Evaluating the Contribution of EuroWordNet and Word Sense Disambiguation to Cross-language Information Retrieval

Paul Clough and Mark Stevenson

University of Sheffield
Regent Court, 211 Portobello Street,
Sheffield, S1 4DP
United Kingdom
Email: p.d.clough@sheffield.ac.uk, marks@dcs.shef.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 which 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 WSD algorithm for this purpose. It was found that this algorithm achieved only around 50% correct disambiguation when compared with manual judgment, however, retrieval performance using the senses it returned was 90% of that recorded using manually disambiguated queries.

1 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. An example is e-commerce where a consumer may be interested in purchasing some computer equipment from another country but does not know how to describe what they want in the relevant 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 [1,2]. 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

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 97–105.

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) [3] is a lexical database which contains possible translations of words between several European languages and was designed for use in CLIR [4]. Section 2 describes the WSD algorithm we use to resolve ambiguity in the retrieval queries. In Section 3 we describe the experiments which were used to determine the improvement in performance which may be gained from using WSD for CLIR the results of which are presented in Section 4. Section 5 described an evaluation of the WSD algorithm used. The implications and conclusions which can be drawn from this work are presented in Sections 6 and 7.

2 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 [5]. 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 thought to be suitable for disambiguating the nouns in retrieval queries.

The algorithm is fully described in [5] and we shall provide only a brief description here. The algorithm makes use of the fact that WordNet synsets are organised into a hierarchy with more general concepts at the top and more specific ones below them. So, for example, `motor vehicle` is less informative than `taxi`. A numerical value is computed for each synset in the hierarchy by counting the frequency of occurrence of its members in a large corpus[1]. This value is dubbed the *Information Content* and is calculated as $\textit{Information Content}(\textit{synset}) = -\log \Pr(\textit{synset})$.

The similarity of two synsets can be found by choosing the synset which is above both in the hierarchy with the highest information content value (i.e. the most specific). By extension of this idea, sets of nouns can be disambiguated by choosing the synsets which return the highest possible total information content value. For each sense a value is returned indicating the likelihood that the sense being the appropriate one given the group of nouns.

3 Experimental Setup

3.1 Test Collection

Evaluation was carried out using past results from the cross-lingual track of TREC6 [6]. 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

[1] We used the British National Corpus which contains 100 million words.

CLIR topics, although four of these (topics 3, 8, 15 and 25) were not supplied with any relevance judgements 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.

3.2 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 [7] is used to identify nouns in the queries. Those are then lemmatised and all potential synsets identified in EWN.[2] 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 [3] was used to identify the set of English WordNet synsets for each term which is equivalent to 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 (see, e.g. [8]). The BM25 function estimates term frequency as Poisson in distribution, and takes into account inverse document frequency and document length. Based on this weighting function, queries are matched to documents using a similarity measure based upon term co-occurrence. Any document containing at least one or more terms from the query is retrieved from the index and a similarity score computed for that document:query pair. Documents containing any number of query terms are retrieved (creating an OR'ing effect) and ranked in descending order of similarity under the assumption that those nearer the top of the ranked list are more relevant to the query than those nearer the bottom.

[2] For these experiments the Spanish lemmatisation was manually verified and altered when appropriate. This manual intervention could be omitted given an accurate Spanish lemmatiser.

3.3 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 [9].

4 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.

Table 1. Results for Spanish retrieval with title and description

synset selection	synset members	recall	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

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 1 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.

Table 2. Results for Spanish retrieval with title only

synset selection	synset members	recall	MAP
gold	all	828	0.2712
	1st	685	0.2192
all	all	735	0.2346
	1st	640	0.1943
1st	all	658	0.2072
	1st	511	0.1689
WSD	all	758	0.2361
	1st	650	0.2007

The experiments were repeated, this time using just the title from the TREC query which represents a shorter query. The results from these experiments are shown in Table 2. The manually annotated queries produces the highest MAP of 0.2712 (77% of monolingual). When the WSD algorithm is used the highest MAP is also recorded when all synset members were chosen. This score was 0.2361 (67% of monolingual). However, when the shorter queries are used the difference between WSD the two naive approaches (choosing the most frequent sense and choosing all senses) is much smaller. This is probably because the reduced

amount of context makes it difficult for the WSD algorithm to make a decision and it often returns all senses.

