Combining NLP and semantics to support a query expansion system

María G. Buey*
IIS Department, University of Zaragoza, Zaragoza, Spain
Angel L. Garrido†
Heraldo de Aragón, Zaragoza, Spain

1 Introduction

Search systems in different contexts are adopting keyword-based interfaces due to their success in traditional web search engines, such as Google or Bing. These systems are usually based on inverted indexes and different ranking policies [Baeza-Yates et al. 1999]. In the context of keyword-based search on structured sources [Yu et al. 2009], most approaches retrieve data that exactly match the user keywords and terms indexed. However, users often do not use the same words. So, there exist a term mismatch problem, that reduce the retrieval effectiveness. The main problems are the information lost (low recall) and the retrieval of non-relevant data (low precision). This problem has recently received a great deal of attention and several approaches have been proposed to solve it [Carpineto and Romano 2012] that are applicable in *Automatic Query Expansion (AQE)* solutions: interactive query refinement, relevance feedback, word sense disambiguation, and search results clustering. One of the most natural and successful techniques is to expand the original query with other words that best capture the actual user intent, or that simply produce a more useful query that is more likely to retrieve relevant documents. In the last years, basic techniques are being used in conjunction with other mechanisms to increase their effectiveness, including the combination of methods, more active selection of information sources, and discriminative policies. AQE is currently considered a promising technique to improve the retrieval effectiveness of document ranking and it is being adopted in commercial applications, especially for desktop and intranet searches. However, very little work has been done to review such studies.

Under these circumstances, we developed a library called SQX-Lib (Semantic Query eXpansion Library) that exploits linguistic and semantic techniques to improve the quality of the keyword-based search process on the relational repositories of the Heraldo Media Group[1]. SQX-Lib is focused on automatically and semantically expanding the scope of the searches, and fine-tune them by taking advantage of the Named Entities (NE) present in the query string. For this purpose, it uses lexical resources and dictionaries, and an unsupervised disambiguation method that looks for the accurate meaning of a word.

2 System overview

The library expands and enriches semantically a given set of user keywords to improve the search process. The semantic expansion process performs three main tasks:

1. *Analysis of the keywords of the query*: SQX-Lib analyzes the keywords introduced as a query. First, it obtains the logical query structure to perform an appropriate construction of the query. It processes the logical operators, parentheses and quotes that could appear in the search string. Then, SQX-Lib performs a morphological analysis of the query string using a NLP tool. Afterwards, the process carries out a second analysis of the query string in order to obtain all the NE. Moreover, it also performs the analysis of the query to obtain exact words, i.e. the words that have been introduced quoted. These words are searched as they have been written in the query string and they are not processed to find their semantics. After this, SQX-Lib removes stop words. Finally, it selects the terms (common names or verbs in our case) that are going to be processed to find their semantics to enrich the input query.
2. *Processing of terms*: SQX-Lib obtains the lemmas of the terms chosen in the previous step using a NLP tool. Then, it searches the sets of synonyms for each term selected, by using the lemmas and semantic resources (dictionaries and lexical thesaurus). In our case, the lexical thesaurus used has been EuroWordNet [Vossen 1998] which is a multilingual resource for several European languages. If the terms have more than one set of synonyms, SQX-Lib performs a disambiguation process to select the most appropriate according to he context, and taking into account an adaptation of the Normalized Google Distance (NGD) [Cilibrasi and Vitanyi 2007].
3. *Expansion of the query*: SQX-Lib reconstructs and expands the query from the relevant data extracted in each of the previous task while keeping the initial logical structure of the query.

3 Conclusions and Future Work

SQX-Lib combines different techniques for information extraction. Whereas there exist many proposals for query expansion, most of them have been applied in experimentation scenarios. Since this is a prototype, designed to give a solution to the particular environment of Heraldo Group, as a future work remains generalizing the solution to enable its evaluation with the standard test packages TREC.

Acknowledgements

Thank you to Raquel Trillo, Sergio Ilarri and HERALDO DE ARAGON S.A. for their support to this work.

References

BAEZA-YATES, R., RIBEIRO-NETO, B., ET AL. 1999. *Modern information retrieval*, vol. 463. ACM press New York.

CARPINETO, C., AND ROMANO, G. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR) 44*, 1, 1.

CILIBRASI, R. L., AND VITANYI, P. M. 2007. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering 19*, 3, 370–383.

VOSSEN, P. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Boston.

YU, J. X., QIN, L., AND CHANG, L. 2009. Keyword search in databases. *Synthesis Lectures on Data Management 1*, 1, 1–155.

*e-mail:mgbuey@unizar.es

†e-mail:algarrido@heraldo.es

[1] http://www.grupoheraldo.com/