The XLDB Group participation at CLEF 2005 ad hoc task

Nuno Cardoso[†], Leonardo Andrade[†], Alberto Simões[*] and Mário J. Silva[†]
[†]Grupo XLDB - Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
[*]Departamento de Informática, Universidade do Minho
{ncardoso, leonardo, mjs} at xldb.di.fc.ul.pt, ambs at di.uminho.pt

Abstract

This paper presents the 2005 participation of the XLDB Group in the CLEF monolingual and bilingual ad hoc tasks for Portuguese. We participated with an improved and extended configuration of the tumba! search engine software. We detail the new features and evaluate their performance.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Design, Measurement, Experimentation

Keywords

CLEF, ad hoc

1 Introduction

In 2004, the XLDB Group made its debut participation in CLEF, on the monolingual ad hoc Portuguese retrieval task [4]. The main goals were to obtain hands-on experience in joint evaluations of information retrieval (IR) and evaluate tumba!, our web search engine [11], on this task. We learned that we had to come up with new approaches and methods, as the strategy for searching and indexing large web collections has to be different for the kind of document collections used in the CLEF ad hoc task.

This year, we embraced the ad hoc task with the objective of evaluating new methods and algorithms for the task:

- Implementation of new logic operators on query strings, to support expanded queries
- Development of new methods for using all the topic information provided and merging the combined result sets.
- Topics translation for submission of English to Portuguese bilingual runs.

This paper is organized as follows: Section 2 describes our system and enumerates the main changes from last year's configuration. In Section 3, we present our evaluation goals and submitted runs. Section 4 presents the results obtained. Section 5 summarises our conclusions.

2 Improvements

One of the main lessons learned from last year's CLEF ad hoc task participation was that IR in large web collections is quite different from IR on small text collections. Simple adjustments to a web search engine aren't sufficient if we want to use all the information provided for each topic instead of just a few terms to query the CLEF ad hoc collection. This motivated the development of a set of new software, to handle properly the task.

We developed a new query expansion module that generates alternative queries from the descriptions given. This module, called QuerCol (Queries Collator) is external to the core tumba! search engine, but has an essential role in the production of the runs we submitted to CLEF in 2005.

We also improved tumba! on its capability to properly rank poorly linked and tagged documents. To rank the results for CLEF 2005, we developed a simplified version of the Okapi weighting algorithm [9], added support for the 'OR' operator in query strings, and implemented new result set merging algorithms.

With these new modules, our group is now taking the first steps to adopt the basic set of components required for serious participation on in this kind of IR task – robust stemming, weighting scheme and blind feedback [2].

In the remainder of this Section, we detail the design of QuerCol, the newly developed query expansion module, and the improvements made to the query processing sub-system of tumba!.

2.1 Query Expansion

The main conclusion of our CLEF 2004 participation was that, in order to achieve higher recall values, we need to expand the title terms into alternative variants, as collections include many documents relevant to the query topic without all the topic terms [4]. So, this year we created multiple queries for each topic, based on synonyms, morphological and lexical expansion of the title terms, and a selection of other terms from the topic description.

Query strings can now include the 'OR' (disjunction) operator, which wasn't supported by the query server that we had in 2004. This enabled us to make extensive use of synonyms and morphological variations of the title terms. Other systems and former CLEF participants, like David Nateau et al, experimented query expansion modules based on the 'OR' operator [7], and that inspired us to start QuerCol.

QuerCol generates queries from a given topic using the following approach:

1. *Eliminate common stop-words and CLEF-related stop-words*. The latter include terms like 'document' and 'relevant', which are frequent in topic descriptions. We obtain these by selecting the top 5 most frequent terms from all topics.

2. *Obtain title concepts.* After stop-word elimination, we assume that all remaining title words are root-terms of Boolean expressions in the disjunctive normal form, each representing a **concept**, which must be present in all query strings derived from the topic. We used *jspell* to expand morphologically the title concepts [1, 12]. *Jspell* is a morphological analyser based on derivation: words are created applying a set of rules over a root term. This way, it is easy to check the root term and apply rules to create word derivations for each title concept. From these, we only pick those having a frequency of least 5 in the collection.

3. *Obtain expanded concepts*. For each topic title, we take the terms as a conjunction query, which is submitted to the tumba! instance indexing the CLEF ad hoc collection. Then, we measure the *tf x idf* value for each term in the topic's set of words, for each document in the obtained result set. We rank the top 8 terms and discard those with a document frequency lower than 5 in the collection. The selected terms are called **expanded concepts**.

