Using COTS Search Engines and Custom Query Strategies at CLEF

David Nadeau, Mario Jarmasz, Caroline Barrière,
George Foster, and Claude St-Jacques

Language Technologies Research Centre
Interactive Language Technologies Group,
National Research Council of Canada
Gatineau, Québec, Canada, K1A 0R6
{David.Nadeau, Mario.Jarmasz, Caroline.Barriere,
George.Foster, Claude.St-Jacques}@cnrc-nrc.gc.ca

Abstract. This paper presents a system for bilingual information retrieval using commercial off-the-shelf search engines (COTS). Several custom query construction, expansion and translation strategies are compared. We present the experiments and the corresponding results for the CLEF 2004 event.

1 Introduction

In our first participation in the Cross-Language Evaluation Forum (CLEF) we entered the French monolingual task as well as the newcomer French to English bilingual tasks. This report is mainly for the latter task although some experimental results are discussed using data from the former. Our research consists in the use of two commercial off-the-shelf (COTS) search engines which we use to perform boolean queries. These search engines do not allow us to perform weighted queries; we attempt to overcome this weakness by developing innovative query strategies. We test our query construction techniques which vary the ways in which the terms are extracted from the topics. We then experiment with various approaches for querying the search engines by combining the terms using the boolean operators. We briefly explore a query expansion approach based on fuzzy logic. Finally, we investigate three different word-for-word translation methods.

We begin by presenting Copernic Enterprise Search (CES) and AltaVista Enterprise Search (AVES), the two COTS search engines used for the 2004 event. Section 3 describes the query term selection process and section 4 describes the steps for constructing the query, i.e. the manner in which the terms and operators are combined. The subsequent sections discuss the query expansion and translation approaches. We end our discussion by stating our conclusions and future work items.

2 Commercial Off-The-Shelf Search Engines

Two commercial search engines were used for our participation at CLEF. Both offer boolean query syntax rather than weighted queries. We realize that this may be a handicap in CLEF-like competitions. Researchers have found strict binary queries to be limiting (Cöster *et al.*, 2003), and most of the best results from previous years rely on systems where each term in a query can be assigned a weight. UC Berkeley performed very well at CLEF 2003 using such a search engine (Chen, 2002). Yet the availability and quality of commercial search engines make them interesting resources which we feel merit proper investigation.

The first search engine that we use is Copernic Enterprise Search (CES), a system which ranked third in the topic distillation task of the Text Retrieval Conference (TREC) held in 2003 (Craswell *et al.*, 2003). Copernic's ranking is based on term frequency, term adjacency and inclusion of terms in automatically generated document summaries and keywords. It performs stemming using a multilingual algorithm akin to Porter's (1980). Copernic also has the ability to handle meta-data and to take it into consideration when performing its ranking calculations. In our experiments we provided CES with the title meta-data which is found in the *TITLE*, *TI* or *HEADLINE* tags depending on the corpus.

The second search engine used is AltaVista Enterprise Search (AVES) which implements algorithms from the renowned AltaVista company. AVES ranking is based on term frequencies and term adjacency. It performs stemming but the exact algorithm is not documented. Meta-data was not taken into consideration for the searches performed with AVES.

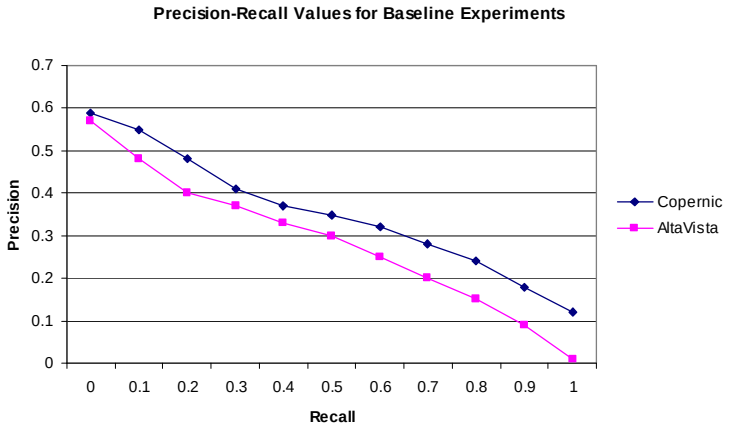

Fig. 1. Precision and recall values for the two search engines using our baseline strategy.

Copernic retrieves more relevant documents than AltaVista for the majority of the configurations which we tested on the 2003 data. This observation holds for the CLEF 2004 data. Figure 1 plots the precision-recall curves for both search engines using the 2004 Monolingual French data. The queries consist of a disjunction of the terms in the topic title. This simple strategy serves as our baseline. Our query strategies are explained in detail in the following sections.

