Improving Cross Language Retrieval with Triangulated Translation

Tim Gollins
Department of Information Studies
University Of Sheffield
Western Bank, Sheffield, S102TN, UK
lip99tjg@sheffiled.ac.uk

Mark Sanderson
Department of Information Studies
University Of Sheffield
Western Bank, Sheffield, S102TN, UK
+44 (0) 114 22 22648
m.sanderson@shef.ac.uk

ABSTRACT

Most approaches to cross language information retrieval assume that resources providing a direct translation between the query and document languages exist. This paper presents research examining the situation where such an assumption is false. Here, an intermediate (or pivot) language provides a means of transitive translation of the query language to that of the document via the pivot, at the cost, however, of introducing much error. The paper reports the novel approach of translating in parallel across multiple intermediate languages and fusing the results. Such a technique removes the error, raising the effectiveness of the tested retrieval system, up to and possibly above the level expected, had a direct translation route existed. Across a number of retrieval situations and combinations of languages, the approach proves to be highly effective.

1. Introduction

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 being searched. This presents a number of challenges, primarily the problem of translation. Almost all the approaches require access to some form of resource be it through machine translation, a bilingual dictionary, or a comparable/parallel corpora to map terms in the query language (the *source*) to terms in the corpus (the *target*) ([15], [10]). If no such resource is obtainable for a particular language pair, translation via an intermediate or *pivot* language may offer the only means of translation: source language into intermediate, intermediate into target. Languages, which are commonly spoken or for which many translation resources exist are likely candidates for pivots (e.g. English, Spanish, or French). This is largely due to cultural factors such as the language being that of a colonial power. It might be expected that translating via an extra language may only be necessary for uncommon languages. However, as shown in various submissions to the CLIR tracks of TREC, researchers often failed to locate resources, either free or commercial, for translating directly between major European languages such as German and Italian ([5], [7], [11], [9], [8], [12]). For CLIR between most language pairs, use of a pivot is likely to be a common experience.

The research cited viewed pivots as an unfortunate necessity: their use allowed retrieval to take place, but at the cost of much introduced error. Using pivots doubles the number of translations performed in a CLIR system, therefore, increasing the likelihood of translation error, caused mainly by incorrect identification of the senses of ambiguous words. Ballesteros [3] researched a transitive scheme and techniques to overcome word ambiguity. She examined the impact of *transitive translation*, discovering that using simple word-by-word transitive translation from Spanish to French via English degraded performance by 91% when compared to *bilingual translation* direct from Spanish to French. Ballesteros attempted to reduce the ambiguity introduced by transitive translation using query structuring and expansion techniques developed in her earlier work [2]. With their application, Ballesteros obtained an average precision for transitive translation at 67% of the monolingual performance in the target language, which compared favourably but was still below, the 79% monolingual performance obtained from a direct translation.

This paper presents an approach to reducing error, which appears to better Ballesteros's work, by combining translations from two different transitive routes, a process known as *lexical triangulation.* To illustrate, we imagine a situation where no direct translation route between German and English exists. However, translations via both Spanish and Dutch are possible (Figure 1). If translating the German word "fisch", a Spanish translation dictionary suggests two terms "pez, pescado" and the Dutch gives "vis". Taking each of these in turn, translating the Spanish terms to English gives "pitch, fish, tar, food fish", while Dutch to English gives "pisces the fishes, pisces, fish". Each of the transitive translations introduced much translation error largely due to word sense ambiguity. If we take the term that is in common from the two transitive translations, we have "fish", a good unambiguous translation of the original German word.

Sigir'01, September 9-12, 2001, New Orleans, Louisiana, USA.

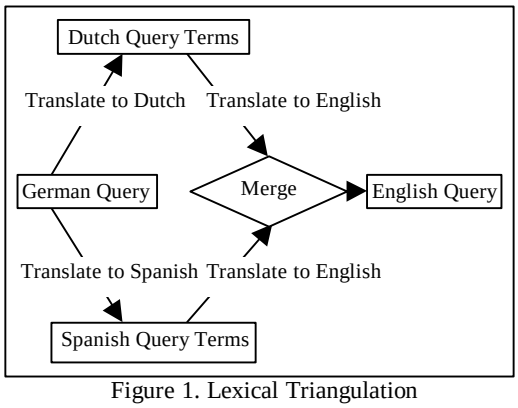

Figure 1. Lexical Triangulation

In the rest of this paper, we discus methodology and resource issues in setting up the experiment, followed by a description of its execution and an analysis of the results. Finally, we discuss the results further, conclude, and look forward to future work.

