. Sundara Rajan and [2]S.P. Rajagopalan
[1]Department of Computer Science, S.R.M Arts and Science College, 61-2nd Main Road,
Baby Nagar, Velachery, Chennai, Pin-600 042, TN India
[2]School of Computer Science, Engineering and Applications, M. G. R. University, Maduravoil,
Chennai 600095, TN, India

Abstract: We propose a mixed language query disambiguation approach by using co-occurrence information from monolingual data only. A mixed language query consists of words in a primary language and a secondary language. Our method translates the query into monolingual queries in either language. Two novel features for disambiguation, namely contextual word voting and 1-best contextual word, are introduced and compared to a baseline feature, the nearest neighbor. Average query translation accuracy for the 2 features improved considerably compared to the baseline accuracy.

Key words: Word Sense Disambiguation (WSD), mixed language query, voting, weighting, pruning, baseline

INTRODUCTION

Online information retrieval is now prevalent because of the ubiquitous World Wide Web. The Web is also a powerful platform for another application/interactive spoken language query systems. Traditionally, such systems were implemented on stand-alone kiosks. Now we can easily use the Web as a platform. Information such as airline schedules, movie reservation, car trading, etc., can all be included in HTML files, to be accessed by a generic spoken interface to the web browser (Fung *et al*., 1998).

Until recently, most of the search engines handle keyword based queries where the user types in a series of strings without syntactic structure. The choice of key words in this case determines the success rate of the search. In many situations, the key words are ambiguous. To resolve ambiguity, query expansion is usually employed to look for additional keywords. We believe that a more useful search engine should allow the user to input natural language sentences. Sentence-based queries are useful because they are more natural to the user and more importantly, they provide more contextual information which are important for query understanding.

To date, the few sentence-based search engines do not seem to take advantage of context information in the query, but merely extracting key words from the query sentence (AskJeeves, 1998). In addition to the need for better query understanding methods for a large variety of domains, it has also become important to handle queries in different languages. Cross-language information retrieval has emerged as an important area as the amount of non-English material is ever increasing (Ballesteros and Croft, 1998; Picchi and Peters, 1998; Davis, 1998).

One of the important tasks of cross-language IR is to translate queries from one language to another. The original query and the translated query are then used to match documents in both the source and target languages. Target language documents are either glossed or translated by other systems. According to Grefenstette (1998) three main problems of query translations are:

- Generating translation candidates.
- Weighting translation candidates.
- Pruning translation alternatives for document matching.

In cross-language IR, key word disambiguation is even more critical than in monolingual IR (Ballesteros and Croft, 1998) since the wrong translation can lead to a large amount of garbage documents in the target language, in addition to the garbage documents in the source language. Once again, we believe that sentence-based queries provide more information than mere key words in cross-language IR. In both monolingual IR and cross-language IR, the query sentence or key words are assumed to be consistently in one language only. This makes sense in cases where the user is more likely to be

Corresponding Author: M. Sundara Rajan, Professor and Head, PG-Department of Computer Science, S.R.M Arts and Science College, 61-2nd Main Road, Baby Nagar, Velachery, Chennai, Pin-600 042, TN India

a monolingual person who is looking for information in any language. It is also easier to implement a monolingual search engine. However, we suggest that the typical user of a cross-language IR system is likely to be bilingual to some extent.

Most web users in the world know some English. In fact, since English still constitutes 88% of the current web pages, speakers of another language would like to find English contents as well as contents in their own language. Likewise, English speakers might want to find information in another language. A typical example is a local user looking for the information of an American movie, s/he might not know the local name of that movie. His/her query for this movie is likely to be in mixed language.

Mixed language query is also prevalent in spoken language. We have observed this to be a common phenomenon in every country. In general, a mixed language consists of a sentence mostly in the primary language with some words in the secondary language. We are interested in translating such mixed language queries into monolingual queries unambiguously. In this study, we propose a mixed language query disambiguation approach which makes use of the co-occurrence information of words between those in the primary language and those in the secondary language.

