Sheffield University CLEF 2000 Submission - Bilingual Track: German to English

Tim Gollins and Mark Sanderson

Department of Information Studies, University of Sheffield, Sheffield, South Yorkshire, UK
m.sanderson@sheffield.ac.uk

Abstract. We investigated dictionary based cross language information retrieval using lexical triangulation. Lexical triangulation combines the results of different transitive translations. Transitive translation uses a pivot language to translate between two languages when no direct translation resource is available. We took German queries and translated then via Spanish, or Dutch into English. We compared the results of retrieval experiments using these queries, with other versions created by combining the transitive translations or created by direct translation. Direct dictionary translation of a query introduces considerable ambiguity that damages retrieval, an average precision 79% below monolingual in this research. Transitive translation introduces more ambiguity, giving results worse than 88% below direct translation. We have shown that lexical triangulation between two transitive translations can eliminate much of the additional ambiguity introduced by transitive translation.

Introduction and Background

Cross Language Information Retrieval (CLIR) addresses the situation where the query that a user presents to an IR system, is not in the same language as the corpus of documents they wish to search. This situation presents a number of challenges (Grefenstette (1998)) but primary amongst these is the problem of crossing the language barrier (Schauble & Sheridan (1997)). Almost all the approaches to this problem require access to some form of rich translation resource to map terms in the query language (the source) to terms in the corpus (the target). "Transitive" CLIR aims to address the situation where there are limited direct translation resources available (Ballesteros (2000)).

A transitive CLIR system translates the source language terms by first translating the terms into an intermediate or "pivot" language and then translating the resulting terms into the target language. Thus, a transitive system could translate a query from German to English via either Dutch, or Spanish.

The main aim of this work is to combine translations from two different transitive routes to discover if this can reduce the ambiguity introduced by transitive translation. Ballesteros suggested the possibility of using this approach in the summary to her recent chapter (Ballesteros (2000)). We have chosen to call this approach "lexical triangulation", see figure 1.

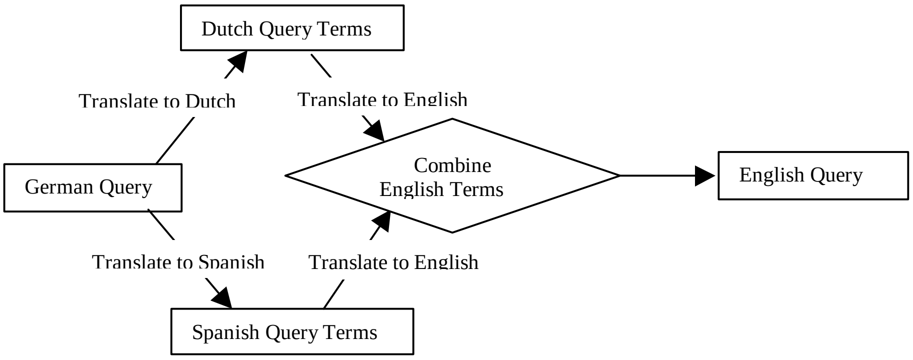


Fig. 1. Lexical triangulation

We have chosen to simulate a Machine-Readable Dictionary (MRD) approach to CLIR. This follows on from the work of Ballesteros & Croft (1996, 1997, 1998), and Ballesteros (2000).

The Experimental Environment

The underlying IR system used in the Sheffield submission was the GLASS system (Sanderson (2000)).

The translation resources were derived from the German, Spanish, Dutch, and English components of EuroWordNet (Vossen (1999)). The data used to lemmatise the German queries was derived from the CELEX German databases.

EuroWordNet

Given that the intention of this work is to examine CLIR using simulated Machine Readable Dictionaries, the choice of EuroWordNet (Vossen (1999)) as the primary translation resource may appear a little strange. The primary basis for this choice was availability[1].

