Query Translation by Text Categorization

Patrick Ruch
SIM, University Hospital of Geneva
24 Micheli du Crest
1201 Geneva, Switzerland
and
LITH, Swiss Federal Institute of Technology
1015 Lausanne, Switzerland
patrick.ruch@sim.hcuge.ch

Abstract

We report on the development of a cross language information retrieval system, which translates user queries by categorizing these queries into terms listed in a controlled vocabulary. Unlike usual automatic text categorization systems, which rely on data-intensive models induced from large training data, our automatic text categorization tool applies data-independent classifiers: a vector-space engine and a pattern matcher are combined to improve ranking of Medical Subject Headings (MeSH). The categorizer also benefits from the availability of large thesauri, where variants of MeSH terms can be found. For evaluation, we use an English collection of MedLine records: OHSUMED. French OHSUMED queries - translated from the original English queries by domain experts- are mapped into French MeSH terms; then we use the MeSH controlled vocabulary as interlingua to translate French MeSH terms into English MeSH terms, which are finally used to query the OHSUMED document collection. The first part of the study focuses on the text to MeSH categorization task. We use a set of MedLine abstracts as input documents in order to tune the categorization system. The second part compares the performance of a machine translation-based cross language information retrieval (CLIR) system with the categorization-based system: the former results in a CLIR ratio close to 60%, while the latter achieves a ratio above 80%. A final experiment, which combines both approaches, achieves a result above 90%.

1 Introduction

Cross Language Information Retrieval (CLIR) is increasingly relevant as network-based resources become commonplace. In the medical domain, it is of strategic importance in order to fill the gap between clinical records, written in national languages and research reports, massively written in English. There are several ways for handling CLIR. Historically, the most traditional approach to IR in general and to multilingual retrieval in particular, uses a controlled vocabulary for indexing and retrieval. In this approach, a librarian selects for each document a few descriptors taken from a closed list of authorized terms. A good example of such a human indexing is found in the MedLine database, whose records are manually annotated with Medical Subject Headings (MeSH). Ontological relations (synonyms, related terms, narrower terms, broader terms) can be used to help choose the right descriptors, and solve the sense problems of synonyms and homographs. The list of authorized terms and semantic relations between them are contained in a thesaurus. A problem remains, however, since concepts expressed by one single term in one language sometime are expressed by distinct terms in another. We can observe that terminology-based CLIR is a common approach in well-delimited fields for which multilingual thesauri already exist (not only in medicine, but also in the legal domain, energy, etc.) as well as in multinational organizations or countries with several official languages. This controlled vocabulary approach is often associated with Boolean-like engines, and it gives acceptable results but prohibits precise queries that cannot be expressed with these authorized keywords. The two main problems are:

- it can be difficult for users to think in terms of a controlled vocabulary, therefore the use of these systems -like most Boolean-supported engines- is often performed by professionals rather than general users;
- this retrieval method ignores the free-text portions of documents during indexing.

1.1 Translation-based approach

A second approach to multilingual interrogation is to use existing machine translation

(MT) systems to automatically translate the queries (Davis, 1998), or even the entire textual database (Oard and Hackett, 1998) (McCarley, 1999) from one language to another, thereby transforming the CLIR problem into a monolingual information retrieval (MLIR) problem.

This kind of method would be satisfactory if current MT systems did not make errors. A certain amount of syntactic error can be accepted without perturbing results of information retrieval systems, but MT errors in translating concepts can prevent relevant documents, indexed on the missing concepts, from being found. For example, if the word *traitement* in French is translated by *processing* instead of *prescription*, the retrieval process would yield wrong results. This drawback is limited in MT systems that use huge transfer lexicons of noun phrases by taking advantage of frequent collocations to help disambiguation, but in any collection of text, ambiguous nouns will still appear as isolated nouns phrases untouched by this approach.

1.2 Using parallel resources

A third approach receiving increasing attention is to automatically establish associations between queries and documents independent of language differences. Seminal researches were using latent semantic indexing (Dumais et al., 1997). The general strategy when working with parallel or comparable texts is the following: if some documents are translated into a second language, these documents can be observed both in the subspace related to the first language and the subspace related to the second one; using a query expressed in the second language, the most relevant documents in the translated subset are extracted (usually using a cosine measure of proximity). These relevant documents are in turn used to extract close untranslated documents in the subspace of the first language. This approach use implicit dependency links and co-occurrences that better approximate the notion of concept. Such a strategy has been tested with success on the English-French language pair using a sample of the Canadian Parliament bilingual corpus. It is reported that for 92% of the English text documents the closest document returned by the method was its correct French translation. Such an approach presupposes that the sample used for training is representative of the full database, and that sufficient parallel/comparable corpora are available or acquired.

