Experiments in Multilingual Information Retrieval using the SPIDER system

Páraic Sheridan, Jean Paul Ballerini
Swiss Federal Institute of Technology (ETH)
CH-8092 Zürich, Switzerland

Abstract

We introduce a new approach to multilingual information retrieval based on the use of thesaurus-based query expansion techniques applied over a collection of comparable multilingual documents. This approach has been built into the SPIDER information retrieval system and has been tested over a large collection of Italian documents. We have shown that the SPIDER system retrieves Italian documents in response to user queries written in German with *better* effectiveness than a baseline system evaluating Italian queries against Italian documents. Although the importantance of the SPIDER stemming algorithm for Italian must be stressed in these results, we have also achieved performance on multilingual retrieval tasks within 32% of the best SPIDER performance on Italian retrieval, by including a relevance feedback loop in the task of multilingual retrieval.

1 Introduction

As far back as the early 1970's Prof. Gerard Salton at Cornell University performed experiments to investigate the feasibility of multilingual information retrieval ([Salton, 1970], [Salton, 1972]); i.e. retrieving documents that may not be written in the same language as the query. He used a manually constructed multilingual thesaurus to assign concept categories to terms, ensuring that equivalent words in the different languages were assigned the same concept categories. These early experiments led to the conclusion that "the effectiveness of the mixed language processing is approximately equivalent to that of the standard process operating within a single language only".

The limiting factor in these experiments however, was that the resources used were constructed wholly manually, thereby making it impossible to scale up to multilingual information retrieval over large heterogeneous document collections. It is only recently, 20 years after the original experiments, that we find the resources becoming available which may enable us to attempt multilingual retrieval with minimal manual intervention.

One development, which has received much attention over the past number of years, and which we feel we may be able to exploit in performing multilingual retrieval, is the use of corpus processing techniques to automatically construct thesaurus-like information structures for use in query expansion (e.g. [Schäuble and Knaus, 1992], [Qiu and Frei, 1993], [Jing and Croft, 1994]). This approach to query expansion generally involves the use of term co-occurence statistics across the document collection to determine a set of document terms related in some sense to the query terms. The user query is then expanded with the document terms most similar to the query terms or query 'concept'. Although to date these query expansion techniques have been used almost exclusively in retrieving English texts, their statistical nature makes them immediately applicable to texts in other languages ([Han et al., 1994]), or even to collections of multilingual documents. More specifically, from the point of view of multilingual retrieval, such techniques applied over a collection of aligned parallel or comparable documents should produce a *translation effect* as a result of query expansion.

The following definitions and distinctions will be useful in describing our approach:

Comparable Documents: "texts which, though composed independently in the respective language communities, have the same communicative function" [Laffling, 1992] (see also [Peters and Picchi, 1996]). To give a concrete example, comparable documents in our news collection are those news documents that report on the same event or topic (in different languages) on the same day, even though they were written independently and are not translations of each other.

Collection of Multilingual documents: A collection of documents where each document contains text in more than one language.

Multilingual document collection: A collection of documents written in different languages, though each individual document may contain text in only *one* language.

Given a comparable multilingual document collection, the comparable documents can be aligned across languages and merged to form a collection of multilingual documents. Alignment may be facilitated by language-independent annotations, dates, or person names (e.g. speaker) associated with each document. This alignment and merging then results in a collection of documents where each document contains text dealing with a particular topic in different languages. Applying thesaurus-based query expansion techniques across this collection of multilingual documents, one finds terms of different languages related to a given topic co-occuring in many documents, therefore resulting in expansion to a multilingual query - the *translation effect* noted above.

In the rest of this paper, we describe our efforts in realising this approach to multilingual retrieval given a collection of German and Italian news stories. In section 2 we describe the use of word normalisation in processing German and Italian texts for retrieval. We introduce our method of automatically constructing corpus-based thesauri in section 3, and we then present an experimental evaluation of our approach to multilingual retrieval in section 4 . We end by discussing our findings and presenting our conclusions.

2 Word Normalisation

Although there has been recently much debate and analysis of the usefulness of word reduction techniques for retrieval of English documents ([Harman, 1987], [Riloff, 1995], [Hull, 1996]), intuition leads us to suspect that word normalisation will lead to much greater improvements in retrieval effectiveness in the morphologically rich and lexically complex European languages we are working with. In fact, recent experiments at ETH Zürich involving retrieval of Italian texts using different variations of word normalisation, compared to retrieval without any word normalisation have demonstrated very substantial improvements when normalisation is used, thereby confirming our intuitions at least for Italian.