2. Methodology and Resources

Here, we discuss the components and resources needed in the experiments of the paper.

The availability of test collection and translation resources was the overriding factor determining our choice of languages. The test written collection was from TREC-8 composed of English documents and queries in a number of European languages. The translation resource was EuroWordNet, a multilingual thesaurus consisting of WordNets for various European languages including those used in TREC CLIR queries [20]. Examination of it suggested that the best choice of query language was German, as its vocabulary coverage in EuroWordNet was reasonable. Further examination indicated that Dutch, Spanish, and Italian were good choices as pivot languages since they offered the next best coverage in EuroWordNet. Thus, the collection used for this investigation was the English corpus from the TREC8 CLIR Track and the 28 German and English queries from the same track for which relevance judgements are available.

2.1 Language Resources and Query Processing

Given that the intention of this work was to show how successful a system could be using minimal language resources, the choice of EuroWordNet may appear a little strange. 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 [14]. It is a sophisticated multi-language resource with its own interlingua. However, for the purposes of the experiments described here, it was treated as a series of simple bilingual dictionaries[1]. EuroWordNet has a small phrase vocabulary, which we anticipated would reduce the effectiveness of our CLIR system. However, as the translation resource is constant across the experiments in the paper, we were confident this would not affect the comparison of triangulation to other CLIR techniques. We return to the issue of vocabulary coverage later in the paper.

2.1.1 Lemmatisation, Translation and other processing

We now describe the processing sequence used on queries in these experiments. Converting the TREC topics into queries, the title, description, and narrative fields were used. All characters were reduced to the lowercase 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. Within any natural language, a word may have different inflected forms without significantly changing its core meaning. This presented a potential issue for a resource such as EuroWordNet. If such a database were to include these multiple forms, it would significantly increase its size without adding much value. The solution adopted was to reduce each word to a canonical form known as a *lemma* ([14], [20]). In order that terms in the German queries were matched with those in EuroWordNet, we *lemmatised* the query terms. The approach chosen was to use a pre-existing language resource to support a table driven process. We identified the CELEX database [16] as the best available resource for this purpose. We augmented this table driven approach by checking if the query term was already in a form that matched with EuroWordNet. In this special case, no further reduction or compound splitting (see below) occurred.

We chose to split German compound words to increase further the chances of finding translations in EuroWordNet. The algorithm looked for any sequence of word forms in CELEX that matched with the whole compound. If the algorithm found such a complete match, it returned the corresponding lemma. The algorithm took account of the use of "s" as "glue" in the construction of German compounds. The approach was based on the description of the word reduction module used successfully by Sheridan and Ballerini[18].

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 list contained all of the German words marked as articles, pronouns, prepositions, conjunctions or interjections in the CELEX database. The translation process used tables derived from EuroWordNet to translate between two languages. In most cases, 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. In later experiments, we also used Italian as a pivot. In all cases, translation was carried out on a term-by-term basis. Any German lemmas with no translation passed through the process unaltered in the hope that they were cognates[2] requiring no translation.

The results of the Dutch and Spanish transitive translations were merged to produce the Dutch - Spanish triangulated translation. The merge process was conducted on an original German lemma-by-lemma basis. We examined two merging schemes, referred to as *strict* and *liberal*. In the strict scheme, only translations common to both routes were used to translate the lemma, in the absence of common translations, the original German lemma was passed through unaltered. Strict is analogous to set intersection. In the liberal scheme, again translations common to both routes

[1] Using the eq_synonym and eq_near_synonym relations to find translations for terms.

[2] Words in one language that have similar spelling and meaning in another.

were preferred, however in the absence of a common translation the terms from both routes were used. Again, if no translation was available on either route, the original German lemma passed through unchanged. Liberal is akin to**if** intersection $=\varnothing$ **then** union **else** intersection.

In our later experiments, which included the third (Italian) transitive translation, the three possible pair-wise triangulations were created. That is between Dutch and Spanish transitive (as above), Dutch and Italian transitive, and Spanish and Italian transitive. In the later experiments, comparing the results of all three transitive translations, we created a final "three-way" triangulated translation.

Again, both strict and liberal merging were examined. In the three-way strict case, translations common to all three routes were used in preference to translations common to any pair of routes. In the three-way liberal case, we extended the three-way strict case so that translations appearing on only one of the transitive routes were used in the absence of any common translations. In both cases original German lemmas with no translations passed through the merge process unaltered.