An analysis of the 2003 data allows us to observe that the use of the title metadata, meaning that the search engine assigns a better score to documents in which query terms are found in the title, accounts for about 20% of the difference between the two systems. It is a reasonable assumption that the remaining 80% difference is due to the different ranking algorithms. Since CES and AVES are commercial products, we use them as black boxes and cannot explain the difference in detail. Query Term Selection

3 Query Term Selection

The query term selection step consists in extracting important keywords from the topic. Each topic consists of a title, a description and a narrative field. Here is an example of a French topic:

```
<top>
<num> C201 </num>
<FR-title> Incendies domestiques </FR-title>
<FR-desc> Quelles sont les principales causes d'incen-
die à la maison ? </FR-desc>
<FR-narr> Les documents pertinents devront mentionner
au moins une des causes possibles d'incendie en géné-
ral ou en référence à un exemple particulier. </FR-
narr>
</top>
```

We investigated various methods for exploiting these fields. Our research focused on the following strategies:

S1. the use of the title in isolation (this is our baseline);
S2. the use of the description in isolation;
S3. the use of the combination of the title and the description;
S4. the use of Extractor (Turney, 2000) keyphrases extracted from all fields;
S5. the use of the title plus the best Extractor keyphrases.

In all cases we removed the words *trouvez*, *documents*, *pertinents* and *informations* from the French topics. These words are not stop words but are commonly used in CLEF topics. Stop words are later discarded as explained in the Query Construction section. Comparison of methods can be found in Figure 2.

We have established that it is not efficient to use the narrative in isolation due to the presence of many unrelated words and because the narrative often contains a sentence explaining what not to find, for example *"Les plans de réformes futures ne sont pas pertinents"*. More sophisticated natural language processing techniques are required to take advantage of these explanations.

Using the information contained in the topics, queries can consist of as little as two words, when using the title in isolation, or tens of words, when Extractor is used to select salient terms from the entire topic.

Exhaustive results are given in the next section but it is worth noting some interesting observations. First, the title in isolation performs well, even if it only contains a few words. Titles are indeed made of highly relevant words. All our best runs are obtained using the words in the title. Furthermore, Extractor is useful for selecting pertinent words from the description and narrative parts. The best term selection strategy we found is the use of title words combined with a number of Extractor keyphrases.

Extractor can select noun phrases from a text. In our experiments, a noun phrase containing *n* words is considered as *n* independent words instead of one lexical unit. It would be worthwhile to investigate if any gains can be obtained by searching for exact matches of these multi-word-units.

4 Query Construction

We perform three major tasks when building our queries. (1) First, we remove stop words from the list of terms based on their frequency in the corresponding CLEF corpus. (2) Then, the terms are again sorted based on their frequency in order to create a query where the rarest word comes first. (3) Finally, we combine words using the boolean *AND* and *OR* operators. Some of our search strategies require several variants of the queries to be sent to the search engines. In this scenario, the first query usually returns a small number of documents. Then, a larger number of documents is obtained by appending the results of a second query, and so on.

Let's study the term filtering step in greater detail. First, the words that do not appear in the corpus are removed. Then, we remove terms that occur above a specified threshold. We determined this threshold, as a percentage of the total number of documents. For example, the very frequent French stop word "*le*" appears in about 95% of documents, while the less frequent stop word "*avec*" appears in about 47% of the documents. We trained our system using the 2003 CLEF data and tested it using the 2002 data. Using these corpora we set our threshold for the exclusion of terms in a query at about 25%.

The second step involves sorting the terms according to their frequencies in the corpora, from least frequent to most frequent. This decision is based on the TF-IDF idea (Salton & Buckley, 1988) which states that a rare, infrequent term is more informative that a common, frequent term. These informative terms allow obtaining precise results. Sorting is useful with the strategy described next.

The last step is to issue the query to the search engine. Here we experimented with two variants. The first, which we use as baseline, is a simple disjunction of all terms. The second, which we call *Successive Constraint Relaxation* (SCR) consists in sending successive queries to the search engine starting with a conjunction of all terms and ending with a disjunction of all terms. The constraints, which are represented by the conjunctions, are replaced with disjunctions term by term, starting by the last term, meaning the least informative, in the query. When necessary, a query containing the previously removed terms is issued to obtain a list of 1000 documents for our results. Here's a sample query for which the constraints are successively relaxed, given the following words with their frequency in the corpus: *incendies* (394), *domestiques* (194), *causes* (1694) and *maison* (4651), SCR issues:

Query 1: domestiques AND incendies AND causes AND maison
Query 2: domestiques AND incendies AND (causes OR maison)
Query 3: domestiques AND (incendies OR causes OR maison)
Query 4: domestiques OR incendies OR causes OR maison

On Clef 2004 data, SCR produces 4% more relevant documents than a simple disjunction. Figure 2 shows the results of our query construction strategies using a disjunction of four to eight terms. Figure 3 shows the same experiments but using SCR. The results are plotted using the CLEF 2003 monolingual-French data. Precision is not plotted here, since experiment is conducted using a fixed (1000) number of documents.