Table 2 also shows that choosing all synset members is a more effective strategy than choosing just the first member. We already noted this with reference to the results form the longer queries (Table 1) although the difference is more pronounced than when the longer queries were used. In fact it can be seen that when the short queries are used choosing all members for each possible synset (i.e. no disambiguation whatsoever) scores higher than choosing just the first member of the manually selected best sense. This shows that these shorter queries benefit far more from greater query expansion and that even correct meanings which are not expanded much do not provide enough information for correct retrieval.

5 Evaluation of WSD

It is important to measure the effectiveness of the WSD more directly than examining CLIR results. Others, such as [10,11], have found that WSD only has a positive effect on monolingual retrieval when the disambiguation is accurate. The manually disambiguated queries were used as a gold-standard aginst which the WSD algorithm we used could be evaluated. Two measures of agreement were computed: strict and relaxed. Assume that a word, w, has n senses denoted as $senses(w)(= w_1, w_2, ...w_n)$ and that one of these senses, w_{corr} (where $1 \leq corr \leq n$), was identified as correct by the human annotators. The WSD algorithm chooses a set of m senses, $wsd(w)$, where $1 \leq m \leq n$. The strict evaluation score for w takes into account the number of senses assigned by the WSD algorithm and if $w_{corr} \in wsd(w)$ the word is scored as $\frac{1}{m}$ (and 0 if $w_{corr} \notin wsd(w)$). The relaxed score is a simple measure of whether the WSD identified the correct senses regardless of the total it assigned and is scored as 1 if $w_{corr} \in wsd(w)$. The WSD accuracy for an entire query is calculated as the mean score for each term it contains.

The two evaluation metrics have quite different interpretations. The strict evaluation measures the degree to which the senses identified by the WSD algorithm match those identified by the human annotators. The relaxed score can be interpreted as the ratio of query words in which the sense identified as correct was not ruled out by the WSD algorithm. In fact simply returning all possible senses for a word would guarantee a score of 1 for the relaxed evaluation, although the score for the strict evaluation would probably be very low. Since it is important not to discard the correct sense for retrieval purposes the relaxed evaluation may be more relevant for this task.

Table 3. Results of WSD algorithm and first sense baseline compared against manually annotated queries

		Score	
Language	**Method**	Strict	Relaxed
English	WSD	0.410	0.546
	1st synset	0.474	
Spanish	WSD	0.441	0.550
	1st synset	0.482	

Table 3 shows the results of the evaluation of the WSD algorithm and baseline method of choosing the first sense against the manually annotated text for both the Spanish and English queries. The baseline scores are identical for each metric since it assigns exactly one sense for each word (the first) and the two metrics only return different scores when the technique assigns more than one sense.

We can see that the evaluation is similar across both languages. The baseline method actually outperforms automatic WSD according to the strict evaluation measure but scores less than it when the relaxed measure is used. We can also see that neither of the approaches are particularly accurate and often rule out the sense that was marked as correct by the human annotator.

However the results from the cross-language retrieval experiments earlier in this Section show that there is generally an improvement in retrieval performance when the WSD algorithm is used. This implies that the relaxed evaluation may be a more appropriate way to judge the usefulness of a WSD algorithm for this task. This idea has some intuitive plausibility it seems likely that for retrieval performance it is less important to identify the sense which was marked correct by an annotator than to try not to remove the senses which are useful for retrieval. It should also be borne in mind that the human annotation task was a forced choice in which the annotator had to choose exactly one sense for each ambiguous query term. In some cases it was very difficult to choose between some of the senses and there were cases where none of the EWN synsets seemed completely appropriate. On the other hand our WSD algorithm tended to choose several senses when there was insufficient contextual evidence to decide on the correct sense.

