EuroWordNet as a Resource for Cross-language Information Retrieval

Mark Stevenson[1] and Paul Clough[2]

(1) Department of Computer Science
(2) Department of Information Studies
University of Sheffield
Regent Court, 211 Portobello Street,
Sheffield S1 4DP
United Kingdom
marks@dcs.shef.ac.uk, p.d.clough@sheffield.ac.uk

Abstract

One of the aims of EuroWordNet (EWN) was to provide a resource for Cross-Language Information Retrieval (CLIR). In this paper we present experiments to test the usefulness of EWN for this purpose via a formal evaluation using the Spanish queries from the TREC6 CLIR test set. All CLIR systems using bilingual dictionaries must find a way of dealing with multiple translations and we employ a word sense disambiguation algorithm for this purpose. Retrieval performance using when the disambiguation algorithm was used was 90% of that recorded using queries which had been disambiguated manually.

Introduction

Cross-language information retrieval (CLIR) is the process of providing queries in one language and returning documents relevant to that query which are written in a different language. This is useful in cases when the user has enough knowledge of the language in which the documents are returned to understand them but does not possess the linguistic skill to formulate useful queries in that language.

A popular approach to CLIR is to translate the query into the language of the documents being retrieved. Methods involving the use of machine translation, parallel corpora and machine readable bilingual dictionaries have all been tested, each with varying degrees of success (Ballesteros et. al., 1998; Jang et. al., 1999). One of the simplest and most effective methods for query translation is to perform dictionary lookup based on a bilingual dictionary. However, the mapping between words in different languages is not one-to-one, for example the English word "bank" is translated to French as "banque" when it is used in the `financial institution' sense but as "rive" when it means `edge of river'. Choosing the correct translation is important for retrieval since French documents about finance are far more likely to contain the word "banque" than "rive". A CLIR system which employs a bilingual dictionary must find a way of coping with this translation ambiguity.

The process of identifying the meanings of words in text is known as word sense disambiguation (WSD) and has been extensively studied in language processing. WSD is normally carried out by selecting the appropriate sense for a context from a lexical resource such as a dictionary or thesaurus but for CLIR it is more appropriate to consider the set of senses as the possible translations of a term between the source and target languages. For example, in an English-to-French CLIR system the word "bank" would have (at least) two possible senses (the translations "banque" and "rive"). By considering the problem of translation selection as a form of WSD allows us to make use of the extensive research which has been carried out in that area.

EuroWordNet (EWN) (Vossen, 1998) is a lexical database which contains possible translations of words between several European languages and was designed for use in CLIR (Gilarranz et. al., 1997). The remainder of this paper is organized as follows: we begin by describing the WSD algorithm used to resolve query ambiguity. We then describe experiments which were used to determine the improvement in performance which may be gained from using WSD for CLIR. We then describe an evaluation of the WSD algorithm and, finally, discuss the implications and conclusions which can be drawn from this work.

Word Sense Disambiguation

One of the main challenges in using a resource such as EWN is discovering which of the synsets are appropriate for a particular use of a word. In order to do this we adapted a WSD algorithm for WordNet originally developed by Resnik (1999). The algorithm is designed to take a set of nouns as context and determine the meaning of each which is most appropriate given the rest of the nouns in the set. This algorithm was used to disambiguate nouns in the retrieval queries. Space restrictions do not allow us to describe the algorithm here but a full description may be found in (Resnik, 1999) and an account of how the approach was adapted to fit into a CLIR system in (Clough and Stevenson, 2004).

Experimental Setup

Test Collection

Evaluation was carried out using past results from the cross-lingual track of TREC6 (Schauble et. al., 1997). We used only TREC6 runs that retrieved from an English language collection, which was the 242,918 documents of the Associated Press (AP), 1988 to 1990. NIST supplied 25 English CLIR topics, although four of these (topics 3, 8, 15 and 25) were not supplied with any relevance judgments and were not used for this evaluation.

The topics were translated into four languages (Spanish, German, French and Dutch) by native speakers who attempted to produce suitable queries from the English version. For this evaluation the Spanish queries were used to evaluate the cross-lingual retrieval and the English queries to provide a monolingual baseline. Spanish was chosen since it provides the most complete and accurate translation resource from the EWN languages. In addition the EWN entries for Spanish tend to have more senses than several of the other languages and is therefore a language for which WSD is likely to be beneficial.