The intention of the EuroWordNet project was to develop a database of WordNets for a number of European languages similar to, and linked with, the Princeton WordNet 1.5 (Vossen (1997)). This effectively makes English the inter lingua that all the other languages link through. One of the intended uses of EuroWordNet was in multi-lingual information retrieval (Vossen (1997)). Gonzalo, et al. (1998) describes a possible implementation.

By developing a series of WordNets for European languages, and linking them to the original Princeton 1.5 WordNet for English, EuroWordNet has created a structure

[1] The Sheffield University Computer Science Department was a collaborator in the EuroWordNet project and Wim Peters of that department kindly made extracts from EuroWordNet available for this research.

similar to the controlled vocabulary thesaurus used by Salton as described by Oard & Dorr (1996). The structure is also very similar to the structure developed by Diekema, et al. (1998). The Princeton WordNet consists of synonyms grouped together to form "synsets", basic semantic relationships link these together to form the WordNet (Vossen (1997), Miller, et al. (2000)). Each synset has a unique identifier (synset-id).

In EuroWordNet, the relationships between the synsets of the various component languages and the Princeton 1.5 WordNet synsets[2] can take many forms. These include, for example, the eq_hyponym[3] relation, which relates more general to more specific concepts (Vossen (1997)).

Our work used EuroWordNet to generate structures to simulate a Machine Readable Dictionary. The only relationships used in the construction of the dictionary tables, were the eq_synonym and eq_near_synonym relationships. These are by far the most restrictive and precise of the possible relationships.

The eq_synonym relationship records the fact that the language synset is synonymous with the WordNet synset. EuroWordNet introduced the eq_near_synonym relationship to record the fact that certain terms that share a common hypernym (more general concept) are closer in meaning than others. In this situation the co-hyponyms (more specific terms) that are closely related are close enough in meaning that they could be used for translation purposes, but are not synonymous and are therefore not in the same synset. This closeness is represented by linking the synsets with an eq_near_synonym relationship (Vossen (1997)).

For each language used from EuroWordNet, two tables were generated. The first mapped lemmas to the synset-ids of the synsets related by eq_synonym or eq_near_synonym. The second maps synset-ids to their constituent lemmas (i.e. related by eq_synonym or eq_near_synonym). As we will explain below, these tables are used to parameterise the translation process.

The translation and processing of queries

Query processing was fully automatic and the queries were generated using all parts of the topics. The queries were passed through a series of processes as follows:

- Parsing - The conversion of the topics to queries which makes use of title, description and narrative fields.
- Normalisation - all characters were reduced to the lower case unaccented equivalents (i.e. "Ö" reduced to "o" and "É" to "e" etc.) in order to maximise matching in both the lemmatisation and translation processes.
- Lemmatisation - The various inflected forms of the query words were reduced to a canonical lemma form to enable matching with the German EuroWordNet translation resources. A table derived from the CELEX German database was

[2] In EuroWordNet terms the Inter Lingual Index or ILI.

[3] The relationships in EuroWordNet have names on the form eq_*relationship_name* the eq_ indicates that the relationship involves some degree of "equality".

used to determine the appropriate lemmata[4] for a word form. German compound words were split using a simple algorithm. The algorithm looks for a series of word forms that will match with the whole compound. If such a complete match is found the corresponding lemmata of the word forms are returned. The algorithm takes account of the use of "s" as "glue" in the construction of German compounds. This approach was based on the description of the word reduction module in Sheridan & Ballerini (1996). All of the CELEX data was normalised to unaccented lower case for matching with the query words.