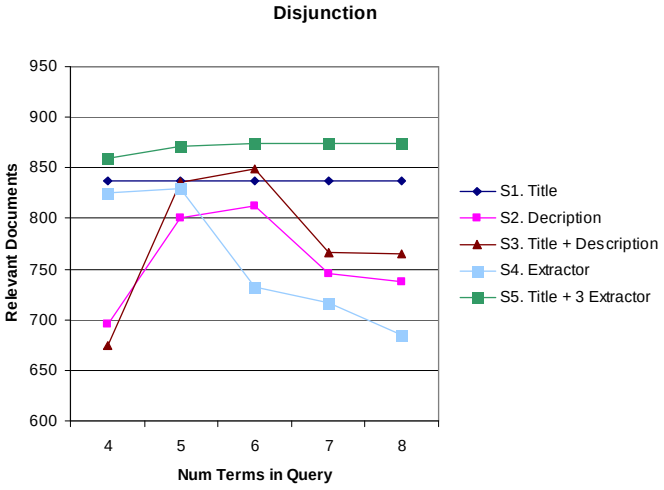

Fig. 2. Results of various term selection strategies using a disjunction.

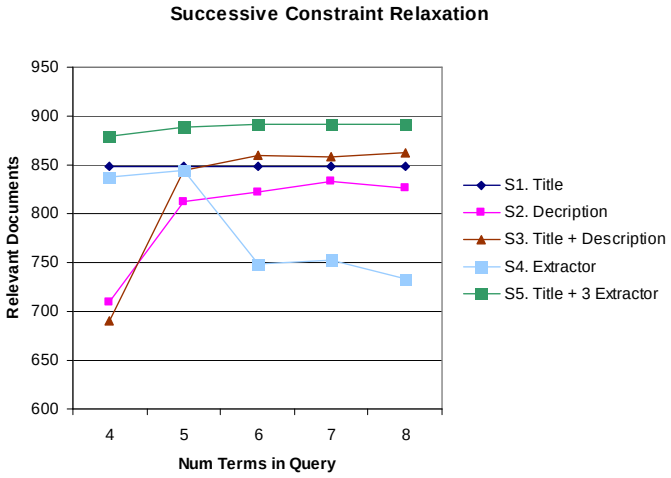

Fig. 3. Results of various term selection strategies using SCR.

We developed our strategies and trained our system using the CLEF 2003 data. We tested all combinations of preceding approaches on the 2002 data. We identified the following methods as being the best query term selection and query construction strategy:

- Use terms from the title plus the three best Extractor keyphrases from the entire topic.
- Remove any words that appear in more than 25% of documents.
- Sort low-frequency first.
- Keep at most 8 terms.
- Issue queries using successive constraint relaxation.

In the bilingual track, all our runs use this combination of strategies.

5 Query Expansion

It has been reported that query expansion using pseudo-relevance feedback generally improves results for Information Retrieval (Buckey and Salton, 1995) and is very effective in CLEF-like settings (Lam-Adesina, 2002). In our experiments, our query expansion strategy relies on a *Pseudo-Thesaurus* construction approach (Miyamoto, 1990) making use of the fuzzy logic operator of *max-min* composition (Klir & Yuan, 1995).

The approach is to take the N-best search engine results (hereafter *N-best corpus*), to extend our initial query with other pertinent words from that corpus as determined

by evaluating their fuzzy similarity to the query words. Texts from *N-best corpus* are segmented into sentences and a term set (*W*) of single words is extracted after filtering prepositions, conjunctions and adverbs from the vocabulary of the DAFLES dictionary (Verlinde et al., 2003). The number of occurrences per sentence for all words is determined. The association between every word pairs is then calculated using the following fuzzy similarity measure:

Let $f(w_{ik})$ be the frequency of the word $w_i \in W$ in the sentence k from the *N-best corpus*.

$$sim(w_i, w_j) = \frac{\sum_k \min\left(f(w_{ik}), f(w_{jk})\right)}{\sum_k \max\left(f(w_{ik}), f(w_{jk})\right)} \quad (1)$$

Among all words, the closest ones to the original query terms were added to our query. We tried adding 1 to 10 terms when building the *N-best corpus* with 5, 10, 25 and 50 documents. This did not improve results, when tested on CLEF 2002 data. The same conclusion holds for 2004.