In order to evaluate the contribution of the WSD algorithm and EWN separately the English and Spanish queries were manually disambiguated by the authors. The possible synsets were identified for each query (for the Spanish queries these were mapped from the Spanish synsets onto the equivalent English ones which would be used for retrieval). A single sense from this set was then chosen for each term in the query.

CLIR System

Our CLIR system employs 3 stages: term identification, term translation and document retrieval. The term identification phase aims to find the nouns and proper names in the query. The XEROX part of speech tagger (Cutting et. al., 1992) is used to identify nouns in the queries. Those are then lemmatised and all potential synsets identified in EWN. For these experiments the Spanish lemmatisation was manually verified and altered when appropriate. This manual intervention could be omitted given an accurate Spanish lemmatiser. For English queries this set of possible synsets were passed onto the WSD algorithm to allow the appropriate one to be chosen. Once this has been identified the terms it contains are added to the final query. (In the next Section we describe experiments in which different synset elements are used as query terms.) For Spanish queries the EWN Inter-Lingual-Index (Vossen, 1998) was used to identify the set of English WordNet synsets for each term which is equivalent to the set of possible translations. For each word this set of synsets was considered to be the set of possible senses and passed to the WSD algorithm which chooses the most appropriate. Non-translatable terms were included in the final translated query because these often include proper names which tend to be good topic discriminators. Document retrieval was carried out using our own implementation of a probabilistic search engine based on the BM25 similarity measure. A more complete system description may be found in (Clough and Stevenson, 2004).

Evaluation Method

We experimented with various methods for selecting synsets from the query terms: all synsets, the first synset and the synset selected by the WSD algorithm. It is worth mentioning here that WordNet synsets are ordered by frequency of occurrence in text and consequently the first synset represents the most likely prior sense. We also varied the number of synset members selected: either the headword (first member of the synset), or all synset terms. In the case of all synset terms, we selected only distinct terms between different synsets for the same word (note this still allows the same word to be repeated within a topic). This was done to reduce the effects of term frequency on retrieval, thereby making it harder to determine how retrieval effectiveness is affected by WSD alone. Preliminary experiments showed retrieval to be higher using distinct words alone. We also experimented with longer queries composed of the TREC6 title and description fields, as well as shorter queries based on the title only to compare the effects of query length with WSD.

Retrieval effectiveness is measured using the *trec_eval* program as supplied by NIST. With this program and the set of relevance documents as supplied with the TREC6 topics, we are able to determine how many relevant documents are returned in the top 1000 rank positions, and the position at which they occur. We use two measures of retrieval effectiveness computed across all 25 topics. The first is recall which measures the number of relevant documents retrieved. The second measure, *mean uninterpolated average precision* (MAP), is calculated as the average precision figures obtained after each new relevant document is seen (Baeza-Yates and Ribero-Neto, 1999).

CLIR Evaluation

The results of cross-lingual retrieval can be placed in context by comparing them against those from the monolingual retrieval using the English version of the title and description as the query. (EuroWordNet was not used here and no query expansion was carried out.) It was found that 979 documents were recalled with a MAP score of 0.3512. These results form a reasonable goal for the cross-lingual retrieval to aim towards.

Synset selection	Synset members	Relevant Retrieved	MAP
gold	all	890	0.2823
	1st	676	0.2459
all	all	760	0.2203
	1st	698	0.2215
1st	all	707	0.2158
	1st	550	0.1994
WSD	all	765	0.2534
	1st	579	0.2073
Monolingual		979	0.3512

Table 1: Results for Spanish retrevial with title and description

Table 1 shows retrieval results after translating the title and description. The first column ("synset selection") lists the methods used to choose the EWN synset from the set of possibilities. "gold" is the manually chosen sense, "all" and "1st" are the two baselines of choosing all possible synsets and the first while "auto" is the senses chosen by the WSD algorithm. The next column ("synset members") lists the synset members which are chosen for query expansion, either all synset members or the first one.

The best retrieval scores for manually disambiguated queries is recorded when all synset members are used in the query expansion which yields a MAP score of 0.2823

(see Table row “gold”, “all”). This is around 80% of the monolingual retrieval score of 0.3512. When WSD is applied the highest MAP score of 0.2534 is achieved when all synset members are selected (Table 1 row “WSD”, “all”). This represents 72% of the MAP score from monolingual retrieval and 90% of the best score derived from the manually disambiguated queries.

In the majority of cases choosing all synset members leads to a noticeably higher MAP score than retrieval using the first synset member. This is probably because the greater number of query terms gives the retrieval engine a greater chance of finding the relevant document. The exception is when all synsets have been selected (see Table 1). In this case the retrieval engine already has a large number of query terms thorough the combination of the first member from all synsets and adding more makes only a slight difference to retrieval performance.