Other approaches are usually based on bilingual dictionaries and terminologies, sometimes combined with parallel corpora. These approaches attempt to infer a word by word transfer function: they typically begin by deriving a translation dictionary, which is then applied to query translation. To synthesize, we can consider that performances of CLIR systems typically range between 60% and 90% of the corresponding monolingual run (Schäuble and Sheridan, 1998). CLIR ratio above 100% have been reported (Xu et al., 2001), however such results were obtained by computing a weak monolingual baseline.

2 Our strategy

Soergel describes a general framework for the use of multilingual thesauri in CLIR (Soergel, 1997), noting that a number of operational European systems employ multilingual thesauri for indexing and searching. However, except for very early work (Salton, 1970), there has been little empirical evaluation of multilingual thesauri in the context of free-text based CLIR, particularly when compared to dictionary and corpus-based methods. This may be due to the expense of constructing multilingual thesauri, but this expense is unlikely to be any more than that of creating bilingual dictionaries or even realistic parallel collections. In fact, it seems that multilingual thesauri can be built quite effectively by merging existing monolingual thesauri, as shown by the current development of the Unified Medical Language System (UMLS[1]).

Our approach to CLIR in MedLine exploit the UMLS resources and its multilingual components. The core technical component of our cross language engine is an automatic text categorizer, which associates a set of MeSH terms to any input text. The experimental design is the following:

1. original English OHSUMED (Hersh et al., 1994) queries have been translated into French queries by domain experts;
2. the OHSUMED document collection is indexed using a standard engine;
3. French queries are mapped to a set of

[1] In our experiments, we used the MeSH as distributed in the 2002 release of the UMLS. See http://umlsks.nlm.nih.gov.

French MeSH terms using an automatic text categorizer;

4. the top-N returned French terms are translated into English MeSH terms, using MeSH *unique identifiers* as interlingua: different values of N terms are tested;
5. these English MeSH terms are concatenated to query the OHSUMED document collection.

2.1 MeSH-driven Text Categorization

Automatic text categorization has been largely studied and has led to an impressive amount of papers. A partial list[2] of machine learning approaches applied to text categorization includes naive Bayes (McCallum and Nigam, 1998), k-nearest neighbors (Yang, 1999), boosting (Schapire and Singer, 2000), and rule-learning algorithms (Apté et al., 1994). However, most of these studies apply text classification to a small set of classes; usually a few hundred, as in the Reuters collection (Hayes and Weinstein, 1990). In comparison, our system is designed to handle large class sets (Ruch et al., 2003): retrieval tools, which are used, are only limited by the size of the inverted file, but 10^{5-6} is still a modest range [3] .

Our approach is data-poor because it only demands a small collection of annotated texts for fine tuning: instead of inducing a complex model using large training data, our categorizer indexes the collection of MeSH terms as if they were documents and then it treats the input as if it was a query to be ranked regarding each MeSH term. The classifier is tuned by using English abstracts and English MeSH terms. Then, we apply the indexing system on the French MeSH to categorize French queries into French MeSH terms. The category set ranges from about 19 936 -if only unique canonic English MeSH terms are taken into account- up to 139 956 -if synonym strings are considered in addition to their canonic class. For evaluating the categorizer, the top 15 returned terms are selected, because it is the average number of MeSH terms per abstract in the OHSUMED collection.

2.2 Collection and Metrics

The mean average precision (noted Av. Prec. in the following tables): is the main measure for evaluating *ad hoc* retrieval tasks (for both monolingual and bilingual runs). Following (Larkey and Croft, 1996), we also use this measure to tune the automatic text categorization system.

Among the 348 566 MedLine citations of the OHSUMED collection[4], we use the 233 445 records provided with an abstract and annotated with MeSH keywords. We tune the categorization system on a small set of OHSUMED abstracts: 1200 randomly selected abstracts were used to select the weighting parameters of the vector space classifier, and the best combination of these parameters with the regular expression-based classifier.