We speculate that for effective German retrieval, particular attention must be paid to the widespread use of single-word compounds. It is quite common for a single German word to represent what would be a multiword term in other languages (e.g. *Bruttoinlandprodukt:* Gross National Product). A feature of these compounds is that they are most often compound *nouns* and as such are likely to be good index terms, so it is important to determine the correct analysis.

We suggest therefore that word normalisation represents an important ingredient to successful multilingual retrieval. In line with this, we have developed word reduction modules for each of German and Italian. Because of the different nature of each language, we have taken a different approach to each, as follows:

2.1 Italian word normalisation

Through analysis of a large corpus sample of Italian words, we determined that it would be possible to formulate a rule-based stemming algorithm for Italian, much like the Porter algorithm [Porter, 1980] that is commonly used for English word reduction. Since Italian is much more morphologically rich than English though, the number of rules required for stemming is much greater than the Porter algorithm for English. The version of our Italian stemmer that we find works best consists of 220 normalisation rules. For ease of understanding, we have specified the rules of our Italian stemmer in the same notation as is used in the well-known Porter algorithm. A small sample of the Italian stemming rules is given here in table 1.

We have evaluated this Italian stemming algorithm over our test collection of 93,229 Italian documents with 65 test topics (events). From the topics we generated sets of test queries of varying lengths in order to investigate whether the relationship between retrieval using stemming and retrieval without stemming varies with query length. Retrieval using the stemming algorithm performed consistently between 75% and 130% better than retrieval without stemming, with the larger performance differences being found when queries were long.

RULES (suffix ZIONE)	EXAMPLES
(*s)crizion[ei] → crivere	iscrizione → iscrivere
(*p)osizion[ei] → orre	deposizione → deporre
(*s)elezion[ei] → elezionare	selezione → selezionare
(*e)lezion[ei] → leggere	elezione → eleggere
((*a)∨(*e)∨(*i))zion[ei] → re	interpolazione → interpolare
(*m)ozion[ei] → uovere	promozione → promuovere
(*d)ozion[ei] → ottare	adozione → adottare
((*b)∨(*n)∨(*t))uizion[ei] → uire	intuizione → intuire

Table 1: Sample rules for Italian Stemming

2.2 German word normalisation

An analysis of our German corpus led us to conclude that such a purely rule-based approach, as we used for Italian, would not be practical for German word normalisation. In particular, we could not see a way to analyse German compounds into their constitutent parts without recourse to a lexicon of some sort. We therefore decided to take a completely dictionary-based approach to German word normalisation. We made use of the German lexicon of the CELEX CD-ROM available from the Linguistic Data Consortium [Baayen et al., 1993], which contains roughly 360,000 German wordforms linked to 51,000 stems. Word normalisation is then a simple lookup procedure, retrieving from the dictionary the lemma for a given wordform.

The productive nature of German compounding however, is such that any dictionary is unlikely to have entries covering all possible word combinations. We therefore augmented our German word reduction module so that, in the case that a word was not found in the dictionary, we would then search for substrings of the word to see if we could find a match for the word through a matching series of substrings (allowing for the *s* that is often added in the middle of German compounds). In such a case, the word reduction module would return the stems of each constituent of the word (e.g. *Abendnachrichtensendungen:* Abend Nachricht Sendung). This approach works well on the whole, but one must watch for cases where, for example a foreign language word, is not in the dictionary and gets split into constituents that happen to be German words (e.g. *Washington:* was hing [hängen] Ton).

3 Similarity Thesauri

A similarlity thesaurus is an information structure representing term similarities which reflect domain knowledge of the collection over which the thesaurus is constructed. A similarity thesaurus is constructed based on *how the terms of the collection are indexed by the documents* ([Qiu, 1995]). A similarity thesaurus is therefore best understood by thinking of exchanging the roles of documents and terms in the traditional view of document retrieval. The documents serve as indexing features and the terms represent retrievable items.

A token τ is a specific occurrence of a term in a document. Let T be the set of all tokens representing an occurrence of a term $\varphi_i \in \Phi$ in a document $d_j \in D$. The function

$$\varphi : T \to \Phi, \tau \mapsto \varphi(\tau)$$

maps the set of all tokens, T, to the indexing vocabulary Φ by assigning every token τ the corresponding indexing feature $\varphi(\tau) := \varphi_i$. The function

$$d : T \to D, \tau \mapsto d(\tau)$$

maps T to the document collection, D, by assigning every token τ the corresponding document $d(\tau) := d_j$. Then, the feature frequency

$$ff(\varphi_i, d_j) := |\{\tau \in T \mid \varphi(\tau) = \varphi_i \wedge d(\tau) = d_j\}|$$