We used an in-house *tf·idf* experimental ranked retrieval system throughout the experiments.

3. Experiment

3.1 The basic experiment

The aim of our first phase of experiments was to test the basic hypothesis of the investigation. Can lexical triangulation reduce the translation error introduced by transitive translation and thus improve retrieval performance? We examined both the strict and liberal merge strategies in separate experiments. Six sets of queries made up the experiment:

1. A control set of TREC8 English queries passed directly to the retrieval system provided a target performance to compare against each CLIR result, referred to as the *monolingual* run.
2. A set of queries generated by translating the German queries to English directly. This *direct* run served as target performance against which to compare the transitive and lexical triangulation runs.
3. A set of queries generated using the triangulation scheme described just above and the strict merging process.
4. Lexical triangulation using liberal merging.
5. A set of queries generated from the transitive translation via Spanish: the *Spanish transitive* run.
6. A set of queries generated from the transitive translation via Dutch: the *Dutch transitive* run.

A modified version of trec_eval (which implements significance tests) was used to compute non-interpolated average precision for all runs. Table 1 summarises the results.

Table 1 Basic Experiment Results

	Av. Prec.	% Below Mono.	% Below Direct
Monolingual	0.289	0	N/A
Direct	0.0549	81**	0
Strict Triangulated	0.0436	85**	21*
Liberal Triangulated	0.0403	86**	27*
Spanish Transitive	0.0106	96**	81**
Dutch Transitive	0.0044	98**	92**

Throughout this work when we report statistical significance, we refer to the Wilcoxon test as implemented in trec_eval. In tables results significant at the 0.05 level are indicated by * and at the 0.01 level by **.

In previous work on direct word-for-word translation, Ballesteros and Croft [1] reported CLIR effectiveness 60% below monolingual. Our comparable results for the direct run indicated performance 81% below monolingual. We believe that this difference in performance arises from two factors. The first is the relatively poor vocabulary coverage of the German EuroWordNet; just under half of the test collection query words were translated. The dictionaries used by Ballesteros and Croft were considerably larger. The other factor concerns the ability to choose the most common sense of a word, this was not attempted using EuroWordNet and resulted in considerable erroneous translations. However, the relatively poor performance of the translation component of our test CLIR system was not a major concern to us, as it remained a constant throughout our experiments. Our prime interest was in comparing the direct, transitive and triangulated translation results against each other. Consequently, we do not repeat the monolingual result in the rest of this paper.

Comparison between the cross language runs revealed that despite the poor performance of both the transitive runs, the triangulated run improved matters considerably. The strict triangulated run improved performance (as measured by average precision relative to the direct run) by some 71% against the Dutch transitive run and by 60% against the Spanish run. The improvements over both the transitive runs were significant at the 0.01 level. The resulting triangulated run was within 21% of the direct run. Although the improvement was not as large, the liberal triangulated run is still considerably better than either of the transitive runs. The strict merging strategy outperformed the liberal by 7.5% (relative to the strict), which was significant at the 0.05 level.

3.2 Pre-translation Pseudo-Relevance Feedback

In their original work on dictionary based CLIR, Ballesteros and Croft [1] introduced pseudo-relevance feedback as a pre-translation step. By using a corpus in the source language, the relevance feedback step expanded the query, the translations of which acted to disambiguate the translations of the other query terms. Ballesteros and Croft found that pre-translation pseudo-relevance feedback strengthened the base for the translation and improved precision. However, they also found the effect limited by the tendency to introduce inappropriate translation terms [1].

We decided, given the resources available to us, and the positive results reported by Ballesteros [3] for transitive translation, to

investigate whether pre-translation pseudo-relevance feedback would combine with lexical triangulation to give further improvements in performance.

3.2.1 Approach

In all of the pre-translation pseudo-relevance feedback experiments, we chose the German SDA corpus from the TREC8 CLIR track as it was from the same time-frame as the TREC8 English corpus. The first task was to optimise the parameters used in pseudo relevance feedback. We conducted a number of experiments to determine the optimal settings for the number of feedback terms to select from how many German documents. Measuring average precision of the triangulated runs, the optimal settings were the top 25 terms from the top 50 documents. The only collection available at the time for evaluating the effect of feedback was the main test collection used in all of the other experiments. In such circumstances, there is a clear danger of over training of the feedback system with a consequent over estimate of the effectiveness of the technique. However, we do not believe that optimising based on our test data affected the aim of the experiment: to determine if triangulation and feedback were beneficial in combination. As shown below the improvements occur across all the runs with the direct run improving to a similar extent to the liberal run. We believe while some over training towards triangulation may have occurred the results are still valid. We repeated the six runs of the previous experiment.