3 Methods

We first present the MeSH categorizer and its tuning, then the query translation system.

3.1 Categorization

In this section, we present the basic classifiers and their combination for the categorization task. Two main modules constitute the skeleton of our system: the regular expression (RegEx) component, and the vector space (VS) component. Each of the basic classifiers implement known approaches to document retrieval. The first tool is based on a regular expression pattern matcher (Manber and Wu, 1994), it is expected to perform well when applied on very short documents such as keywords: MeSH terms do not contains more than 5 tokens. The second classifier is based on a vector space engine[5]. This second tool is expected to provide high recall in contrast with the regular expression-based tool, which should privilege precision. The former component uses tokens as indexing units and can be merged with a thesaurus, while the latter uses stems (Porter). Table 1 shows the results of each

[2] See http://faure.iei.pi.cnr.it/~fabrizio/ for an updated bibliography.

[3] In text categorization based on learning methods, the scalability issue is twofold: it concerns both the ability of these data-driven systems to work with large concept sets, and their ability to learn and generalize regularities for rare events: (Larkey and Croft, 1996) show how the frequency of concepts in the collection is a major parameter for learning systems.

[4] As for queries, we use the corrected version of the OHSUMED queries. For 5 of the 106 OHSUMED queries relevant document sets are not known so only 101 queries were used.

[5] The IR engine, which has used last year for TREC (Ruch et al., 2004), and the automatic categorization toolkit are available on the author's pages: `http://lithwww.epfl.ch/~ruch/softs/softs.html`

System or parameters	Relevant retrieved	Prec. at Rec. = 0	Av. Prec.
RegEx	3986	.7128	.1601
lnc.atn	3838	.7733	.1421
anc.atn	3813	.7733	.1418
ltc.atn	3788	.7198	.1341
ltc.lnn	2946	.7074	.111

Table 1: Categorization results. For the VS engine, *tf.idf* parameters are provided: the first triplet indicates the weighting applied to the "document", i.e. the concept, while the second is for the"query", i.e. the abstract. The total number of relevant terms is 15193.

classifiers.

Regular expressions and MeSH thesaurus. The regular expression search tool is applied on the canonic MeSH collection augmented with the MeSH thesaurus (120 020 synonyms). In this system, string normalization is mainly performed by the MeSH terminological resources when the thesaurus is used. Indeed, the MeSH provides a large set of related terms, which are mapped to a unique MeSH representative in the canonic collection. The related terms gather morpho-syntactic variants, strict synonyms, and a last class of related terms, which mixes up generic and specific terms: for example, *Inhibition* is mapped to *Inhibition (Psychology)*. The system cuts the abstract into 5 token-long phrases and moves the window through the abstract: the edit-distance is computed between each of these 5 token sequence and each MeSH term. Basically, the manually crafted finite-state automata allow two insertions or one deletion within a MeSH term, and ranks the proposed candidate terms based on these basic edit operations: insertion costs 1, while deletion costs 2. The resulting pattern matcher behaves like a term proximity scoring system (Rasolofo and Savoy, 2003), but restricted to a 5 token matching window.

Vector space classifier. The vector space module is based on a general IR engine with *tf.idf*[6] weighting schema. The engine uses a list of 544 stop words.

As for setting the weighting factors, we observed that cosine normalization was especially effective for our task. This is not surprising, considering the fact that cosine normalization performs well when documents have a similar length (Singhal et al., 1996). As for the respective performance of each basic classifiers, table 1 shows that the RegEx system performs better than any *tf.idf* schema used by the VS engine, so the pattern matcher provide better results than the vector space engine for automatic text categorization. However, we also observe in table 1 that the VS system gives better precision at high ranks ($Precision_{at\ Recall=0}$ or *mean reciprocal rank*) than the RegEx system: this difference suggests that merging the classifiers could be a effective. The *idf* factor seems also an important parameter, as shown in table 1, the four best weighting schema use the *idf* factor. This observation suggests that even in a controlled vocabulary, the *idf* factor is able to discriminate between content and non-content bearing features (such as *syndrome* and *disease*).