- German Stop Word Removal - A stopword list, generated from the CELEX German database, was used to remove words in the query that carried little meaning and would otherwise introduce noise to the translation. The stop-word lists contain all of the German words marked as articles, pronouns, prepositions, conjunctions or interjections in the CELEX database.
- Translation - The translation process used tables derived from EuroWordNet to translate between two languages. The lemma to synset-id table for the first language and the synset to lemma table for the second language were used to map words in the first language to words in the second. All the possible translations through the intermediate synset-ids were returned. Three different translations were created for each query: a direct German to English translation, a transitive translation using Spanish as the intermediate language, and a transitive translation using Dutch as the intermediate language.
- Merging - The results of the two transitive translation routes were merged to produce a fourth translation, the triangulated translation. The merge process was conducted on an "original German Lemma" by "original German Lemma" basis. The translations from each route for each lemma were compared and only translations common to both routes were used to translate the lemma.
- Retrieval – the translation and merging process produced four different versions of the queries translated into English, these were submitted to the GLASS IR system which had been used to index the English corpus. The GLASS system normalised both documents and queries to lower case, and removed any English stopwords (using a standard English stop word list). Porter stemming (Porter (1980)) was used on both the queries and the collection. No special processing was used on the corpus.

The Experimental story

We submitted four official runs to the CLEF evaluation process.

- A "bilingual" run (shefbi), generated from the direct translation from German to English

[4] The wordform to lemma table is a many-to-many mapping as a wordform may be a valid inflection of more than one lemma.

- A “Spanish transitive” run (shefes), generated from the transitive translation using Spanish as the intermediate.
- A “Dutch transitive” run (shefnl), generated from the transitive translation using Dutch as the intermediate.
- And a “triangulated” run (sheftri), generated from the result of merging of the two transitive translations.
- Only the triangulated run (sheftri) was judged and contributed to the relevance judgement pool.

In order to provide a baseline for comparison we conducted an additional English monolingual run using the same parsing and retrieval processes. This unofficial run is presented below to enable comparisons to be made.

In summary, the experimental conditions were as follows:

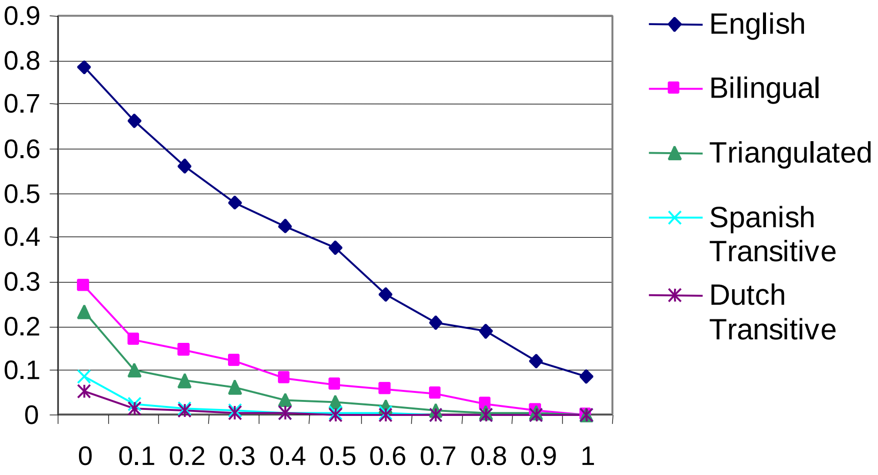
Experimental Variable	Value for this experiment
Queries	CLEF 2000 CLIR, German and English
Corpus	LA Times 1994- CLEF Collection
Relevance Judgements	CLEF 2000 pool
Corpus and Query Stemming	Yes, Porter based
Lemmatiser	Yes, including German Compound Splitting
German Stop-words removed pre-translation	Yes, all articles, pronouns, prepositions, conjunctions or interjections from the CELEX German database.
Translation	Simulated Dictionary based, using lookup-tables derived from EuroWordNet eq_synonym and eq_near_synonym relations.
Merging Strategy for Lexical triangulation	Only translations common to both transitive routes.