Collected in Table 2 are the results of these experiments together with the results from the previous experiment for comparison.

Table 2: Pre-translation Pseudo Relevance Feedback.

	Without Feedback	With Feedback	% Improvement
Direct	0.0549	0.0771	41
Strict	0.0436	0.0718	65
Liberal	0.0403	0.0642	59*
Spanish Transitive	0.0106	0.029	173*
Dutch Transitive	0.0044	0.0154	250

For the strict merging strategy, we observe that although the two transitive runs improved considerably with the introduction of the pseudo-relevance feedback, the improvement provided by triangulation remains. The triangulated run outperforms the Spanish and Dutch transitive runs by 59% and 78% respectively. Both differences are significant at the 0.05 level. The triangulated run comes within 7% of the direct run and the difference is not significant.

For the liberal merging strategy, the picture is similar with the triangulated run outperforming the Spanish and Dutch transitive runs by 55% and 76% respectively (both significant at the 0.05 level). However, in this case the lower result for the triangulated run means that direct outperforms triangulated by 20% (significant at the 0.05 level).

We observe that after pre-translation pseudo relevance feedback the strict merging strategy now outperforms the liberal by over 10% (relative to the strict). This result is not significant.

3.3 Multiple Transitive Translation

We have observed the positive results from the triangulated runs generated by merging the results of two different transitive translations. The investigation moved on to determine if merging the results of three transitive translations (now including Italian) would improve matters. In addition to the cross language runs conducted in the earlier experiments we examined a further 4 runs:

- A *three-way triangulated run.*
- A Dutch and Italian triangulated run *Dutch – Italian triangulated.*
- An Italian and Spanish triangulated run *Spanish – Italian triangulated.*
- An *Italian transitive* run.

Again, we examined both a strict and a liberal merging strategy in these experiments.

Table 3 below shows the average precision of the four triangulated runs using the two different merge strategies with the direct run and Italian transitive run for comparison (results from new runs indicated in bold). The final column shows the percentage change of the strict strategy relative to the liberal. Negative values favour the liberal strategy.

Table 3: Multiple transitive translation results

	Liberal	Strict	% Change
Direct	0.0549	0.0549	0
Three-way triangulated	**0.0558**	**0.038**	**-31**
Dutch – Spanish triangulated	0.0403	0.0436	8*
Dutch – Italian triangulated	**0.0306**	**0.0272**	**-11**
Spanish – Italian triangulated	**0.031**	**0.0357**	**15**
Italian transitive	**0.0026**	**0.0026**	**0**

All three of the triangulated runs offer considerable improvements over their constituent transitive runs and these improvements are all significant at the 0.05 level. The three-way run outperforms all of the two branch triangulated runs in the liberal merge experiment by between 27% and 50% (relative to the three-way run). This result is significant at the 0.05 level. This remains true in the strict experiments for the Dutch – Italian and Spanish – Italian triangulated runs; however, this result is not statistically significant.

In the case of the liberal merge strategy, the three-way triangulated run out-performs even the direct run by some 1.6%. Although the result is not statistically significant, it is perhaps surprising and is discussed further in Section 4.

The results and analysis above indicate the triangulated scheme using three routes produces superior average precision to the two route schemes. The advantage of any triangulated scheme over its constituent transitive routes is also very clear, an observation that reinforces the previous outcomes of the Dutch and Spanish experiments. With the introduction of the third triangulation, the detrimental effects of the translation error introduced by transitive translation (over direct translation) appear to be eliminated.

3.4 Combining Pre-translation Expansion with Multiple Transitive Translation

The final set of experiments attempted to discover if the three-branch triangulated scheme successfully combines with the pseudo relevance feedback, to give even better performance. We conducted two experiments; using the optimum expansion settings

determined by the Dutch – Spanish triangulated run as above. The experiments followed the same pattern as the multiple transitive translation experiments described above.

Table 4 below shows the average precision of the eight runs under the two different merge strategies and the percentage change relative to the runs without feedback.

