Query Expansion of Zero-Hit Subject Searches: Using a Thesaurus in Conjunction with NLP Techniques

Sarantos Kapidakis[1], Anna Mastora[1], and Manolis Peponakis[2]

[1]Laboratory on Digital Libraries & Electronic Publishing, Archives & Library Science Department, Ionian University, Corfu, Greece
{sarantos, mastora}@ionio.gr
[2]National Hellenic Research Foundation / National Documentation Centre, Athens, Greece
epepo@ekt.gr

Abstract. The focus of our study is zero-hit queries in keyword subject searches and the effort of increasing recall in these cases by reformulating and, then, expanding the initial queries using an external source of knowledge, namely a thesaurus. To this end, the objectives of this study are two-fold. First, we perform the mapping of query terms to the thesaurus terms. Second, we use the matched terms to expand the user's initial query by taking advantage of the thesaurus relations and implementing natural language processing (NLP) techniques. We report on the overall procedure and elaborate on key points and considerations of each step of the process.

Keywords: Query expansion, Thesaurus, Zero-hit queries, NLP techniques

1 Introduction

The focus of our study is zero-hit queries in keyword subject searches in an effort to increase recall by reformulating and, then, expanding the initial queries using an external source of knowledge, namely a thesaurus, and taking advantage of natural language processing (NLP) techniques. In case of zero-hit queries query expansion methods based on sets of retrieved results (implicit relevance feedback) cannot be implemented. Building on this fact, we chose to use a handmade thesaurus to expand the initial queries taking advantage of the relations identified within a thesaurus' structure without letting the users interfere in the process.

In order to proceed to query expansion we first allocate an entry point within the knowledge base, i.e. match the initial queries to a term from the thesaurus. Exact string matching is unlikely to be successful in highly inflectional languages, like Greek, because of the various forms a word can take. Additionally, research has shown that typing errors are also responsible of delivering zero-hit queries [1]. To overcome the identified obstacles we used techniques for natural language processing, namely spelling, lemmatizing, removal of stop words, accent and case processing. The database and the thesaurus underwent the same processing where needed. Finally, we derived candidate expansion terms by considering the related, parallel, narrower and broader terms of the allocated entry point in the thesaurus moving one level towards each direction.

The remaining sections elaborate on the overall procedure and report on key points and considerations of each step of the process.

This is a post print version of an article originally published in:
Kapidakis, S., Mastora, A. & Peponakis, M., 2012. Query Expansion of Zero-Hit Subject Searches: Using a Thesaurus in Conjunction with NLP Techniques. In P. Zaphiris et al., eds. *Theory and Practice of Digital Libraries*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 433–438. Available at: http://dx.doi.org/10.1007/978-3-642-33290-6_48.

2 Aims and Objectives

The aim of the study is to improve the recall of user queries which returned zero hits by expanding them through multi-step implementation of natural language processing techniques and hand-made thesaurus browsing. To this end, the objectives of this study are twofold. First, we perform the matching of query terms to the thesaurus terms. Second, we use the mapped terms to expand the user's initial query terms by taking advantage of the thesaurus relations.

3 Related Work

In [2] is stated that the most critical language issue for retrieval effectiveness is the terms mismatch problem: the indexers and the users often do not use the same words. We say that this is undoubtedly one of the major reasons for a system's poor performance, especially as the outcome of subject searching [3, 4]. Agreeing with [5] we also suggest that the Subject facet is the only facet with semantic relations between its terms, making it the most suitable facet for our method and experimental setup.

In [6] is clarified that *query reformulation* involves either the restructuring of the original query or by adding new terms, while *query expansion* is limited to adding new terms to the original query. Three types of query expansion are discussed in literature: manual, automatic, and interactive (i.e., semiautomatic, user-mediated, or user-assisted). In [7] is stated that thesauri have been recognized as a useful source for enhancing search-term selection for query formulation and expansion. In their study they mention that in 50% of the searches where additional terms were suggested from the thesaurus, users stated that they had not been aware of the terms at the beginning of the search. In [8], about the performance comparison of thesaurus relationships in automatic versus interactive query expansion is concluded that synonyms and narrower terms are good candidates for automatic expansion, while related (associative) terms are better candidates for interactive expansion, leaving the report on broader terms rather equivocal.