Classifiers' fusion. The hybrid system combines the regular expression classifier with the vector-space classifier. Unlike (Larkey and Croft, 1996) we do not merge our classifiers by linear combination, because the RegEx module does not return a scoring consistent with the vector space system. Therefore the combination does not use the RegEx's edit distance, and instead it uses the list returned by the vector space module as a *reference* list (RL), while the list returned by the regular expression module is used as *boosting* list (BL), which serves to improve the ranking of terms listed in RL. A third factor takes into account the length of terms: both the number of characters (L_1) and the number of tokens (L_2, with $L_2 > 3$) are computed, so that long and compound terms, which appear in both lists, are favored over single and short terms. We assume that the reference list has good recall, and we do not set any threshold on it. For each concept t listed in the RL, the combined Retrieval Status Value ($cRSV$, equation 1) is:

$$cRSV_t = \begin{cases} RSV_{VS}(t) \cdot Ln(L_1(t) \cdot L_2(t) \cdot k) & \text{if } t \in BL, \\ RSV_{VS}(t) & \text{otherwise.} \end{cases}$$

The value of the k parameter is set empirically. Table 2 shows that the optimal *tf.idf* parameters (*lnc.atn*) for the basic VS classifier does not provide the optimal combination

[6] We use the SMART representation for expressing statistical weighting factors: a formal description can be found in (Ruch, 2002).

Weighting function concepts.abstracts	Relevant retrieved	Prec. at Rec. = 0	Av. Prec.
Hybrids: tf.idf + RegEx			
ltc.lnn	4308	.8884	.1818
lnc.lnn	4301	.8784	.1813
anc.ntn	4184	.8746	.1806
anc.ntn	4184	.8669	.1795
atn.ntn	3763	.9143	.1794

Table 2: Combining VS with RegEx.

with RegEx. Measured by mean average precision, the optimal combination is obtained with *ltc.lnn* settings (.1818) [7], whereas *atn.ntn* maximizes the $Precision_{at\ Recall=0}$ (.9143). For a general purpose system, we prefer to maximize average precision, since this is the only measure that summarizes the performance of the full ordering of concepts, so ltc.lnn factors will be used for the following CLIR experiments.

3.2 Translation

To translate user queries, we transform the English MeSH mapping tool described above, which attributes MeSH terms to English abstracts in a French mapping tool for mapping French OHSUMED queries into French MeSH terms. The English version of the MeSH is simply replaced by the accented French version (Zweigenbaum and Grabar, 2002) of the MeSH. We use the weighting schema and system combination (ltc.lnn + RegEx) as selected in the above experiments, so we assume that the best weighting schema regarding average precision for mapping abstracts to MeSH terms is appropriate for categorizing OHSUMED queries. The only technical differences concern: 1) the thesaural resources, 2) the stemming algorithm. The former are provided by the Unified Medical Lexicon for French consortium (Zweigenbaum et al., 2003) and contains about 20000 French medical lexemes, with synonyms, while the latter is based on Savoy's stemmer (Savoy, 1999).

An additional parameter is used, in order to avoid translating too many irrelevant concepts, we try to take advantage of the concept ranking. Depending on the length of the query, a balance must be found between having a couple of high precision concepts and missing an important one. To evaluate this aspect we do not select the top 15 terms, as in text categorization, but we vary this number and we allow different thresholds: 1, 2, 3, 5, 10, and 25. Finally, by linear regression, we also attempt to determine a linear fit between the length of the query (in byte) and the optimal threshold.

System	Av. precision	CLIR Ratio (%)
MLIR (baseline)	.2406	100
THR-3	.1925	80.0
MT	.1637	59.7
THR-3 + MT	.2209	91.8
THR-F	.1978	82.2

Table 3: Average precision and CLIR ratio.

4 Results and Discussion

Evaluations are computed by retrieving the first 1000 documents for each query. In figure 1, we provide the average precision of each CLIR run depending on the threshold value. The maximum of the average precision is reached when three MeSH terms are selected per query (0.1925), but we can notice that selecting only two terms is as effective (0.19). On the contrary, selecting the unique top returned term is not sufficient (average precision is below 0.145), and adding more than three terms smoothly degrade the precision, so that with 25 terms, precision falls below 0.15. Table 3 compares the results to the baseline, i.e. the score of the monolingual information retrieval system (MLIR). The relative score (CLIR Ratio) of the system which selects only three terms is 80% (THR-3), and should be contrasted with the score obtained by the MT system[8] (59.7%). In the same table, we observe that using a linear function (THR-F), to compute the number of terms to select, results in a very modest improvement as compared to using the best performing static value (82.2% vs. 80%): it means that using a dynamic threshold is not really more effective than translating only the top 3 MeSH concepts. This moderate effectiveness may be due to the fact that OHSUMED queries roughly have a similar length. In contrast, we could expect that querying with very short (one word) and very long queries (querying by documents) could justify the use of a length-dependent threshold.