Table 4: Multiple transitive translation with feedback

	Liberal,	% Change feedback	Strict	% Change feedback
Direct	0.0771	41		
Three-way triangulated	0.0781	40	0.0609	60
Dutch – Spanish triangulated	0.0642	59*	0.0718	65
Dutch – Italian triangulated	0.0529	73*	0.0449	64*
Spanish – Italian triangulated	0.0553	78*	0.0572	60
Spanish Transitive	0.029	173*		
Dutch Transitive	0.0154	250		
Italian Transitive	0.0136	420*		

We observe that, as before, there was a general increase in performance across all of the runs and experiments with the introduction of pre-translation pseudo-relevance feedback. The three-way triangulated run again outperformed the direct run by 1.3%, again not statistically significant.

In both cases, the results indicate that the two techniques of three-way lexical triangulation and pre-translation pseudo-relevance feedback reinforce each other to improve performance. This is in contrast to Ballesteros [3] who reported inconclusive results when applying pseudo-relevance feedback to Spanish-English-French transitive translation.

4. Discussion

The concept behind lexical triangulation is the technique of cancelling out random noise overlaid on a signal by comparing two different sources for the same signal, which have different noise characteristics. Essentially the signal will be that which is common to both paths. In our case, the signal is the set of "concepts" represented by the original query and the noise comes from the omissions and erroneous additional words added by translation. This noise introduces ambiguity. Removing the noise reduces the ambiguity. The results of our experiments support the concept of noise cancellation.

In the related field of spoken document retrieval, Singhal et al [19] used a similar technique to lexical triangulation, of merging the outputs of several different recognisers (which “translate” speech into text), to improve word-recall. McCarley [13] used the combined evidence of retrieval against translated documents and retrieval with translated queries to gain excellent Cross Language retrieval results. Bartell et al [4] used multiple retrieval strategies of monolingual retrieval fusing the results to produce better effectiveness. Such methods are used in metasearch engines [17].

It is reasonable to expect that results may vary depending on the choice of transitive route. Noise cancellation works best if the noise introduced on the two routes is independent and random, suggesting that the pivots should be as dissimilar as possible. Of the three language pairs used in the multiple transitive experiments, we speculated that Spanish and Italian, both romance languages, were more linguistically in common than they were with the Germanic Dutch. Therefore, we expected the best performance from the two pairings of Dutch with Spanish and Dutch with Italian rather than Spanish and Italian. This result was not observed, however. The vocabulary coverage of EuroWordNet varies greatly over these languages and this most likely had the strongest influence on the CLIR system’s effectiveness.

On the question of triangulation possibly improving retrieval effectiveness over direct translation: Ballesteros [3] observed that although transitive translation introduced more ambiguity, the *translation-recall* (which she called percentage of terms recovered) increased from 0.45 to 0.548. She speculated that combining evidence from several transitive translations might increase performance. We observed a similar increase in translation-recall for transitive translation, from some 0.54 to 0.67. Such an increase may explain why the multiple triangulated runs produced a small improvement over direct.

5. Conclusions

Many applications of CLIR rely on large bilingual translation resources for required language pairs. However, research funding by such projects as TIDES[1], indicates that there is a need, within intelligence organisations at least, for CLIR systems using poor translation resources and pivots. Our work strongly suggests that a lexical triangulation approach to transitive translation can have a beneficial effect on retrieval. Lexical triangulation eliminated the difference in retrieval between transitive translated queries and equivalent direct translated queries - a result not previously demonstrated. In addition, triangulating between three different transitive translations was more beneficial than any of the three possible pairs of triangulations were. There was some suggestion in the results that the three-way triangulated queries may have outperformed the direct translation.

In drawing our conclusions, we are conscious of the characteristics of the TREC-8 CLIR collection, in particular the lack of manual runs contributing to the relevance judgement pools and the small size of the pools [6]. However, given the consistency of the results, and the comments of Braschler, Peters, and Schäuble [6] that the ranking of algorithmic variants are usually reliable under such circumstances, we feel our conclusions are reasonable.

Different language pairings do appear to affect the outcomes of triangulated translations, but this effect appears to be more related to the relative size of the language resources used than an inherent property of the languages themselves. It is also clear from the results that the choice of merge strategy in lexical triangulation is important to maximise results but we were unable to demonstrate a single best algorithm for all environments.

[1] http://www.darpa.mil/ito/research/tides/

6. Future work

Our results showed the potential for lexical triangulation to outperform direct translation; in cases where both triangulation and direct translation are possible, combining the multiple translation routes might be beneficial.

The merging process analysed translations on a word-by-word basis. Taking account of the context of other translations in the query could improve the quality of noise reduction. One approach to achieving this is to defer merging until after retrieval has taken place and fuse document rankings instead.