MATERIALS AND METHODS

Mixed language query translation is halfway between query translation and query disambiguation in that not all words in the query need to be translated. There are two ways to use the disambiguated mixed language queries. In one scenario, all secondary language words are translated unambiguously into the primary language, and the resulting monolingual query is processed by a general IR system. In another scenario, the primary language words are converted into secondary language and the query is passed to another IR system in the secondary language. Our methods allows for both general and cross-language IR from a mixed language query. To draw a parallel to the three problems of query translation, we suggest that the three main problems of mixed language disambiguation are:

- Generating translation candidates in the primary language.
- Weighting translation candidates.
- Pruning translation alternatives for query translation.

Co-occurrence information between neighboring words and words in the same sentence has been used in phrase extraction (Fung and Wu, 1994) phrasal translation (Smadja *et al*., 1996), target word selection (Liu and Li, 1997), domain word translation (Fung and Lo, 1998), sense disambiguation (Yarowsky, 1995) and even recently for query translation in cross-language IR as well (Ballesteros and Croft, 1998). Co-occurrence statistics is collected from either bilingual parallel and non-parallel corpora (Fung and Lo, 1998) or monolingual corpora (Liu and Li, 1997; Yarowsky, 1995). As we noted in Fung and Lo (1998) and Fung *et al*. (1998) parallel corpora are rare in most domains. We want to devise a method that uses only monolingual data in the primary language to train co-occurrence information.

Translation candidate generation: Without loss of generality, we suppose the mixed language sentence consists of the words $S = \{E_1, E_2, \ldots C, \ldots, E_n\}$, where C is the only secondary language word*. Since in our method we want to find the co-occurrence information between all E_i and C from a monolingual corpus, we need to translate the latter into the primary language word E_c. This corresponds to the first problem in query translation/translation candidate generation. We generate translation candidates of C via an online bilingual dictionary. All translations of secondary language word C, comprising of multiple senses, are taken together as a set $\{E_a\}$.

*In actual experiments, each sentence can contain multiple secondary language words.

Translation candidate weighting: Problem two in query translation is to weight all translation candidates for C. In our method, the weights are based on co-occurrence information. The hypothesis is that the correct translations of C should co-occur frequently with the contextual words E_i and incorrect translation of C should co-occur rarely with the contextual words. Obviously, other information such as syntactical relationship between words or the part-of-speech tags could be used as weights too. However, it is difficult to parse and tag a mixed language sentence. The only information we can use to disambiguate C is the co-occurrence information between its translation candidates $\{E_a\}$ and $E_1, E_2, \ldots, E_n$.

Translation candidate pruning: The last problem in query translation is selecting the target translation. In our approach, we need to choose a particular E_c from E_{ci}. We call this pruning process translation disambiguation. We present and compare three unsupervised statistical methods in this study. The first base-line method is similar to Ballesteros and Croft (1998) and Smadja *et al*. (1996),

where we use the nearest neighboring word of the secondary language word *C* as feature for disambiguation. In the second method, we choose all contextual words as disambiguating feature. In the third method, the most discriminative contextual word is selected as feature.

Baseline: Single neighbouring word as disambiguating feature: The first disambiguating feature we present here is similar to the statistical feature in Smadja *et al.* (1996) and Ballesteros and Croft (1998) namely the co-occurrence with neighboring words. We do not use any syntactic relationship as in Dagan and Itai (1994) because such relationship is not available for mixed-language sentences. The assumption here is that the most powerful word for disambiguating a word is the one next to it.

Based on mutual information, the primary language target word for C is chosen from the set $\{E_{ci}\}$. E_y is taken to be either the left or the right neighbor of our target word.