In a last experiment, we try to combine the two translation strategies: the translation provided by selecting three terms is simply added to the translation provided by the MT system. In table 3, a significant improvement (THR3 +

[7] For the augmented term frequency factor (noted a, which is defined by the function $\alpha + \beta \times (tf/max(tf))$, the value of the parameters is $\alpha = \beta = 0.5$.

[8] The SysTran system was used.

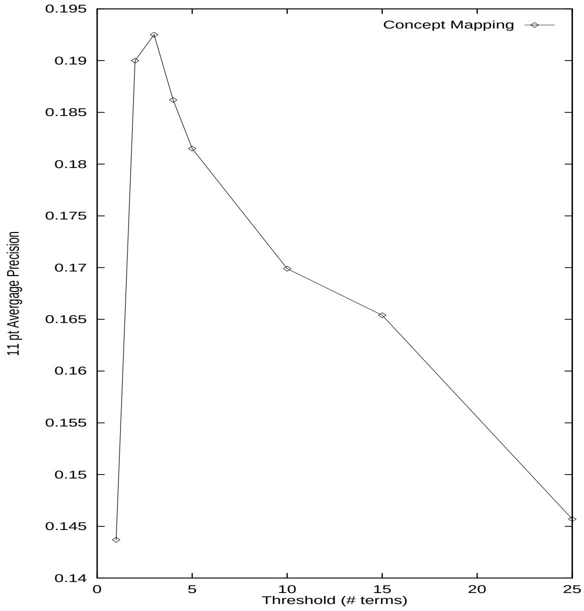

Figure 1: Average precision: different number of terms are translated by concept mapping.

MT = 91.8%) is observed as compared to each single strategies. It seems to confirm that at least some of the words, which are not translated or not properly translated by the text categorizer are well translated by the commercial system.

For example, if we consider a French query such as “anémie - anémie ferriprive, quel examen est le meilleur” (OHSUMED ID = 97: “anemia - iron deficiency anemia, which test is best”), the ranked list of English MeSH term returned by the categorizer is (most similar terms first, with N = 3): *anemia*; *anemia, iron-deficiency*; *anemia, neonatal.* We also observe that an important word like *test* is missing from the list of terms, while on the opposite a less relevant term like *anemia, neonatal* is provided. Now, if we consider the translation supplied by MT, the above query becomes “weaken - weakens ferriprive, which examination is the best”: although this translation is far from perfect, it is interesting to remark that part of the sense expressed by the word *test* in the English query can be somehow found in words such as *examination* and *best.* Further, it is also of interest to notice that most of the erroneously translated content (*weaken* - *ferriprive*) is very unlikely to affect the document retrieval for this query: *ferriprive* as a French word will be ignored, while *weaken* is of marginal content.

Volk et al. (2002) works with a related collection but using German queries, they observe that morphological analysis was effective and report on a CLIR ratio above 80% (MLIR = 0.3543; CLIR = 0.2955). Directly related to our experiments, Eichmann et al. (1998) use the same benchmarks and similar terminological resources, but rely on a word-by-word transfer lexicon constructed from the UMLS. The average precision of their system using French queries is 0.1493, what results in a CLIR ratio of 62% [9]. Because we use the same benchmarks and resources and because our monolingual baselines are quite similar, the methodological difference must be underlined: while Eichmann and al. rely on a word to word transfer lexicon, our system aims at breaking the *bag of word* limitation by translating multiwords terms. Finally, we also observe that the combined system is able to take advantage of existing multilingual vocabulary without assuming any prior terminological knowledge from the user, so that usual problems associated with controlled vocabularies (cf. the introduction) are mutually solved in the proposed architecture.

5 Conclusion and future work

We have presented a cross language information retrieval engine, which capitalizes on the availability of multilingual controlled vocabulary to translate user requests. The system relies on a text categorizer, which maps queries into a set of predefined concepts. The automatic text categorizer is tuned to perform a keyword assignment task before being used to translate French queries into English MeSH terms. For OHSUMED queries, optimal precision is obtained when selecting three MeSH terms, but results are improved when the system is merged with a commercial machine translation system, what suggest that text categorization can be opportunely combined with other query translation approaches. As future investigation, we plan to take into account the retrieval status value obtained by each of the ranked MeSH terms instead of simply setting a threshold on the ranking of the terms.