A possible explanation of the lack of improvement with our query expansion algorithm may be that search engines using only boolean queries may not be able to take advantage of these expanded terms. The extra words added to the queries can be unrelated to the topic, and should have a smaller weight than the initial query terms. The *Pseudo-Thesaurus* gives confidence levels for its expanded list of terms, but we were not able to incorporate this information into our final queries. More investigation is needed to understand why our query expansion attempt failed.

6 Query Translation

A critical part of bilingual information retrieval is the translation of queries, or conversely the translation of the target documents. In our experiments we decided to translate the queries using three different methods. As a baseline we use the free Babel Fish translation service (Babel Fish, 2004). We compare this to (1) an automatic translation method which relies on TERMIUM Plus ® (Termium, 2004), an English-French-Spanish terminological knowledge base which contains more than 3 500 000 terms recommended by the Translation Bureau of Canada and (2) a statistical machine translation technique inspired by IBM Model 1 (Brown *et al.*, 1993), which we call BagTrans. BagTrans has been trained on part of the Europarl Corpus and the Canadian Hansard. The following sections present Termium, BagTrans and, finally, the results obtained by all systems.

The terms stored in Termium are arranged in records, each record containing all the information in the database pertaining to one concept, and each record dealing with only one concept alone (Leonhardt, 2004). Thus the translation task becomes one of word sense disambiguation, where a term must be matched to its most relevant

record; this record in turn offers us standardized and alternative translations. A record contains a list of subject fields and entries, all of which are in English and French, and some of which are in Spanish as well. Entries include the main term, synonyms and abbreviations. The translation procedure attempts to find an overlap between the subject fields of the terms in the query so as to select the correct record of the word which is being translated. If none is found, then the most general term is selected. Generality is determined by the number of times a term appears across all records for a given word, or, if the records themselves do not provide adequate information, generality is determined by the term frequency in a terabyte-sized corpus of unlabeled text (Terra and Clarke, 2003). When a word is not contained in Termium, then its translation is obtained using Babel Fish. More details about the translation procedure using Termium can be found in (Jarmasz and Barrière, 2004).

Given a French word, BagTrans assigns probabilities to individual English words that reflect their likelihood of being the translation of that word and then uses the most probable word in the English query. The probability of an English word *e* is then calculated as the average over all French tokens *f* in the query of the probability p(*e*|*f*) that *e* is the translation of *f*. Translation probabilities p(*e*|*f*) are derived from the standard bag-of-words translation IBM Model 1, and estimated from parallel corpora using the EM algorithm. Two different parallel corpora were used in our experiments: the Europarl corpus, containing approximately 1M sentence pairs and 60M words; and a segment of the Hansard corpus, containing approximately 150,000 sentence pairs and 6.5M words.

Figure 4 shows the precision and recall curves for the three translation techniques as measured using the CLEF 2004 data. The automatic translations strategies with Babel Fish and BagTrans do not perform any word sense disambiguation, whereas the ones using Termium attempt to disambiguate the senses by determining the context from the other terms in the query. Note that Termium found more relevant documents than Babel Fish but its precision-recall curve is lower.

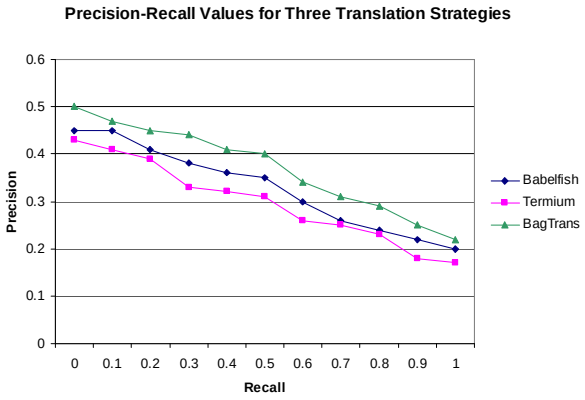

Fig. 4. Comparison of the three translation strategies.