When translating queries, it would appear that using Resnik's algorithm to disambiguate query terms improves retrieval performance when compared against choosing all possible senses or the first (most likely) senses to disambiguate.

The experiments were repeated, this time using just the title from the TREC query, representing a shorter query. The results of this experiment were similar to those when the title and description were used. It was also found that the shorter queries benefit far more from query expansion. These experiments are fully reported in (Clough and Stevenson, 2004).

Retrieval Results for Individual Queries

Averaged scores across topics often hide interesting patterns as large scores affect the mean value. Generalising across topics does not help us to find out which queries are performing well or poorly. Figure 1 shows the MAP scores across the English baseline queries (title and description) and their translated versions (using all synset words from automatic disambiguation). As expected most of the Spanish queries perform worse than the monolingual retrieval using the English queries. Perhaps surprising are queries which actually produce better retrieval results than the original baseline (topics 4, 16, 18, 19 and 24). This is not uncommon in CLIR when the translation includes additional words not in the query (a kind of query expansion), or use words which better discriminate between topics. For example, the original English version of query 24 contains the following terms: “teddy bears”, “popularity”, “world wide”. The Spanish equivalent for this system setting, however, includes the following terms: “bear”, “teddy bear”, “plush”, “felt”, “popularity”, “earth”, “globe”, “world”. The number of query terms is much greater than the original English version because in some cases the WSD algorithm was uncertain about its sense selection and therefore selected all senses. But for this query several of the possible synsets are quite similar and, while some did not match the gold standard sense selection, they contained terms which were useful for the retrieval engine.

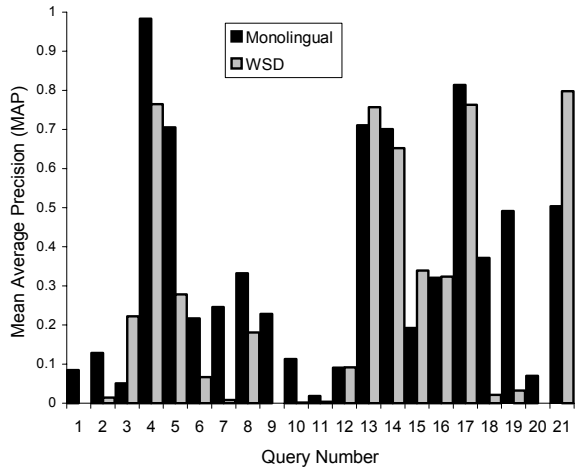

Figure 1:MAP scores across queries for English baseline (“monolingual”) and Spanish automatic translation (“WSD”)

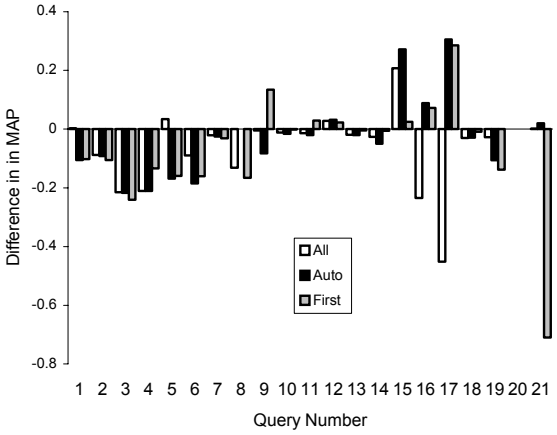

Figure 2: MAP scores for all synsets, first synset and synset selected by WSD compared with manually chosen baseline. MAP scores were computed using all words in chosen synsets for retrieval.

Figure 2 shows the differences in MAP scores for each Spanish query, based on the scores obtained using manual disambiguation (“gold” – see Table 1), selecting all senses (“all”), selecting the first sense (“first”) and selecting senses automatically (“auto”). In this figure the MAP score for each of these methods was calculated from retrieval using all synset members. The percentage difference is calculated as *MAP(system) -MAP(gold)*, where a value of 0 indicates no difference between selecting a suitable sense manually or using an automatic selection method, a positive score indicating queries in which the automatic selection has performed better than the gold, and a negative score indicating the gold has performed better.