denotes the number of occurrences of φ_i in d_j, and the document frequency

$$df(\varphi_i) := |\{d_j \in D \mid \exists \tau \in T : \varphi(\tau) = \varphi_i \wedge d(\tau) = d_j\}|$$

denotes the number of documents containing the feature φ_i at least once. Quite a large number of retrieval methods and adjoint methods, like access structures, are based on the structure

$$< T, \Phi, D; ff, df >$$

consisting of the sets, T, Φ and D and the functions ff and df introduced above. If we *exchange the roles* of documents and indexing features, i.e. if we replace φ_i, d_j, φ, d, Φ, D by d_j, φ_i, d, φ, D, Φ, we obtain the *dual structure* ([Schäuble and Knaus, 1992]):

$$< T, \widehat{\Phi}, \widehat{D}; \widehat{ff}, \widehat{df} >$$

where $\widehat{\Phi} := D$ consists of indexing features (i.e. documents), $\widehat{D} := \Phi$ consists of retrievable items (i.e. terms), and the functions

$$\begin{aligned} \widehat{ff}(d_j, \varphi_i) &= |\{\tau \in T \mid d(\tau) = d_j \wedge \varphi(\tau) = \varphi_i\}| \\ &= ff(\varphi_i, d_j) \\ \widehat{df}(d_j) &= |\{\varphi_i \in \Phi \mid \exists \tau \in T : d(\tau) = d_j \wedge \varphi(\tau) = \varphi_i\}| \\ &\approx \text{length of } d_j \end{aligned}$$

If we use any retrieval method in the dual space, we obtain a *similarity thesaurus* which returns for every term a list of terms ranked in decreasing order of similarity (relevance) to the query term.

For instance, we have used the retrieval method:

$$sim(\varphi_i, \varphi_k) := \sum_{d_j \in \varphi_i \cap \varphi_k} w(\varphi_i, d_j) * w(\varphi_k, d_j)$$

where $w(\varphi_i, d_j)$ is given by

$$\frac{\left(0.5 + 0.5 * \frac{\widehat{ff}(d_j, \varphi_i)}{Max\widehat{ff}(d_j)}\right) * \widehat{idf}(d_j)}{\sqrt{\sum_{d_j \in \varphi_i} \left(\left(0.5 + 0.5 * \frac{\widehat{ff}(d_j, \varphi_i)}{Max\widehat{ff}(d_j)}\right) * \widehat{idf}(d_j)\right)^2}}$$

This formula encapsulates some of our intuitions about the behaviour of similar terms, analogous to assumptions usually made about terms and documents in the usual plane of the dual space (where documents are indexed by terms).

- A short document plays a more important role in determining the meaning of a term than a long document, as a long document is more likely to deal with more than one topic. If two terms co-occur in a short document, the probability that the two terms are similar is greater than if those terms co-occur in a long document. We therefore take the inverse of $\widehat{df}(d_j)$ defined above (which is related to document length).
- The greater the number of occurrences of an indexing feature (document) in an information item (term), the higher the probability that the document contributes to the meaning of the term. We therefore use the definition of $\widehat{ff}(d_j, \varphi_i)$ given above. This should not be taken as an absolute however, rather we should also consider the frequency with which that feature indexes other items, so we include the factor $Max\widehat{ff}(d_j)$ in our computation.

Note that the weight is also cosine normalised so that if a term is described by many different documents then the weight of a document d_j representing this term will be smaller than if d_j was one of only a few indexing features for this term.

4 Evaluation

4.1 Document Collection

Our multilingual document collection was provided by the Schweizerische Depeschen Agentur (SDA), the Swiss news agency. For the experiments reported here we used a sample of documents reporting on news events for the ten months from June 17th 1994 up to April 25th 1995. For each day during this period we have three files, corresponding to the whole day's news reports in French, German and Italian. Although coverage of each of these three languages is included in our work, the experiments reported here involve only the German and Italian parts of the collection. Each daily news file typically contains between 200 and 300 stories. We consider each individual news story as a document in our collection. For example, the ten months of Italian news gives us a collection of 93,229 documents (130MB of text). The documents are relatively short, with an average length of 112 tokens.