In [9] is reported that the improvement in expansion increases when adding up to 20 terms, and reaches a plateau, then the improvement begins to decrease when more than 50 terms are added. The expanded queries, however, still perform better than the original queries. They also indicate that expanding a query with 30 to 40 top ranked terms seems to be the safest method with respect to the collection targeted in their evaluation. Additionally, the user may get confused if the system retrieves documents that do not contain the original query terms. But, using a thesaurus gives the user confidence and security that her needs are met, as reported in [10], while in [7] is mentioned that narrower and related terms taken together constitute approximately 60% of the query-expansion terms selected by users.

The final publication is available at:
http://link.springer.com/chapter/10.1007/978-3-642-33290-6_48

4 Methodology

The data analyzed in this paper was gathered from the log files produced during an in vitro experiment. More details about the experiment scenario can be found in [1].

We have to make clear at this point that the participants did not counteract with the thesaurus at any step of submitting their queries. Moreover, the subjects contained in the database were not formulated according to the thesaurus; they were free keywords provided by the cataloguers. The thesaurus was a tool used for our data post-processing.

4.1 The tools: Thesaurus, Speller and Lemmatizer

For our study we used the EuroVoc[1] thesaurus which is widely-used, multidisciplinary and available in 22 European languages, Greek being among them. We used a licensed version which is available in a variety of formats such as XML and SKOS/XML. The version we used contained 6,797 descriptors and 3,554 non-descriptors. In terms of the relationships identified in the thesaurus, we report on the *Equivalence* (UF and USE) and *Hierarchical* (Broader (BT) and Narrower (NT)).

For spell checking and correcting the queries we used Aspell, v.0.60.6[2]. Aspell is a utility program that connects to the Aspell library so that it can function as an ispell -a replacement, as an independent spell checker, as a test utility to test out Aspell library features, and as a utility for managing dictionaries used by the library. The Aspell library contains an interface allowing other programs direct access to its functions and therefore reducing the complex task of spell checking to simple library calls.

We also used the ilsp_nlp[3] tool, which lemmatizes Greek texts. The input is an XCES document with POS-tagged tokens and the output is an XCES document with lemmas assigned to each token. The tool is developed by the Greek Institute for Language and Speech Processing and is freely available through a web interface.

5 Query Matching and Expanding

The participants submitted 2,116 queries in total while 749 of them returned zero hits, which we further processed. Our first main task was to match the queries to the EuroVoc terms. Then we proceeded to expanding the allocated terms taking advantage of the thesaurus relations. Figure 1 depicts the basic steps of the process.

[1] Reproduced and adapted from the original language editions of the *Eurovoc Thesaurus (Edition 4.3)* © European Communities, 2008. //eurovoc.europa.eu/ (last accessed April 2012)

[2] Aspell, v.0.60.6. ©2000-2004 by Kevin Atkinson, //aspell.net/ (last accessed April 2012)

[3] ilsp_nlp, //nlp.ilsp.gr/soaplab2-axis/ (last accessed April 2012)

The final publication is available at:
http://link.springer.com/chapter/10.1007/978-3-642-33290-6_48