6 Discussion

The WSD algorithm's approach of only choosing senses when there is sufficient evidence suits this task well. However, the WSD results also highlight a serious limitation of EWN for CLIR. EWN's semantics are based on ontological semantics using the hyponymy relationship. That is, the EWN synset hierarchy contains information about the type of thing something is. So, for example, it tells us that "car" is a type of "motor vehicle". However, many types of useful semantic information are missing. One example is discourse and topic information. For example, "tennis player" (a hyponym of person) is not closely related to "racket", "balls" or "net" (hyponyms of artifact). Motivated by this example, Fellbaum [12] dubbed this the "tennis problem". This information is potentially valuable for retrieval where one aim is to identify terms which model the topic of the query.

Others, including [1,13,14], have used word co-occurrence statistics to identify the most likely translations and this could be considered a form of translation. This approach seems promising for CLIR since it returns words which occur together in text and these are likely to be topically related. This approach has potential to be developed into a WSD algorithm which could be applied to EWN.

There has been some disagreement over the usefulness of WSD for monolingual retrieval (see, for example, [11,15]). In particular [10,11] 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 the WSD methods were hindered by a lack of context and were 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.

7 Conclusions

The results presented in this paper show that WSD is useful when CLIR was being carried out using EWN. The WSD algorithm used was not highly accurate on this particular task however it was able to outperformed two simple baselines and did not appear to adversely effect the retrieval results.

In future work we plan to experiment with different languages which are supported by EWN to test whether the differences in lexical coverage of the various EWNs have any effect on retrieval performance. One of the authors has already shown that combining WSD algorithms can be a useful way of improving their effectiveness for ontology construction [16]. We plan to test whether similar techniques could be employed to improve the automatic disambiguation of queries.

Acknowledgments

The authors are grateful for advice from Mark Sanderson and Wim Peters of Sheffield University. The work described here was supported by the EPSRC-funded Eurovision project at Sheffield University (GR/R56778/01).

References

1. Ballesteros, L., Croft, W.: Resolving ambiguity for cross-language retrieval. In: Research and Development in Information Retrieval. (1998) 64–71.
2. Jang, M., Myaeng, S., Park, S.: 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), College Park, MA (1999) 223–229.
3. Vossen, P.: Introduction to EuroWordNet. Computers and the Humanities **32** (1998) 73–89 Special Issue on EuroWordNet.
4. Gilarranz, J., Gonzalo, J., Verdejo, F.: Language-independent text retireval with the EuroWordNet Multilingual Semantic Database. In: Proceedings of the Second Workshop on Multilinguality in the Software Industry: the AI contribution, Nagoya, Japan (1997) 9–16.
5. Resnik, P.: 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. Kluwer Academic Press (1999) 77–98.
6. Schaüble, P., Sheridan, P.: Cross-Language Information Retrieval (CLIR) Track Overview. In Voorhees, E., Harman, D., eds.: The Sixth Text REtrieval Conference (TREC-6), Gaithersburg, MA (1997) 31–44.
7. Cutting, D., Kupiec, J., Pedersen, J., Sibun, P.: A practical part-of-speech tagger. In: Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy (1992) 133–140.

8. Robertson, S., Walker, S., Beaulieu, M.: Okapi at TREC-7: automatic ad hoc, filtering VLC and interactive track. In: NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7), Gaithersburg, MA (1998) 253–264.
9. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley Longman Limited, Essex (1999).
10. Krovetz, R., Croft, B.: Lexical ambiguity and information retrieval. ACM Transactions on Information Systems **10** (1992) 115–141.
11. Sanderson, M.: Word sense disambiguation and information retrieval. In: Proceedings of the 17th ACM SIGIR Conference, Dublin, Ireland (1994) 142–151.
12. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database and some of its Applications. MIT Press, Cambridge, MA (1998).
13. Gao, J., Nie, J., He, H., Chen, W., Zhou, M.: Resolving query translation ambiguity using a decaying co-occurence model and syntactic dependence relations. In: Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retreival, Tampere, Finalnd (2002) 183–190.
14. Qu, Y., Grefenstette, G., Evans, D.: Resolving translation ambiguity using monolingual corpora. In: Cross Language Evaluation Forum 2002, Rome, Italy (2002).
15. Jing, H., Tzoukermann, E.: 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), Seattle, WA (1999) 90–96.
16. Stevenson, M.: Augmenting Noun Taxonomies by Combining Lexical Similarity Metrics. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING-02), Taipei, Taiwan (2002) 953–959.