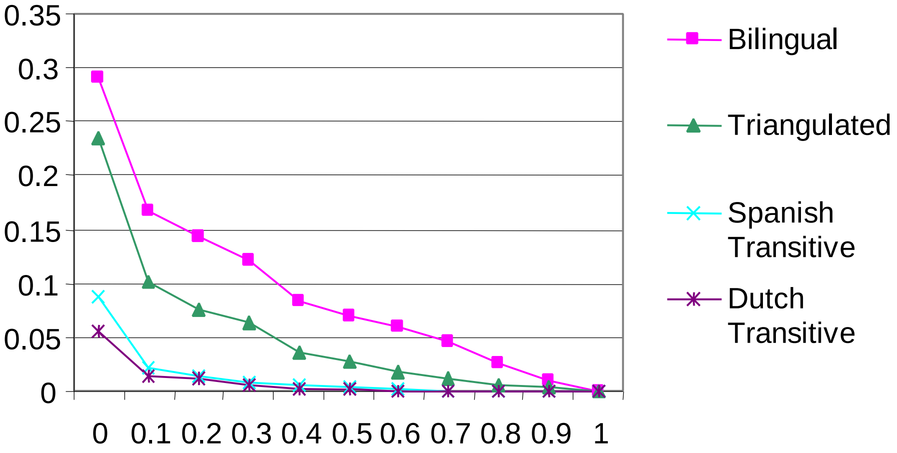
Results

The table below shows the average precision for the five runs that made up the CLEF experiment. Only the cross language runs were submitted to the CLEF, and of those, only the triangulated run contributed to the pooled results.

	Porter, Intersection
English	0.3593
Bilingual (shefbi)	0.0856
Triangulated (sheftri)	0.0458
Spanish Transitive (shefes)	0.0098
Dutch Transitive (shefnl)	0.007

The standard 11-point recall and precision curves for the five runs are shown below, the second graph shows only the four cross language runs.





Analysis

Comparing the average precision of the monolingual run with the bilingual run we see that the bilingual run is some 76%[5] below the monolingual. This compares to the 60% below worst case reported by Ballesteros & Croft (1996) when considering word by word dictionary based Spanish to English CLIR.

Taking next the two transitive runs, we observe a differential of -88% in the case of the Spanish transitive run and -92% in the case of the Dutch transitive run relative to the bilingual run. Both of these results are statistically significant at the 0.01 level under both the sign and Wilcoxon tests. These figures are in line with the -92%

[5] Statistically significant at the 0.01 level under both the sign and Wilcoxon tests.

differentials reported by Ballesteros (2000) for transitive retrieval of Spanish – French CLIR with English as the pivot compared to Spanish – French direct translation.

Comparing the triangulated run with the two transitive runs reveals the expected improvement in performance. The differentials for the two transitive runs relative to the triangulated run are -79% for the Spanish transitive run and -85% for the Dutch transitive. Both of these figures are statistically significant at the 0.01 level under both the sign and Wilcoxon tests.

There is also a statistically significant differential of -47% between the triangulated run and the bilingual in favour of the bilingual. This significance is at the 0.01level under both the sign and Wilcoxon tests.

Conclusion

In summary, these results support the results of Ballesteros (2000) with respect to the behaviour of transitive translation in CLIR. They also support the hypotheses we set out to prove that lexical triangulation has the beneficial effect of improving the results from transitive translation in dictionary based CLIR.

This work made use of relatively rich resources in the form of EuroWordNet. However, it remains to be seen if these results could be repeated using the poorer quality resources that are likely to be available for translating between less common pairs of languages.

As Samuel Johnson said "Dictionaries are like watches; the worst is better than none, and the best cannot be expected to be quite true." (Gendreyzig (2000))

Bibliography

Ballesteros, L. & Croft, B. (1996). "Dictionary methods for cross-lingual information retrieval". In: *Database and Expert Systems Applications. 7th International Conference, DEXA '96 Proceedings*. Springer-Verlag Berlin, Germany. [Online]. Available: http://cobar.cs.umass.edu/pubfiles/ir-98.ps.gz [23/03/2000].

Ballesteros, L. & Croft, W. B. (1997). "Phrasal translation and query expansion techniques for cross-language information retrieval". In: *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 84 - 91. Association for Computing Machinery. [Online]. Available: http://www.acm.org/pubs/articles/proceedings/ir/258525/p84-ballesteros/p84-ballesteros.pdf [29/02/2000].