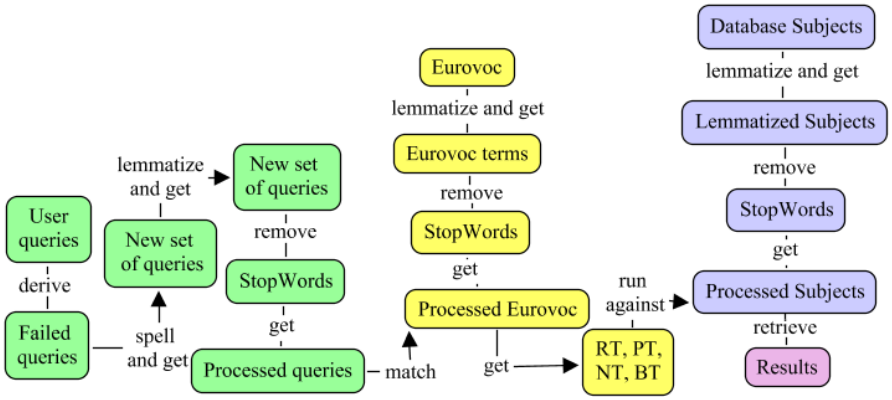

Fig. 1. Overview of the procedure followed

5.1 Matching Queries to EuroVoc Terms

In order to check if the query terms derived from the zero-hit queries were valid words we spell checked and corrected them with Aspell. The tool seems to work adequately, even for named entities, which is often a challenge. A decision we had to make concerned how many correction suggestions we would consider for each identified error. We decided, judging from a selective evaluation of the retrieved results, to use the first suggestion as the most likely to be correct.

During the execution of the search tasks, the participants submitted various forms of the word(s) representing their information need. Let us take for example the case of "Greenhouse effect". Within this nominal compound the word *Greenhouse* stands for "Θερμοκήπιο" *(Thermokēpio)*. We recorded the submission of the following forms of this word: *θερμοκήπιο (thermokēpio)*, *θερμοκηπίου (thermokēpíou)*, *θερμοκηπίων (thermokēpíōn)*, *θερμοκήπια (thermokēpia)*, one misspelled "*θερμοκηπείου*" (thermokēpeíou) and one truncated "*θερμοκήπ**" (thermokēp*). In order to eliminate the variant types of the words due to inflections we lemmatized the words.

A significant issue we also had to deal with was about accents. They play an important role in Greek language but we faced a rather intriguing situation. Most IR systems do not consider them while searching but tools for language processing do consider them for the analysis they perform because a change in the accent position may lead to a change to the word semantics. So, we have to carry accents along the processing and drop them at a convenient point. We followed a similar approach for capitalized words. Finally, we dropped several allocated stop words.

As far as the matching between terms is concerned, the ideal situation is met when all words in the query term match a EuroVoc term. For multi-word EuroVoc terms, this is highly unlikely. In this process, we accept as matching to EuroVoc terms all terms that contain at least one word from the user query. If the EuroVoc term contains additional words, we do not really broaden the query semantics very much, because in order to consider that a EuroVoc term, matching the user query, is contained in a record, all of its words must be present in the record. For example the user submitted the term "organic food" and we located "organic law" and "organic farming" (because of "organic") as well as "food production", "food fat", "pet food" (because of "food"). A record will match this term if it includes at least one EuroVoc term from each of the two sets. Following

The final publication is available at:
http://link.springer.com/chapter/10.1007/978-3-642-33290-6_48

these decisions, our pilot metrics showed that the 1,346 words we imposed to matching with the EuroVoc, 810 of them gave us a positive match.

Truncation in the query terms is also considered during matching, but introduces more problems in the transformation procedures. A truncated word may seem misspelled, and not be lemmatized, but should be left unchanged, as the replacement correct word will in many cases fail to match the words that the original word did.

5.2 Expanding Queries

The second step involves utilizing the thesaurus structure and relations in order to produce the expansion terms. For each matching term, we currently follow all Related, Broader and Narrower relations for one level, and at the end we add the "Use For" relations for all the identified terms. All terms derived this way are accepted alternatives to the original word, and in order to be considered as matching a bibliographic record, all of its words are required to be present in this record. If more than one query words are replaced by a set of EuroVoc terms, at least one term from each set must match the bibliographic record, as well as all words that did not lead to any EuroVoc terms. Again, our pilot metrics show that for 810 mapped words to the thesaurus, we got 44,341 EuroVoc candidate-for-expansion terms in total.