It is important to note that the news collection is not at all parallel. Not only are the news stories written independently in each language, but we have also found that different stories are reported in each of the languages each day. For example, the German news files have a concentration on happenings in northern Switzerland, where German is spoken, while the Italian reports have a southern focus and also include many stories local to nothern Italy. It is the case however that major international news items are reported in each language, and usually on the same day. The SDA news corpus also provides us with a method of obtaining a *rough* alignment of comparable or related stories across languages. Each news item in the SDA collection is annotated with a list of (usually 3 or 4) language-independent descriptors from a controlled vocabulary. These descriptors usually describe the location of the news event and give a rough categorisation. An example of some descriptors is given in table 2. The vocabulary of descriptors is made up of about 50 topic descriptors, 25 Swiss placenames, 10 area (continent) codes, plus roughly 250 country codes.

fin	finance	**zh**	Zürich
kul	culture	**be**	Bern
umw	environment	**ge**	Geneva
c1	United States	**c4ger**	Germany
c4	European Union	**c4ire**	Ireland
c7	Africa	**c4ita**	Italy

Table 2: Sample controlled vocabulary descriptors

As part of our extensive investigation of Italian retrieval, we have constructed a test collection consisting of the 93,229 Italian SDA news documents with 65 queries and relevance judgements. In building this test collection we used the fact that we were dealing with news texts, and chose as queries completely unpredicted world events so as to limit the space of documents to be examined for relevance: no news document issued prior to an unpredicted event can be relevant. This approach also implies a very strict definition of relevance. The relevance judgements were done by a native Italian in such a way that only news stories explicitly reporting on the specified event were considered relevant for a given query topic. For the following experiments we used 50 queries consisting of an average of 5 terms per query. The German and Italian queries used in our experiments are included as an appendix.

4.2 Document Alignment

Since our approach to multilingual retrieval relies on using query expansion across a collection of multilingual documents, we had to align related news stories of the SDA collection and merge them into a single collection. There were two keys to alignment in the SDA collection; the date of the news story and the language-independent descriptors assigned to stories. For each day of news, we generated a set of files corresponding to each descriptor assigned to a story on that day. We did this for both German and Italian news. Note that as a result of this process, a single news story can be included in several files (one for each descriptor assigned to that story) and each file may include several news stories (each story on that day with that descriptor). For example, the file *240894.zh* would include all news stories from 24th August 1994 that had been given the descriptor *zh*. Similarly, a news story on that day that had been annotated with the descriptors *c1 mil* would be included into both files *240894.c1* and *240894.mil*.

We consider these date/descriptor files to be a rough alignment of comparable documents of the SDA collection - i.e. the file *240894.mil* from the German documents and *240894.mil* of Italian documents, both contain news stories dealing with military issues of 24th August 1994 and so probably contain some overlap in covering the military events of that day. Since the news collections are independent in each language though, there will also be cases where a story is reported in one language but not in the other.

Our collection of multilingual documents was created by simply merging the date/descriptor files for each language. Whenever there existed a file with the same date and descriptor in both German and Italian, the two files were merged into one. These multilingual files where then treated as individual documents for query expansion. Because of the relatively few descriptors occurring each day, and perhaps because of a large number of stories only being reported in one language, our collections of about 90,000 (each) Italian and German documents merged to create a collection of only 10,293 bi-lingual documents. Since these documents usually contain several news stories from a given day, and contain the aligned stories in two languages, they are considerably longer than the individual Italian documents; an average length of 2770 compared to 112 tokens. This collection of 10,293 bi-lingual documents was used for constructing the similarity thesaurus to be used for query expansion in our experiments.

4.3 Experimental Approach

Since we have already done experiments in Italian retrieval and already have figures for retrieval performance of Italian queries against Italian documents, we decided to use this as our basis for comparison. To introduce the multilingual aspect, we submitted 50 German queries to our Italian subcollection and compared the results to those obtained by submitting equivalent Italian queries.

Our experiments were run using the SPIDER information retrieval system [Schäuble, 1993] which has a client-server architecture. The simplest way to run our experiments was to use two different servers, one giving access to the collection of multilingual documents and one giving access to the Italian documents. Given this architecture, multilingual retrieval proceeds as follows:

1. Read a German query
2. Submit the German query to the multilingual document server for query expansion
3. Receive back an expanded query containing both German and Italian terms similar to the original query
4. Filter the expanded query through an Italian wordlist and extract the first x Italian terms.
5. Submit the Italian query to the Italian document server for query evaluation
6. Receive back a ranked list of Italian documents

Note that we used the filter step since only Italian documents were to be retrieved and this gave us a way to regulate the size of queries submitted to the Italian server.

4.4 Results

Given the experimental approach outlined above, we ran several batches of multilingual experiments with Italian queries of varying lengths produced from the query expansion (x above). We also determined two different baselines against which to compare performance of multilingual retrieval. We can, for example, compare performance of the multilingual SPIDER system against a baseline retrieval system performing Italian retrieval (without our Stemming algorithm). Alternatively we can compare performance of the SPIDER system retrieving Italian documents to Italian queries versus German queries (where our stemming algorithm is also used for Italian retrieval). The results of evaluating both German and Italian queries against our colleciton of 93,229 Italian documents are presented in table 3 and table 4. A comparison between the best multilingual performance, and the performance of Italian queries retrieving Italian documents is given in the following recall-precision graph (figure 1).