Voting: Multiple contextual words as disambiguating feature: The baseline method uses only the neighboring word to disambiguate C. The intuition for choosing the nearest neighboring word E_y as the disambiguating feature for C is based on the assumption that they are part of a phrase or collocation term and that there is only one sense per collocation (Dagan and Itai, 1994). However, in most cases where C is a single word, there might be some other words which are more useful for disambiguating C. In fact, such long-distance dependency occurs frequently in natural language (Rosenfeld, 1995). Another reason against using single neighboring word comes from (Gale and Church, 1994) where it is argued that as many as 100,000 context words might be needed to have high disambiguation accuracy. Yarowsky (1995) all use multiple context words as discriminating features. We have also demonstrated in our domain translation task that multiple context words are useful (Fung and Lo, 1998; Fung and McKeown, 1997). Based on the above arguments, we enlarge the disambiguation window to be the entire sentence instead of only one word to the left or right. We use all the contextual words in the query sentence.

Suppose there are n primary language words in $S = E_1, E_2, ..., C, ... E_n$, as shown in Fig. 1, we compute mutual information scores between all E_{ci} and all E_j where E_{ci} is one of the translation candidates for C and E_j is one of all n words in S. A mutual information score matrix is shown in Table 1 where MI_{jci} is the mutual information score between contextual word E_j and translation candidate E_{ci}. For each row j the largest scoring MI_{jci} receives a vote. The rest of the row get zero's. At the end

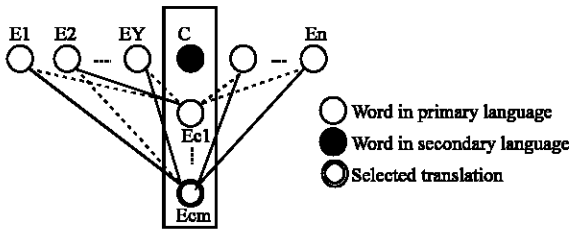

Fig. 1: Voting for the best translation

Table 1: Mutual information between all translation candidates and words

	E_{C1}	E_{C2}	...	E_{Cm}
E_1	$MI1_{c1}$	$MI1_{c2}$	...	$MI1_{c,}$
E_2	$MI2_{c2}$	$MI2_{c2}$	...	$MI2_{cm}$
...				
E_j	$MI1_{jc2}$	MI_{jc2}	...	MI_{jcm}
...				
E_n	Mi_{nc1}	MI_{nc2}	...	MI_{ncm}

we sum up all the one's in each column. The column i receiving the highest vote is chosen as the one representing the real translation.

Best contextual word as disambiguating feature: In the above voting scheme, a candidate receives either a one vote or a zero vote from all contextual words equally no matter how these words are related to C. As an example, in the query "Please show me the latest movie of Jacky Chan", the and Jacky are considered to be equally important. We believe however, that if the most powerful word is chosen for disambiguation, we can expect better performance. This is related to the concept of "trigger pairs" in Rosenfeld (1995) and Singular Value Decomposition in Shutze (1992).

In Dagan and Itai (1994) syntactic relationship is used to find the most powerful "trigger word". Since syntactic relationship is unavailable in a mixed language sentence, we have to use other type of information. In this method, we want to choose the best trigger word among all contextual words.

We compute the disambiguation contribution ratio for each context word E_j. For each row j in Table 1, the largest MI score MI_{jcf} and the second largest MI score MI_{jcs} are chosen to yield the contribution for word E_j, which is the ratio between the two scores.

$$\text{Contribution}(E_j, E_{ci}) = \frac{MI_{jcf}}{MI_{jcs}}$$

If the ratio between MI_{jcf} and MI_{jcs} is close to one, we reason that E_j is not discriminative enough as a feature for disambiguating C. On the other hand, if the ratio between MI_{iefi} and MI_{iesi} is noticeably greater than one, we can use