There are many ways in which our Termium and BagTrans translation systems can be improved. None have been customized or trained in particular for this CLEF competition. Since our search engines use boolean operators, an incorrect translation can have a big impact on the results. As we do not take context into consideration when using Babel Fish or BagTrans, it is not surprising that the translations are often incorrect. Termium, on the other hand, is a governmental terminological database and it may contain only specific senses of a word, which might be more correct in some official sense, yet less popular. Termium suffers from being normative. We will continue to pursue automatic machine translation methods which can be trained on specific corpora like BagTrans and which take into account the correct word senses for our future participations at CLEF.

7 Conclusion and Future Work

In our first participation in the Cross-Language Evaluation Forum (CLEF) we participated in the French monolingual and French to English bilingual tasks. We use two COTS search engines and implement various query strategies. The best setup we found consists in creating a query using all words of the topic title plus the 3 best keyphrases of Extractor. We filter stop words based on their frequency in the corpus. Then we sort terms from the rarest to the most frequent. We retrieved documents by issuing successive queries to the search engine, starting with a conjunction of all terms and gradually relaxing constraint by adding disjunction of terms. For the bilingual aspect, BagTrans, a statistical model based on IBM Model 1, yields the best results.

Two main points need more investigation. The first one is our unsuccessful use of the pseudo-relevance feedback. We believe that a strict boolean search engine may be problematic for this kind of algorithm. Indeed, the insertion of only one irrelevant term may lead to irrelevant documents. A weighted query may be the key to smoothen the impact of those terms, especially when our pseudo-relevance feedback algorithm has the ability to output confidence values.

Another pending question is why Termium found many more documents than Babel Fish while the latter present a higher precision-recall curve. We believe it means that Termium did not rank the relevant documents as well as the other strategies did. The explanation, though, remains unclear.

For our next participation, we plan to use a search engine which can perform weighted queries. We'll concentrate on pseudo-relevance feedback, known to be useful at CLEF. We should also add a third language to our translation models to participate in another bilingual track.

8 Acknowledgements

Thanks to Roland Kuhn and Peter Turney for thorough reading and helpful comments.

References

Babel Fish, (2004), *Babel Fish Translation*. http://babelfish.altavista.com/ [Source checked August 2004].

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics, 19(2)*:263-311.

Buckley, C. and Salton, G. (1995), Optimization of relevance feedback weights, *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval*, 351-357.

Chen, A. (2002), *Cross*-Language Retrieval Experiments at CLEF 2002, *CLEF 2002*, *Cross-Language Evaluation Forum*.

Cöster, R., Sahlgren, M. and Karlgren, J. (2003), Selective compound splitting of Swedish queries for Boolean combinations of truncated terms, *CLEF 2003*, *Cross-Language Evaluation Forum*.

Craswell, N., Hawking, D., Wilkinson R. and Wu M. (2003), Overview of the TREC 2003 Web Track, *The Twelfth Text Retrieval Conference, TREC-2003*, Washington, D. C.

Jarmasz, M. and Barrière, C. (2004), *A Terminological Resource and a Terabyte-Sized Corpus for Automatic Keyphrase in Context Translation*, Technical Report, National Research Council of Canada.

Klir, G. J. and Yuan B. (1995), *Fuzzy Sets and Fuzzy Logic*, Prentice Hall: Upper Saddle River, NJ.

Lam-Adesina, A.M., Jones, G.J.H. (2002), Exeter at CLEF 2001: Experiments with Machine Translation for bilingual retrieval, *CLEF 2001, LNCS 2406*, Peters, C., Braschler, M., Gonzalo, J. and Kluck, M. (Eds.), Springer, Germany.

Leonhardt, C. (2004). *Termium ® History*. http://www.termium.gc.ca/site/histo_e.html [Source checked August 2004].

Miyamoto, S. (1990). *Fuzzy Sets in Information Retrieval and Cluster Analysis*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Porter, M. F. (1980). An Algorithm for Suffix Stripping. *Program*, 14(3): 130-127.

Salton, G. and Buckley, C. (1988), Term-weighting approaches in automatic text retrieval, *Information Processing and Management: an International Journal*, 24 (5): 513-523

Termium (2004), *The Government of Canada's Terminology and Linguistic Database*. http://www.termium.com/ [Source checked August 2004].

Terra, E. and Clarke, C.L.A.. (2003), Frequency estimates for statistical word similarity measures. *In Proceedings of the Human Language Technology and North American Chapter of Association of Computational Linguistics Conference 2003* (HLT/NAACL 2003), Edmonton, Canada, 244 – 251.

Turney, P.D. (2000), Learning Algorithms for Keyphrase Extraction, *Information Retrieval*, 2 (4): 303-336.

Verlinde, S., Selva, T. & GRELEP (Groupe de Recherche en Lexicographie Pédagogique) (2003). *Dafles*