Multilingual Retrieval: German Queries		
Total Relevant Documents: 1418		
Italian Query Length	**# Retr.**	**Avg. Prec.**
10	525	0.212
25	**649**	**0.278**
50	638	0.275

Table 3: German Queries on Italian Documents

Italian Retrieval: Italian Queries		
Total Relevant Documents: 1418		
Description	**# Retr.**	**Avg. Prec.**
No Stemming	488	0.231
SPIDER Stemming	898	0.527

Table 4: Italian Queries on Italian Documents

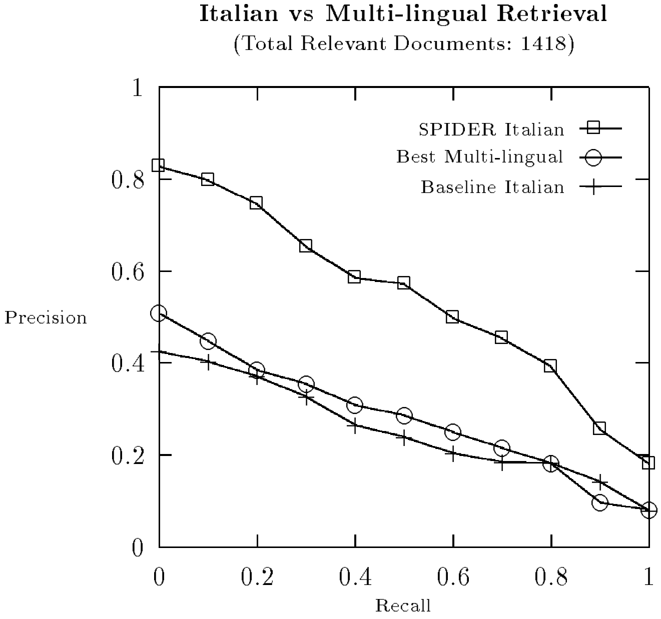

Figure 1: Comparison of Multilingual vs Italian Retrieval: 50 queries over 93,000 Italian documents

Comparing the performace of the best multilingual SPIDER configuration, where German queries (expanded and then filtered to 25 Italian terms) retrieve Italian documents, against plain Italian retrieval without the use of our stemming algorithm, we find that our system performs 23% better than the baseline on the measure of average precision. In this comparison the multilingual SPIDER system also retrieves 33% more relevant documents than the baseline Italian system. If we level the playing field however, and compare retrieval performance of German queries to that of Italian queries, where stemming is used in both cases, then the Italian queries perform significantly better (87%). Note that our stemming algorithm provides a 128% improvement over Italian retrieval without stemming. This is consistent with our earlier experiments on Italian retrieval.

A feature of our multilingual retrieval is that it consistently retrieves more relevant documents than Italian queries, even when the multilingual queries result in lower average precision values. We believe this to be an artifact of the approach we take to multilingual retrieval - query expansion. Query expansion is recognised as a technique to boost recall, since it serves to pad the query with related terms from the document collection and works on the premise that such related terms will help to find more documents relevant to the query. A further, though less desireable, effect of the expansion approach is that terms usually found to be similar to a submitted query tend to be *generally* related, especially when computed similar to the entire query concept. This is likely to provide the explanation as to why, although our multilingual retrieval experiments show good recall, we have difficulty attaining higher levels of precision. This problem is further exacerbated by the nature of our particular test collection. The queries are based on specific world events and relevant documents deal with exactly those events. When query expansion is applied to a German query about an earthquake in Kobe, Japan, we are likely to get similar terms about earthquakes and terms related to events in Japan. When the resulting Italian query is evaluated we are likely to get documents about earthquakes other than the one in Kobe and documents about other news events in Japan, though in both cases these are nonrelevant documents.

4.5 Interactive Multilingual Retrieval

An important consideration in evaluating a multilingual retrieval system should be the role of the user, since a multilingual system is most likely to be used in an interactive setting. If a user is entering German queries but is willing to retrieve Italian documents, then it likely means that the user cannot write well in Italian but has sufficient comprehension to determine if an Italian document is relevant or not. In this scenario, it does not seem unreasonable to expect the user to do a quick survey of the retrieved documents, mark some relevant documents, and perform a single feedback loop. We can then consider this user interaction as another step in performing multilingual retrieval and include it in our evaluation.