4. *Compute the similarity between the title concepts and the expanded concepts*. For instance, if the title concepts are *shark* and *attack*, and the term *strike* is selected as an expanded concept, we want to relate it to the *attack* concept, to create a query like *shark attack OR shark strike*. We used a database of term co-occurrences of Portuguese terms developed by the Porto node of Linguateca, built from two Portuguese corpora, CETEMPublico[10] and WPT 03 [6]. In the example above, we queried the

database for the top-20 terms that co-occur after the term *shark*. If *strike* is in the result, we can say that the two terms belong to the same concept, and we add *strike* to the *attack* concept term list.

If an expanded concept isn't associated to a concept, it is later added to the query string as a disjunction. This means that expanded concepts don't influence the result set lists, but contribute to weighting the documents containing them.

5. *Query string generation.* In the end, each title concept is defined as a list of terms, selected both from the expanded concepts and from the morphological expansions of the initial title terms. With all the lists of concepts for each topic, we compute all term combinations as a $m \times n$ matrix of m concepts $\times n$ term list size for each concept, and finally we merge them with disjunction operators to generate a single query string.

For the English to Portuguese bilingual ad hoc task, we devised the two following approaches:

1. Using the Babelfish web translation tool [14]. The topic strings were extracted and sent one at a time to the translator and the translations replaced the original topic strings.

2. Using Example Based Machine Translation (EBMT) methods in parallel corpora [13]. The translations were made from a translation memory built from multilingual thesauri freely available on the Internet (EuroVoc, Unesco thesaurus and others). The thesauri have not only simple term entries but also multi-word entries that help in the translation of some word sequences. The translation memory was then used to translate word sequences of the topics file. Words without a corresponding entry in the translation memory were individually translated using Babelfish.

2.2 Weighting and Ranking

Sidra [5] is the indexing and ranking system used in the tumba! web search engine. Sidra provides support for "all the terms" searches, exact phrase queries and field searches that restrict result sets to a specific subdomain or document format. Sidra was primarily designed to rank web documents, as its original ranking function relied mainly in metadata such as links' anchor text, URL strings and page titles. However, it performs poorly when handling document collections with scarce metadata, such as the CLEF ad hoc collection. Sidra does not perform term stemming. The index terms are all the single works, indexed by a full inverted file.

To improve the performance of Sidra on CLEF, we made two major enhancements:

1. Implement a weighting function based on term frequency, to tackle the absence of meta-data.

2. Develop support for disjunction of terms expressions as queries to handle expanded queries created by QuerCol.

As weighting function, we implemented a simplified Okapi BM25 formula, without relevance information [8]. This way, the weighting effect is very similar to a simple *tf*idf* weighting function. The initial idea was to develop a baseline for a future implementation of a full BM 25 schema, which is not yet available.

Query strings submitted to Sidra are no longer interpreted as AND expressions of terms. Term expressions with the logic 'OR' operator can only be accepted in the Disjunctive Normal Form. Given that the Sidra query servers handle each conjunction as a simple query, support for the 'OR' operator consisted in devising strategies for merging the result sets of ranked documents obtained in each sub-query. We used two simple approaches:

Weight Merge: The final result set is obtained by sorting the weights of each result on the combined result set. The final weight of a document present in more than one result set is the sum of the weights of the document in each result set.

Round-Robin Merge: The final result set is generated by sorting the result sets by the weight of the top ranked document in the result set. Then, documents are picked from each result set using a round-robin rule. Documents already picked to the merged result set are ignored.

3 Runs

For the ad hoc task, we submitted 5 runs for the Portuguese monolingual ad hoc task (4 regular runs plus one mandatory run) and 4 for the English to Portuguese bilingual ad hoc task. As we were testing implementations of the 'OR' operator on tumba!, we selected the result set merging methods as a parameter to measure which produced better results. Hence, we applied the Weight Merge algorithm to half the runs plus the mandatory run, and Round Robin Merge to the other half (see Table 1).

Monolingual				
Query	Manual		Automatic	
Fusion	Weight	Round Robin	Weight	Round Robin
Run	XLDBTumba01	XLDBTumba05	XLDBTumba02 XLDBTumba09	XLDBTumba06

Bilingual				
Query	EBMT translation		Babelfish Translation	
Fusion	Weight	Round Robin	Weight	Round Robin
Run	XLDBTumba03	XLDBTumba07	XLDBTumba04	XLDBTumba08

Table 1: Runs submitted to the ad hoc task

In the monolingual task, we created runs XLDBTumba01 and XLDBTumba05 by manually adding all kinds of synonyms and morphological expansions to the queries that seemed reasonable. We used it as a baseline for evaluation against other submitted runs. For runs XLDBTumba02 and XLDBTumba06, QuerCol automatically generated the queries. We aimed at obtaining result sets of the same level of quality as for manually created runs, as QuerCol used the same query creation approach. XLDBTumba09 is a mandatory run, with query strings automatically generated from the topics' title and description fields only.