[9] They report on surprisingly better results (CLIR ration = 71%) for Spanish queries and suggest that French is more difficult to translate than Spanish !

Acknowledgements

The study has been supported by the Swiss National Foundation (Grant 3200-065228). I would like to thank Arnaud Gaudinat for his technical help and Dr. Paul Fabry (MD) for the translation of OHSUMED queries.

References

C Apté, F Damerau, and S Weiss. 1994. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*, 12(3):233–251.

M Davis. 1998. Free resources and advanced alignment for cross-language text retrieval. In *In proceedings of The Sixth Text Retrieval Conference (TREC6)*.

S Dumais, T Letsche, M Littman, and T Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In D Hull and D Oard, editors, *AAAI Symposium on Cross-Language Text and Speech Retrieval*.

D Eichmann, M Ruiz, and P Srinivasan. 1998. Cross-Language Information Retrieval with the UMLS Metathesaurus. pages 72–80.

P Hayes and S Weinstein. 1990. A system for content-based indexing of a database of news stories. *Proceedings of the Second Annual Conference on Innovative Applications of Intelligence*.

W Hersh, C Buckley, T Leone, and D Hickam. 1994. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *SIGIR*, pages 192–201.

L Larkey and W Croft. 1996. Combining classifiers in text categorization. In *SIGIR*, pages 289–297. ACM Press, New York, US.

U Manber and S Wu. 1994. GLIMPSE: A tool to search through entire file systems. In *Proceedings of the USENIX Winter 1994 Technical Conference*, pages 23–32, San Fransisco CA USA, 17-21.

A McCallum and K Nigam. 1998. A comparison of event models for Naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*.

J McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval. *ACL*.

D Oard and P Hackett. 1998. Document translation for cross-language text retrieval at the university of Maryland. In *In Proceedings of The Sixth Text Retrieval Conference (TREC6)*.

Y Rasolofo and J Savoy. 2003. Term proximity scoring for keyword-based retrieval systems. In *ECIR*, pages 101–116.

P Ruch, R Baud, and A Geissbühler. 2003. Learning-Free Text Categorization. *LNAI 2780*, pages 199–208.

P Ruch, C Chichester, G Cohen, G Coray, F Ehrler, H Ghorbel, H Müller, and V Pallotta. 2004. Report on the TREC 2003 Experiment: Genomic Track. In *TREC-12*.

P Ruch. 2002. Using contextual spelling correction to improve retrieval effectiveness in degraded text collections. *COLING 2002*.

G Salton. 1970. Automatic processing of foreign language documents. *JASIS*, 21(3):187–194.

J Savoy. 1999. A stemming procedure and stopword list for general french corpora. *Journal of the American Society for Information Science*, 50(10):944–952.

R Schapire and Y Singer. 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

P Schäuble and P Sheridan. 1998. Cross-language information retrieval (CLIR) track overview. In *In Proceedings of The Sixth Text Retrieval Conference (TREC6)*.

A Singhal, C Buckley, and M Mitra. 1996. Pivoted document length normalization. *ACM-SIGIR*, pages 21–29.

D Soergel. 1997. Multilingual thesauri in cross-language text and speech retrieval. In D Hull and D Oard, editors, *AAAI Symposium on Cross-Language Text and Speech Retrieval*.

M Volk, B Ripplinger, S Vintar, P Buitelaar, D Raileanu, and B Sacaleanu. 2002. Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval. *Int J Med Inf*, 67 (1-3):75–83.

J Xu, A Fraser, and R Weischedel. 2001. Cross-lingual retrieval at bbn. In *TREC*.

Y Yang. 1999. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1:67–88.

P Zweigenbaum and N Grabar. 2002. Restoring accents in unknown biomedical words: application to the french mesh thesaurus. *Int J Med Inf*, pages 113–126.

Pierre Zweigenbaum, Robert Baud, A Burgun, F Namer, E Jarrousse, N Grabar, P Ruch, F Le Duff, B Thirion, and S Darmoni. 2003. Towards a Unified Medical Lexicon for French. *MIE 2003*.