We investigated this option by making use of the interactive SPIDER client ([Knaus et al., 1996]), which presents relevant documents to the user with potentially relevant passages highlighted for attention. The user can then perform relevance feedback by marking *passages* as relevant. In the feedback loop, the system then simply expands the original query with all features from the marked relevant passages.

Using the full 50 German queries, multilingual retrieval was performed as before, and the user was presented with a ranked list of 25 document titles. The user was asked to skim the titles presented, to view documents that seemed to be relevant and mark relevant passages, and to perform a feedback loop. We then evaluated performance on the top 100 retrieved documents after feedback and compared this to the performance of the original German query without relevance feedback. The results of this comparison are given in figure 2.

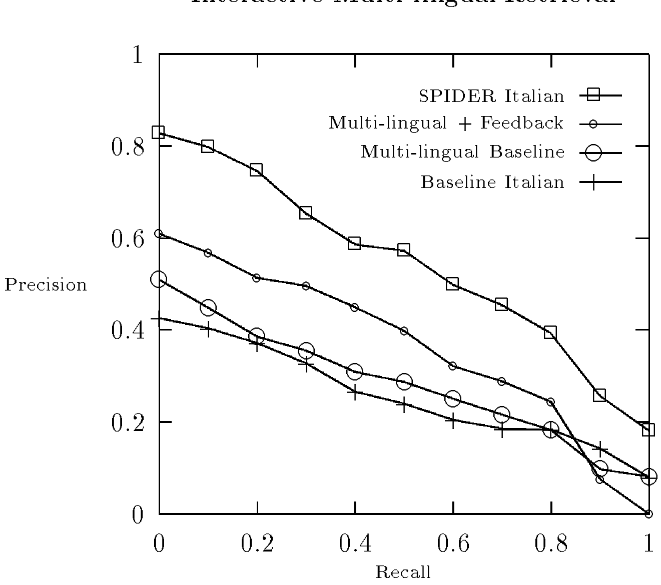

Figure 2: Interactive Multilingual Retrieval: 50 queries over 93,000 Italian documents

This figure shows that the relevance feedback loop has provided a 29% performance increase over the automatic multilingual retrieval without feedback. At this level, multilingual retrieval with a feedback loop is performing 55% better than running Italian queries without our stemming algorithm, and is now performing within 32% of the level of Italian retrieval with full use of stemming.

5 Discussion

We feel that the results presented here demonstrate the usefulness of our approach to multilingual information retrieval. We acknowledge however that we have not yet demonstrated the *general* applicability of the approach. The experiments reported here relate to the task of performing retrieval over a collection of comparable documents. Since the availability of a comparable corpus is central to our approach, the current experimental environment must be considered as an ideal. We must still investigate whether we can demonstrate effective retrieval when the collection we are searching is disjoint from the collection used to build the multilingual similarity thesaurus. Although we can claim that our approach is likely to be equally successful in searching, for example, a comparable collection of Swiss parliament reports (once the similarity thesaurus can be constructed over the collection being searched), we do not yet know how successful we can be when using a similarity thesaurus constructed over Swiss news agency stories to submit German queries to a collection of Italian financial reports, for example.

An important area of further research therefore, is the use of multilingual similarity thesauri constructed off-line over one collection and then used performing multilingual retrieval over another collection. We feel that we can still offer acceptable retrieval performance as long as the similarity thesaurus is built on a collection dealing with the same topic as the collection to be searched. We will have to determine experimentally how well we can perform when searching a large heterogeneous document collection. This is a very important question for future research, to establish the general applicability of this appraoch to multilingual information retrieval.

One source of potential improvement for the performance of our similarity thesauri may be a move from word-based indexing to phrases. If we can identify multi-word phrases and then adjust our similarity thesaurus so that it can identify similar phrases, then we may be able to achieve a much better translation effect. Indeed, working with German may be an advantage in this regard, since our word normalisation module already provides us with many multi-word phrases that occur as compounds (*Bäckerkonditorenmeisterverband*). This may also render our (pre-computed) similarity thesauri more robust when used in different domains, as multi-word terms tend to be much less ambiguous across domains than single words.

The current experiments have served to reinforce our belief in the importance of word normalisation for retrieval in European languages. We have now performed many experiments with Italian retrieval on our SDA collection and, depending on the length of queries, the use of our Italian stemming algorithm consistently provides betwen 75% and 130% improvements over retrieval without stemming. We will therefore continue to place substantial emphasis on achieving good retrieval performance in each individual language as foundation stones to good multilingual retrieval. We are continuing our research into German word normalisation and we have recently started development of a module for French normalisation.