A non-technical issue of use of pivots that must be examined is a study of existing translation resources to determine the range of resources available to researchers and users of CLIR systems. Such a study will help identify good candidate pivot languages.

7. Acknowledgements

We wish to thank Wim Peters of the Computer Science department at Sheffield University, whose knowledge of EuroWordNet and CELEX helped the research immensely.

8. Bibliography

[1] Ballesteros, L. and Croft, B., Dictionary methods for cross-lingual information retrieval. in *Database and Expert Systems Applications. 7th International Conference, DEXA '96 Proceedings*, (Zurich, Switzerland, 1996), Springer-Verlag Berlin, Germany.

[2] Ballesteros, L. and Croft, W.B., Resolving ambiguity for cross-language retrieval. in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Melbourne, Vic., Australia, 1998), ACM Press.

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

[4] Bartell, B.T., Cottrell, G.W. and Belew, R.K., Automatic combination of multiple ranked retrieval systems. in *Proceedings of the 17th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, (1994), ACM Press, 173 - 181.

[5] Braschler, M., Kan, M., Schäuble, P. and Klavans, J.L., The Eurospider Retrieval System and the TREC-8 Cross-Language Track. in *NIST Special Publication 500 - 246: The Eighth Text REtrieval Conference (TREC-8)*, (Gaithersburg, MD., 1999), NIST.

[6] Braschler, M., Peters, C. and Schäuble, P., Cross-Language Information Retrieval (CLIR) Track Overview. in *NIST Special Publication 500 - 246: The Eighth Text REtrieval Conference (TREC-8)*, (Gaithersburg, MD., 1999), NIST.

[7] Fluhr, C., Schmit, D., Ortet, P., Elkateb, F. and Gurtner, K., SPIRIT-W3: A Distributed Cross-Lingual Indexing and Search Engine. in *Inet 97*, (Kuala Lumpur, 1997), The Internet Society.

[8] Franz, M., McCarley, J.S. and Roukos, S., Ad hoc and Multilingual Information Retrieval at IBM. in *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7)*, (Gaithersburg, MD., 1998), NIST.

[9] Gey, F.C., Jiang, H., Chen, A. and Larson, R.R., Manual Queries and Machine Translation in Cross-language Retrieval and Interactive Retrieval with Cheshire II at TREC-7. in *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7)*, (Gaithersburg, MD., 1998), NIST.

[10] Grefenstette, G. (ed.), *Cross Language Information retrieval*. Kulwer Academic Publishers, Norwell, Mass, 1998.

[11] Hiemstra, D. and Kraaij, W., Twenty-One at TREC-7: Ad-hoc and Cross-language track. in *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7)*, (Gaithersburg, MD., 1998), NIST.

[12] Littman, M.L., Jiang, F. and Keim, G.A., Learning a language-independent representation for terms from a partially aligned corpus. in *Machine Learning. Proceedings of the Fifteenth International Conference (ICML'98)*, (Madison, WI, USA, 1998), Morgan Kaufmann Publishers, San Francisco.

[13] McCarley, J.S., Should we Translate the Documents or the Queries in Cross-language Information Retrieval. in *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, (College Park, MD., 1999), 208 - 214.

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

[15] Oard, D.W. and Dorr, B.J. A Survey of Multilingual Text Retrieval, Institute for Advanced Computer Studies and Computer Science Department University of Maryland, College Park, MD, 1996 http://www.clis.umd.edu/dlrg/filter/papers/mlir.ps.

[16] Piepenbrock, R. CELEX, The Dutch Centre for Lexical Information, CELEX, 2000 http://www.kun.nl/celex/.

[17] Selberg, E. and Etzioni, O. The MetaCrawler Architecture for Resource aggregation on the web. *IEE Expert, 12* (1). 8 - 14.

[18] Sheridan, P. and Ballerini, J.P., Experiments in multilingual information retrieval using the SPIDER system. in *Proceedings of the 1996 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 96*, (Zurich, Switz, 1996), ACM Press, 58 - 65.

[19] Singhal, A., Choi, J., Hindle, D. and Pereira, F., AT&T at TREC-6 :SDR Track. in *NIST Special Publication 500-226: The Sixth Text REtrieval Conference (TREC-6)*, (Gaithersburg, MD., 1997), NIST, 227 - 232.

[20] Vossen, P. EuroWordNet Building a multilingual database with wordnets for several European languages., University of Amsterdam, 1999 http://www.hum.uva.nl/~ewn/.