It can be seen that in the majority of cases manual disambiguation gives better results than either of the three approaches. For the majority of queries the three approaches do not perform much worse than the gold standard obtained through manual disambiguation and in some cases outperform it. Across all queries using WSD is the best of the three approaches but the MAP scores for the other two are not much worse. However Figure 2 shows that the two baseline approaches ("all" and "first") can perform very badly on individual queries. For example, for queries 19 and 20 choosing all synsets performs noticeably worse than either using WSD or choosing the first synset. However, for query 24 choosing the first synset performs particularly poorly.

An interesting feature of the WSD algorithm is that it varies the number of senses selected depending on its confidence of picking the correct one. This can be seen in the barchart in the cases where either selecting the first sense or all senses gives better results. It appears that in most cases the WSD algorithm is acting like one of these approaches (i.e. it either selects several senses, or it selects one sense which happens to be the first), this variable selection method giving better performance than using just one of the strategies. Of course, there are cases when all methods do badly, in these queries using the WSD algorithm or not does not appear to make much difference (for the better or worse). There appear few cases in Figure 1 when the WSD algorithm gives an increase or decrease on its own, rather it tends to act like one of the other approaches.

Conclusion

There has been some disagreement over the usefulness of WSD for monolingual retrieval (see, for example, (Sanderson, 1994; Jing and Tzoukermann, 1999). In particular Krovetz and Croft (1992) and Sanderson (1994) showed that WSD had to be accurate to be useful for monolingual retrieval. However, the results presented here imply that this is not the case for CLIR since pervious experiments (Clough and Stevenson, 2004) showed that the WSD algorithm was not particularly accurate. The reason for this difference may be that retrieval algorithms actually perform a similar purpose to WSD algorithms in the sense that they attempt to identify instances of words being used with the relevant meanings. WSD algorithms therefore need to be accurate to provide any improvement. The situation is different for CLIR where identifying the correct translation of words in the query is unavoidable. This can only be carried out using some disambiguation method and the results presented here suggest that some disambiguation is better than none for CLIR.

In future work we plan to experiment with different retrieval engines, in particular those that support structured queries. This will allow us to make use of the information contained in EuroWordNet in a principled way which may benefit retrieval.

Acknowledgments

The work described here was supported by the EPSRC-funded Eurovision project at Sheffield University (GR/R56778/01).

References

Baeza-Yates, R., Ribeiro-Neto, B. (1999) Modern Information Retrieval. Addison Wesley Longman Limited, Essex

Ballesteros, L., Croft, W. (1998) Resolving ambiguity for cross-language retrieval (pp 64-71) In Research and Development in Information Retrieval

Clough, P. and Stevenson, M. (2004) Evaluating the Contribution of EuroWordNet and Word Sense Disambiguation to Cross-language Retrieval In Proceedings of the Second International Global WordNet Conference (GWC-2004) (pp 97-105) Brno, Czech Republic

Cutting, D., Kupiec, J., Pedersen, J., Sibun, P. (1992) A practical part-of-speech tagger. In: Proceedings of the Third Conference on Applied Natural Language Processing (pp 133-140) Italy

Gilarranz, J., Gonzalo, J., Verdejo, F. (1997) Language-independent text retrieval with the EuroWordNet Multilingual Semantic Database In: Proceedings of the Second Workshop on Multilinguality in the Software Industry: the AI contribution (pp 9 -16), Nagoya, Japan

Jang, M., Myaeng, S., Park, S. (1999) Using mutual information to resolve query translation ambiguities and query term weighting. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99) (pp 223-229), College Park, MA

Jing, H., Tzoukermann, E. (1999) Information retrieval based on context distance and morphology. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99) (pp. 90-96) Seattle, WA

Krovetz, R., Croft, B. (1992) Lexical ambiguity and information retrieval. ACM Transactions on Information Systems 10:115-141

Resnik, P. (1999) Disambiguating Noun Groupings with Respect to WordNet senses In Armstrong, S., Church, K., Isabelle, P., Manzi, S., Tzoukermann, E., Yarowsky, D., eds.: Natural Language Processing using Very Large Corpora (pp77-98) Kluwer Academic Press

Sanderson, M. (1994) Word sense disambiguation and information retrieval. In: Proceedings of the 17th ACM SIGIR Conference (pp 142-151), Dublin, Ireland

Schaüble, P., Sheridan, P. (1997) Cross-Language Information Retrieval (CLIR) Track Overview In Voorhees, E., Harman, D., eds.: The Sixth Text Retrieval Conference (TREC-6) (pp 31-44) Gaithersburg, MA

Vossen, P. (1998) Introduction to EuroWordNet. Computers and the Humanities 32:73--89 Special Issue on EuroWordNet