On the bilingual task, the goal of our participation was to have a preliminary evaluation of the EBMT systems being developed at the Braga node of Linguateca.

4 Results

Run label	Retrieved	Relevant	Ret_rel	Avg. Prec.	R-Prec.	Overall Prec.	Overall Recall
XLDBTumba01	12595	2904	1675	29,0%	34,3%	13,3%	57,7%
XLDBTumba02	5546	2904	986	19,7%	23,2%	17,8%	34,0%
XLDBTumba05	12595	2904	1666	24,0 %	30,6%	13,2%	57,4%
XLDBTumba06	5546	2904	985	18,1%	22,5%	17,8%	34,0%
XLDBTumba03	4875	1991	605	5,8%	8,0%	12,4%	30,4%
XLDBTumba04	6774	2156	299	5,5%	7,4%	4,4%	13,9%
XLDBTumba07	4875	1991	617	4,7%	7,2%	12,6%	31,0%
XLDBTumba08	6774	2156	301	5,3%	7,4%	4,4%	14,0%
XLDBTumba09	6521	2904	989	19,4%	22,9%	15,2%	34,0%

Table 2: Overall results on all runs

Figure 1 and Table 2 show the obtained results. One of our main goals was to compare the two result sets merging strategies, and in the end the Weight merge method outperformed the Round-Robin method. A deeper analysis on the results will provide valuable hints on the result set merging mechanism to implement for disjunctive queries.

Manual query creation (runs 01 and 05) performed better than automatic query creation (runs 02 and 06). Further analysis on the obtained results will also provide good hints for improving QuerCol to narrow

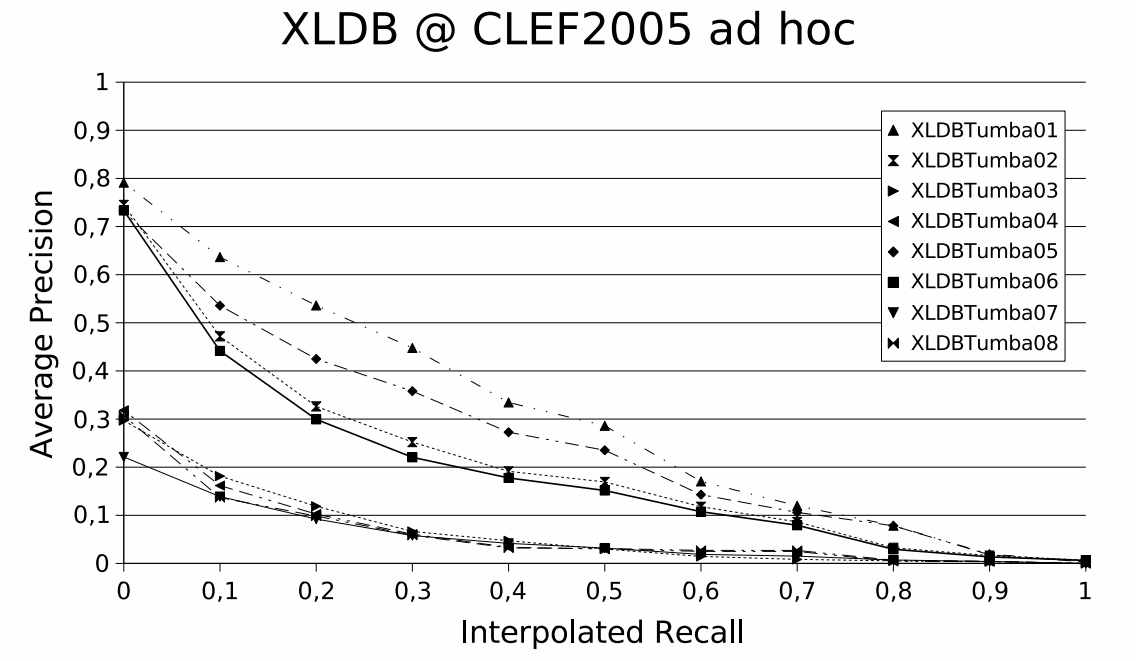

Figure 1: Results of the XLDB Group on monolingual and bilingual ad hoc tasks

the difference.

The results of the monolingual runs are much better than the bilingual. This is likely a consequence of some poor translations. We concluded that we were using thesauri with less quality than expected. As we have overlaps (alternative translations coming from different thesauri), some of the used translations came from the wrong thesaurus and were the source of the bad translation results. Table 2 shows that the runs using EMBT translation obtained more relevant results with less retrieved documents, which is an encouraging result.