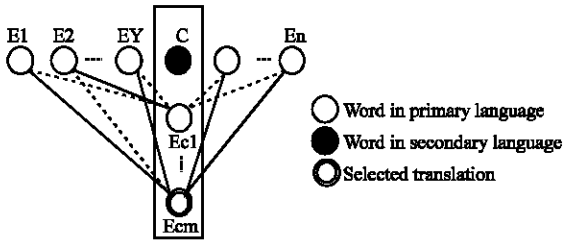

Fig. 2: The best contextual word as disambiguating feature

E_y as the feature to disambiguate $\{E_{ci}\}$ with high confidence. We choose the word E_y with maximum contribution as the disambiguating feature and select the target word E_{ci}, whose mutual information score with E_y is the highest, as the translation for C.

$$r = \arg\max_i MI(E_y, E_{ci})$$

This method is illustrated in Fig. 2. Since E_2 is the contextual word with highest contribution score, the candidate E_i is chosen so that the mutual information between E_2 and E_{ci} is the largest.

Evaluation experiments: The mutual information between co-occurring words and its contribution weight is obtained from a monolingual training corpus. We evaluate our methods for mixed language query disambiguation on an automatically generated mixed-language test set. No bilingual corpus, parallel or comparable, is needed for training. To evaluate our method, a mixed-language sentence set is generated from the monolingual corpus.

Some English words in the original sentences are selected randomly and translated into secondary language words manually to produce the testing data. These are the mixed language sentences. The ratio of secondary language words in the sentences varies from 10-65%.

We carry out three sets of experiments using the three different features we have presented in this study. In each experiment, the percentage of primary language words in the sentence is incrementally increased at 5% steps, from 35-90%. We note the accuracy of unambiguous translation at each step.

Evaluation results: One advantage of using the artificially generated mixed-language test set is that it becomes very easy to evaluate the performance of the disambiguation/ translation algorithm. The experimental results are shown in Fig. 3. The horizontal axis represents the percentage of English words in the testing data and the vertical axis represents the translation accuracy. Translation accuracy is the ratio of the number of secondary language

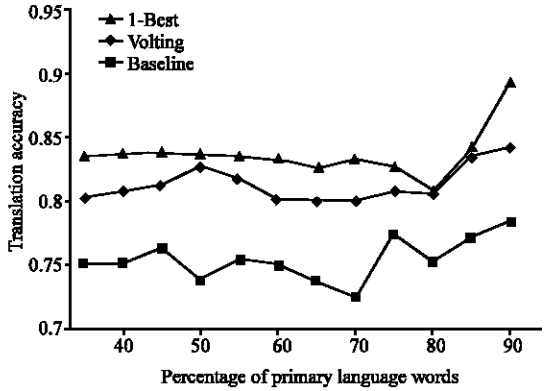

Fig. 3: 1-best is the most discriminating feature

words disambiguated correctly over the number of all secondary language words present in the testing sentences. The three different curves represent the accuracies obtained from the base-line feature, the voting model and the 1-best model.

We can see that both voting contextual words and the 1-best contextual words are more powerful discriminant than the baseline neighboring word. The 1-best feature is most effective for disambiguating secondary language words in a mixed-language sentence.

CONCLUSION

Mixed-language query occurs very often in both spoken and written form, especially in Asia. Such queries are usually in complete sentences instead of concatenated word strings because they are closer to the spoken language and more natural for user. A mixed-language sentence consists of words mostly in a primary language and some in a secondary language. However, even though mixed-languages are in sentence form, they are difficult to parse and tag because those secondary language words introduce an ambiguity factor. To understand a query can mean finding the matched document, in the case of web search, or finding the corresponding semantic classes, in the case of an interactive system. In order to understand a mixed-language query, we need to translate the secondary language words into primary language unambiguously.

In this study, we present an approach of mixed-language query disambiguation by using co-occurrence information obtained from a monolingual corpus. Two new types of disambiguation features are introduced, namely voting contextual words and 1-best contextual word. These two features are compared to the baseline feature of a single neighboring word.