6 Conclusion and Future Considerations

Queries with zero hits are not a fruitful outcome for implementing relevance feedback techniques for query expansion. The use of techniques for language processing can be effective towards improving recall rates when it comes to highly inflectional languages, like Greek. Lemmatizing appears to be a more appropriate process though more computationally intensive. The improvement accomplished so far is shown in Table 1. For each subsequent step, the new set of queries derived from the previous step was re-run to the database. The overall conclusion is that by implementing the proposed framework, 219 (29.07%) of the initial 749 failed queries were dealt with successfully, meaning that they now lead to returning results, instead of zero hits.

Table 1. Improvement during transformation stages after database runs

State of Queries	***Initial***	***Spelled***	***Lemmatized***	***Expanded***
No of zero-hit queries	749	676	624	531

The spell checking and correcting seem to be of considerable assistance towards the improvement of recall but requires extensive work on data interfacing and transformations in order to derive quantitative results, since we had to combine tools (for spelling, lemmatizing, thesaurus browsing and query matching) from various creators having different interfaces and functions.

We do have to consider further options and take more meaningful decisions in selecting candidate expansion terms available in the thesaurus so as to deal with the problem of ambiguity and avoid any semantic drifts caused by arbitrary

The final publication is available at:
http://link.springer.com/chapter/10.1007/978-3-642-33290-6_48

matches. Nominal compounds, named entities and acronyms are also part of our research in terms of properly recognizing them during our processing. Another consideration would be to deal with some exceptional cases, like when replacing a word, in order to avoid missing matches of it, it would be useful to also keep the original word together with the suggested alternatives, as if it was one more alternative. Finally, we plan to measure the exact recall rates at each stage of our method using a test collection.

Acknowledgment. This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II.

References

1. Mastora, A., Kapidakis, S. & Monopoli, M., 2011. Failed Queries: a Morpho-Syntactic Analysis Based on Transaction Log Files. In *First Workshop on Digital Information Management. Corfu* (Greece), pp. 1–7. Available at: http://eprints.rclis.org/handle/10760/15845 [Accessed April, 2012].
2. Carpineto, C. & Romano, G., 2012. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Comput. Surv.*, 44(1), p.1:1–1:50.
3. Lau, E.P. & Goh, D.H.-L., 2006. In Search of Query Patterns: a Case Study of a University OPAC. *Information Processing and Management: an International Journal*, 42(5), pp. 1316–1329.
4. Villén-Rueda, L. et al., 2007. The Use of OPAC in a Large Academic Library: A Transactional Log Analysis Study of Subject Searching. *The Journal of Academic Librarianship*, 33(3), pp. 327-337.
5. Hollink, L., Malaisé, V. & Schreiber, G., 2010. Thesaurus enrichment for query expansion in audiovisual archives. *Multimedia Tools Appl.*, 49(1), pp.235–257.
6. Selvaretnam, B. & Belkhatir, M., 2011. Natural language technology and query expansion: issues, state-of-the-art and perspectives. *Journal of Intelligent Information Systems*. Available at: http://dx.doi.org/10.1007/s10844-011-0174-3/ [Accessed April, 2012].
7. Shiri, A. & Revie, C., 2006. Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. *Journal of the American Society for Information Science and Technology*, 57(4), pp.462–478.
8. Greenberg, J., 2001. Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology. *Journal of the American Society for Information Science and Technology*, 52(6), pp.487-98.
9. Mandala, R., Tokunaga, T. & Tanaka, H., 2000. Query expansion using heterogeneous thesauri. Information Processing & Management, 36, pp.361–378.
10. Fang, H., 2008. A Re-examination of Query Expansion Using Lexical Resources. *In proceedings of ACL-08: HLT*, p.139–147.