Although it may be tempting to consider our reliance on the availability of comparable or parallel corpora a disadvantage of our approach to multilingual retrieval, it is important to bear in mind that all methods of multilingual retrieval rely on *some* source of information, either a parallel corpus or a transfer lexicon or dictionary. In fact our approach has the advantage over lexicon-based multilingual retrieval that it is inherently 'multi-directional'. Once we index and align our multilingual SDA collection, for example, our query-expansion approach immediately delivers us the ability to query in any of the three languages and retrieve documents in any of the three languages. For example, instead of using the similarity thesaurus module to, 'retrieve for an input query in German the most similar Italian terms (those terms which occur in the Italian part of the SDA collection)', we can just as easily give an Italian query and use the similarity thesaurus to retrieve the most similar German terms. Since the similarity thesaurus module is completely statistically-based, it is unimportant which language is input or output. A lexicon-based approach may need a total of six transfer dictionaries to cover all the possibilities for three languages, depending on the type of translation offered (if the transfer dictionaries were bi-directional then the requirement is reduced to three). Although all the experiments report here involved retrieving Italian documents for German queries, the exact same configuration could have been used to enter Italian queries and retrieve German documents, just by filtering the expanded query through a German wordlist instead of an Italian one. We believe that this multi-directionality is an important advantage of our approach.

Acknowledgements: We would like to acknowledge the input of several of our collegues to this work. We thank Erasmus Peukert for help with German word reduction and providing the German queries, Martin Wechsler for help in configuring the SPIDER system for these experiments and the Schweizerische Depeschen Agentur (SDA) for making large quantities of their news available to us. We also acknowledge the contribution of the Linguistic Data Consortium, not only to our work, but to linguistic research in general, by making dictionaries and corpora available to the research community. Our thanks also go to the SIGIR reviewers who made helpful comments on our first draft of this paper.

References

[Baayen et al., 1993] Baayen, R., Piepenbrock, R., and van Rijn, H. (1993). The CELEX Lexical Database. Technical report, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

[Han et al., 1994] Han, C., Fujii, H., and Croft, W. B. (1994). Automatic Query Expansion for Japanese Text Retrieval. Technical report, Departement of Computer Science, University of Massachusetts, Amherst.

[Harman, 1987] Harman, D. (1987). A Failure Analysis on the Limitations of Suffixing in an Online Environment. In *ACM SIGIR Conference on R&D in Information Retrieval*, pages 102–108.

[Hull, 1996] Hull, D. (1996). Stemming algorithms - a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1):70–84.

[Jing and Croft, 1994] Jing, Y. and Croft, W. B. (1994). An Association Thesaurus for Information Retrieval. Technical Report 94-17, Departement of Computer Science, University of Massachusetts, Amherst.

[Knaus et al., 1996] Knaus, D., Mittendorf, E., Schäuble, P., and Sheridan, P. (1996). Highlighting Relevant Passages for Users of the Interactive SPIDER Retrieval System. In *TREC-4 Proceedings*.

[Laffling, 1992] Laffling, J. (1992). On Constructing a Transfer Dictionary for Man and Machine. *Target*, 4(1):17–31.

[Peters and Picchi, 1996] Peters, C. and Picchi, E. (1996). From Parallel to Comparable Text Corpora. to appear in Proceedings EURALEX '96, Goteborg, 13-18 August 1996.

[Porter, 1980] Porter, M. F. (1980). An Algorithm for Suffix Stripping. *Program*, 14(3):130–137.

[Qiu, 1995] Qiu, Y. (1995). *Automatic Query Expansion Based on a Similarity Thesaurus.* PhD thesis, Swiss Federal Institute of Technology.

[Qiu and Frei, 1993] Qiu, Y. and Frei, H. (1993). Concept Based Query Expansion. In *ACM SIGIR Conference on R&D in Information Retrieval*, pages 160–169.

[Riloff, 1995] Riloff, E. (1995). Little Words Can Make a Big Difference for Text Classification. In *ACM SIGIR Conference on R&D in Information Retrieval*, pages 130–136.

[Salton, 1970] Salton, G. (1970). Automatic Processing of Foreign Language Documents. *Journal of the American Society for Information Science*, pages 187–194.

[Salton, 1972] Salton, G. (1972). Experiments in Multi-Lingual Information Retrieval. Technical Report 72-154, Cornell University, Ithaca, New York.

[Schäuble, 1993] Schäuble, P. (1993). SPIDER: A Multiuser Information Retrieval System for Semistructured and Dynamic Data. In *ACM SIGIR Conference on R&D in Information Retrieval*, pages 318–327.