The baseline method uses only the neighboring word to disambiguate C. The assumption is that the

neighboring word is the most semantic relevant. This method leaves out an important feature of nature language: long distance dependency. Experimental results show that it is not sufficient to use only the nearest neighboring word for disambiguation. The performance of the voting method is better than the baseline because more contextual words are used. The results are consistent with the idea in Gale and Church (1994) and Yarowsky (1995). In our experiments, it was found that the 1-best contextual word is even better than multiple contextual words. This seemingly counter-intuitive result leads us to believe that choosing the most discriminative single word is even more powerful than using multiple contextual words equally. We believe that this is consistent with the idea of using "trigger pairs" in Rosenfeld (1995) and Singular Value Decomposition in Shutze (1992). We can conclude that sometimes long-distance contextual words are more discriminating than immediate neighboring words, and that multiple contextual words can contribute to better disambiguation. Our method using multiple disambiguating contextual words can take advantage of syntactic information even when parsing or tagging is not possible, such as in the case of mixed-language queries. Other advantages of our approach include:

- The training is unsupervised and no domain-dependent data is necessary.
- Neither bilingual corpora or mixed-language corpora is needed for training.
- It can generate monolingual queries in both primary and secondary languages, enabling true cross-language IR.

In our future research, we plan to analyze the various discriminating words contained in a mixed language or monolingual query to find out which class of words contribute more to the final disambiguation. We would also like to test the significance of the co-occurrence information of all contextual words between themselves in the disambiguation task. Finally, we plan to develop a general mixed-language and cross-language understanding framework for both document retrieval and interactive tasks.

REFERENCES

AskJeeves, 1998. http://www.askjeeves.com.

Davis, M., 1998. Free resources and advanced alignment for cross-language text retrieval. In Proceedings of the 6th Text Retrieval Conference (TREC-6), NIST, Gaithersburg, MD.

Eugenio Picchi and Carol Peters, 1998. Cross-Language Information Retrieval: A System for Comparable Corpus Querying. In Gregory Grefenstette, (Ed.). Cross-Language Information Retrieval, Kluwer Academic Publishers, pp: 81-92.

Frank Smadja, Kathleen McKeown and Vasileios Hatzsivassiloglou, 1996. Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics, 21: 1-38.

Gregory Grefenstette, 1998. Cross-language Information Retrieval. Kluwer Academic Publishers.

Hinrich Shutze, 1992. Dimensions of meaning. In Proceedings of Supercomputing.

Ido Dagan and Alon Itai, 1994. Word sense disambiguation using a second language monolingual corpus. In Computational Linguistics, pp: 564-596.

Lisa Ballesteros and W. Bruce Croft, 1998. Resolving ambiguity for cross-language retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp: 64-71.

Pascale Fung, Cheung Chi Shuen, Lam Kwok Leung, Liu Wai Kat and Lo Yuen Yee, 1998. A speech assisted online search agent (salsa). In ICSLP.

Pascale Fung and Dekai Wu, 1994. Statistical augmentation of a Chinese machine-readable dictionary. In: Proceedings of the 2nd Annual Workshop on Very Large Corpora, Kyoto, Japan, pp: 69-85.

Pascale Fung and Kathleen McKeown, 1997. Finding terminology translations from non-parallel corpora. In The 5th Annual Workshop on Very Large Corpora, Hong Kong, pp: 192-202.

Pascale Fung and Yuen Yee Lo, 1998. An IR approach for translating new words from non-parallel, comparable texts. In Proc. 36th Ann. Conf. Assoc. Computational Linguistics, Montreal, Canada, pp: 414-420.

Rony Rosenfeld, 1995. A Corpus-Based Approach to Language Learning. Ph.D. thesis, Carnegie Mellon University.

William A. Gale and Kenneth W. Church, 1994. Discrimination decisions in 100,000 dimensional spaces. Current Issues in Computational Linguistics: In honour of Don Walker, pp: 429-550.

Xiaohu Liu and Sheng Li, 1997. Statistic-based target word selection in English-Chinese machine translation. Journal of Harbin Institute of Technology.

Yarowsky, D., 1995. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Conference of the Association for Computational Linguistics. Assoc. Computational Linguistics, pp: 189-196.