The relative performance of the best of our runs compared to other groups' submissions is close to the median. There are a few queries where our performance is much worse than the median for reasons that we have yet to find. However, given that in 2005 our weighting algorithm was very simple, we believe that an improvement here would likely raise the performance level of our software in future evaluations.

5 Conclusion

The results obtained this year represent show a major improvement since last year. This comes as a direct consequence of the changes made to our IR system. Some of the developments for this CLEF task will be incorporated in the next version of tumba!

We have also identified further improvements, like extending QuerCol with a Portuguese stemmer. This would create better term expansions and improve the 'clustering' of terms from the same concept. QuerCol's generated queries also revealed some flaws that we need to amend, as there are concepts with more than one term that shouldn't be handled separately (for instance, *Bill Clinton*). Some morphological expansions of title terms might also produce misleading variations. Finally, we could also incorporate the software developed for our participation in GeoCLEF 2005 to expand geographic names in queries [3].

6 Acknowledgements

We would like to thank to Daniel Gomes, who managed the tumba! repository instance created for supporting our participation in this joint evaluation. Thanks also to the developers of the tumba! search engine and the Portuguese language tools used to assemble the runs. Our participation was partly financed by the Portuguese Fundação para a Ciência e Tecnologia through grants POSI / PLP / 43931 / 2001 (Linguateca) and POSI / SRI / 40193 / 2001 (GREASE).

References

[1] João José Almeida and Ulisses Pinto. Jspell – a module for generic natural language lexical analysis. In *Actas do X Encontro da Associação Portuguesa de Linguística*, pages 1–15, Évora, 1994. in Portuguese. `http://www.di.uminho.pt/~jj/pln/jspell1.ps.gz`.

[2] Martin Braschler and Carol Peters. *Cross-Language Evaluation Forum: Objectives, Results, Achievements.*, volume 7, chapter 1-2, pages 7–31. Kluwer Academic Publishers, January 2004.

[3] Nuno Cardoso, Bruno Martins, Leonardo Andrade, Marcirio Chaves, and Mário J. Silva. The XLDB Group at GeoCLEF 2005. In C. Peters, editor, *Working Notes for the CLEF 2005 Workshop*, Wien, Austria, 21-23 September 2005.

[4] Nuno Cardoso, Mário J. Silva, and Miguel Costa. The XLDB Group at CLEF 2004. In C. Peters, editor, *Working Notes for the CLEF 2004 Workshop*, Bath, UK, 15-17 September 2004.

[5] Miguel Costa. Sidra: a flexible web search system. Master's thesis, Faculdade de Ciências da Universidade de Lisboa, November 2004.

[6] Bruno Martins and Mário J. Silva. A statistical stydy of the wpt-03 corpus. Technical Report DI/FCUL TR-04-1, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa, April 2004.

[7] David Nateau, Mario Jarmasz, Caroline Barrière, George Foster, and Claude St-Jacques. Using COTS Search Engine and Custom Query Strategies at CLEF. In C.Peters, editor, *Working Notes for the CLEF 2004 Workshop*, Bath, UK, 15-17 September 2004.

[8] S. E. Robertson, S. Walker, S. Jones, and M. M. Hancock-Beaulieu. Okapi at TREC-3. In D. K. Harman, editor, *IST Special Publication 500-225: Overview of the Third Text REtrieval Conference (TREC 3)*, pages 109–126, Gaithersburg, MD, USA, 1995. Department of Commerce, National Institute of Standards and Technology.

[9] Stephen E. Robertson, Steve Walker, S. Jones, and Micheline Hancock-Beaulieu M. Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text Retrieval Conference (TREC-3)*, pages 109–126, Springfield, Virginia, USA, 1996.

[10] Paulo Rocha and Diana Santos. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, "Atibaia, São Paulo, Brasil.

[11] Mário J. Silva. The Case for a Portuguese Web Search Engine. In *Proceedings of the IADIS International Conference WWW/Internet 2003, ICWI 2003*, pages 411–418, Algarve, Portugal, 5-8 November 2003. IADIS.

[12] Alberto M. Simões and João José Almeida. Jspell.pm – a morphological analysis module for natural language processing. In *Actas do XVII Encontro da Associação Portuguesa de Linguística*, pages 485–495, Lisbon, 2001. In Portuguese.

[13] Harold Somers. Review article: Example based machine translation. *Machine Translation*, 14(2):113–157, 1999.

[14] BabelFish Web Translation Tool. http://babelfish.altavista.com.