[Schäuble and Knaus, 1992] Schäuble, P. and Knaus, D. (1992). The Various Roles of Information Structures. In *Gesellschaft für Klassifikation, Annual Meeting.*

50 Italian Queries	50 German Queries
pub terroristi Loughinisland	Pub Loughinisland Nordirland getötet
due donne uccise periferia Algeri	Algerien Frauen Stadtrand umgebracht
Algeri sette italiani sono stati uccisi Jijel (Djendjen)	Italiener Terroranschlag Jijel
Algeri, cinque tecnici russi sono stati uccisi Oued	russische Arbeiter Terroranschlag Algerien
Algeri esplosione auto-bomba parcheggio albergo-ristorante	Vier Tote Restaurant Algier
Buenos Aires esplosione locali Asociacion Mutual Israelita	Zehn Tote jüdische Buenos Aires
Algeri, terroristi sequestrato airbus Air France	Air France Flughafen Algier Bewaffneten gekapert
cubani clandestinamente imbarcazione naufragio	Kubanische Flüchtlinge Schiffbruch vermisst
Panama esplosione aereo	Flugzeug stürzt Panama
Colpo di stato nel Gambia militare	Militär-Aufstand in Gambia
Naziskin devastano Buchenwald danneggiano	Gedenkstätte KZ Buchenwald geschändet
Londra esplosione City ambasciata israeliana	Schwere Explosion israelischen Konsulats London
Madrid autobomba esplode	Autobombenexplosion Madrid
turco ucciso fiamme	Türke Bremen erschlagen angezündet
Algeri terremoto	Erdbeben in Algerien
aereo Marocco precipita	Flugzeug Marokko abgestürzt
uccisi somali ONU	indische UNO-Soldaten Somalia getötet
Pittsburgh precipita aereo	Flugzeugabsturz bei Pittsburgh
affondamento traghetto estone	Estnisches Passagierfähre gesunken
aereo schiantato Casa bianca	Kleinflugzeug Rasen Weissen Hauses abgestürzt
suicidio collettivo membri setta Svizzera occidentale	Kollektiver Sektenselbstmord Westschweiz
A Tel Aviv attentato autobus.	Busexplosion Tel Aviv
maltempo Egitto Durunka (Dronka) deposito petrolio fuoco	Sintflutartige Regenfälle gypten
”Achille Lauro” incendiata	Achille Lauro brennt
affonda traghetto Manila	Fähre Manila gesunken
carolina nord cade aereo	Flugzeugabsturz Nordkarolina
Anversa incendio Switel	Hotelbrand Antwerpen
Marin (NE), Papiliorama fiamme	Tropengarten ”Papiliorama” Raub Flammen
fuoco palazzo trenta piani Toronto	Kanada Feuer dreissigstoeckigem Haus
Terremoti giappone forte scossa	Beben Stärke erschüttert Japan
esplosione tribunale federale Oklahoma City	USA, Explosion Gerichtsgebaeude Oklahoma
Grecia terrorismo esplode bomba ambasciata Arabia Saudita quartiere	Griechenland Terrorismus saudischen Botschaft Bombe explodiert
Giappone intossicati Yokohama gas	Japan Gasvergiftung Yokohama
Colombia tecnici italiani rapiti	Kolumbien, zwei italienische Techniker entfuehrt.
aereo precipitato esploso americano	Acht Tote bei Militärflugzeugabsturz in den USA
suicidio Cheyenne Brando figlia Marlon	Tochter von Marlon Brando erhängte sich Cheyenne
tokyo avvelenamento metropolitana	Giftgasunglück in der U-Bahn von Tokio
danneggiato oleodotto Siberia petrolio	Pipeline Sibirien gebrochen Öl ausgeflossen
quattro feriti attentati dinamitardi corsica	Vier Verletzte bei Bombennacht auf Korsika
uomo spara ufficio postale montclair	Vier Tote Schiesserei Postamt USA Montclair
Italia bambino americano morto tentativo rapina	Italien amerikanisches Kind starb versuchtem Raub
incendio parlamento Belfast	Flammen zerstören alten Parlamentssaal in Belfast
incendio ospedale disabili Lüdenscheid	Lüdenscheid Feuer in Sanatorium - Brandstiftung
Canada fuoco palazzo trenta piani	Kanada, Feuer in dreissigstoeckigem Haus
incendi Burbank USA centro-occidentale	Burbank Bränden Mittleren Westen USA
Islanda valanga nord occidentale	Lawinenunglück nordwestlichen Island
quattro bambini morti incendio Londra	vier Kinder sterben bei Brand in London
Russia vittime incendio albergo Irkutsk	Moskau Hotelbrand in Irkutsk
Turchia ucciso esponente partito sinistra	Tuerkei, Vertreter einer linken Partei ermordet
Cina incendio hotel morti	Tote bei Hotelbrand in China