Ballesteros, L. & Croft, W. B. (1998). "Resolving ambiguity for cross-language retrieval". In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Association for Computing Machinery. [Online]. Available: http://www.acm.org/pubs/articles/proceedings/ir/290941/p64-ballesteros/p64-ballesteros.pdf [29/02/2000].

Ballesteros, L. A. (2000). "Cross Language Retrieval via transitive translation". In: Croft, W. B. (ed.) *Advances in Information Retrieval: Recent Research from the CIIR*, pp. 203 - 234 Kulwer Academic Publishers.

Diekema, A., Oroumchian, F., Sheridan, P. & Liddy, E. D. (1998). "TREC-7 Evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French". In:

Voorhees, E. M. & Harman, D. K. (eds.), *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7).* NIST. [Online]. Available: http://trec.nist.gov/pubs/trec7/t7_proceedings.html [15/02/2000].

Gendreyzig, M. (2000). *Collection of Web-Dictionaries*, [Online]. LEO - Link Everything Online. Available: http://dict.leo.org/dict/dictionaries.en.html [24/08/2000].

Gonzalo, J., Verdejo, F., Peters, C. & Calzolari, N. (1998). "Applying EuroWordNet to cross-language text retrieval", *Computers and the Humanities,* **32**(2-3)**,** pp 185-207

Grefenstette, G. (1998). "Problems and approaches to Cross Language Information Retrieval", *Proceedings of the Asis Annual Meeting,* **35,** pp 143-152

Miller, G. A., Chodorow , M., Fellbaum , C., Johnson-Laird, P., Tengi, R., Wakefield, P. & Ziskind, L. (2000). *WordNet - a Lexical Database for English,* [Online]. Cognitive Science Laboratory, Princeton University. Available: http://www.cogsci.princeton.edu/~wn/w3wn.html [23/08/2000].

Oard, D. W. & Dorr, B. J. (1996). A Survey of Multilingual Text Retrieval. (Report). Institute for Advanced Computer Studies and Computer Science Department University of Maryland. [Online]. Available: http://www.clis.umd.edu/dlrg/filter/papers/mlir.ps [09/03/2000]

Porter, M. F. (1980). "An algorithm for suffix stripping". In: Sparck Jones, K. & Willett, P. (eds.), *(1997) Readings in Information Retrieval.*, pp. 313 - 316. San Francisco: Morgan Kaufmann.

Sanderson, M. (2000). *GLASS,* [Online]. Dr Mark Sanderson. Available: http://dis.shef.ac.uk/mark/GLASS/ [25/07/2000].

Schauble, P. & Sheridan, P. (1997). "Cross-Language Information Retrieval (CLIR) Track Overview". In: Voorhees, E. M. & Harman, D. K. (eds.), *NIST Special Publication 500-226: The Sixth Text REtrieval Conference (TREC-6).* NIST. [Online]. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html [15/02/2000].

Sheridan, P. & Ballerini, J. P. (1996). "Experiments in multilingual information retrieval using the SPIDER system". In: Frei, H. P. (ed.) *Proceedings of the 1996 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 96*, pp. 58 - 65. Association for Computing Machinery. [Online]. Available: http://www.acm.org/pubs/articles/proceedings/ir/243199/p58-sheridan/p58-sheridan.pdf [29/02/2000].

Vossen, P. (1997). "EuroWordNet: A Multilingual Database for Information Retrieval". In: *THIRD DELOS WORKSHOP Cross-Language Information Retrieval*, pp. 85-94. European Research Consortium For Informatics and Mathematics. [Online]. Available: http://www.ercim.org/publication/ws-proceedings/DELOS3/Vossen.pdf [01/03/2000].

Vossen, P. (1999). *EuroWordNet Building a multilingual database with wordnets for several European languages.,* [Online]. University of Amsterdam. Available: http://www.hum.uva.nl/~ewn/ [28/02/2000].