Query Expansion using Wikipedia and DBpedia

Nitish Aggarwal and Paul Buitelaar

Unit for Natural Language Processing, Digital Enterprise Research Institute,
National University of Ireland, Galway
firstname.lastname@deri.org

Abstract. In this paper, we describe our query expansion approach submitted for the Semantic Enrichment task in Cultural Heritage in CLEF (CHiC) 2012. Our approach makes use of an external knowledge base such as Wikipedia and DBpedia. It consists of two major steps, concept candidates generation from knowledge bases and the selection of K-best related concepts. For selecting the K-best concepts, we ranked them according to their semantic relatedness with the query. We used Wikipedia-based Explicit Semantic Analysis to calculate the semantic relatedness scores. We evaluate our approach on 25 queries from the CHiC Semantic Enrichment dataset.

Keywords: Query Expansion, Explicit Semantic Analysis, ESA ranking, Wikipedia and DBpedia

1 Introduction

With the enormous amount of information emerging on the Web, the gap between vocabularies used in indexed documents and user queries has been increased. To fill this gap, many query expansion methods such as dictionary-based query expansion [4], and knowledge-based [2] query expansion, have been studied. The query expansion task can be defined by semantic enrichment of a query with its semantically related concepts. With these complementary concepts, additional relevant documents, which may not contain the keywords provided in that query, can be retrieved. For instance, a given query "Hiroshima"can retrieve documents where the keyword Hiroshima directly appears but not the documents, that only contain related concepts such as atomic bomb, Nagasaki or Etajima.

One possible semantic enrichment of a query can be achieved by using the Wikipedia or DBpedia. Wikipedia is a freely available large knowledge resource built by a collaborative effort of voluntary contributors. DBpedia [3] contains a large ontology describing more than 3.5 millions instances extracted from the Wikipedia info-boxes. Also, it is connected to several other linked data repositories on the Semantic Web. Therefore, our approach uses Wikipedia to retrieve the K-best related concepts to the query. We use Wikipedia and DBpedia to

generate the concept candidates, and then rank them according to the semantic relatedness score given by the Wikipedia-based Explicit Semantic Analysis (ESA) [6]. ESA is an approach that calculates the semantic relatedness scores between words or phrases, and uses them to augment ranking functions. Egozi et. al. [5] used the ESA to rank the documents. Our approach takes inspiration from the same to rank all concept candidates.

In this work, we present an approach for semantic enrichment of queries using Wikipedia, DBpedia, and ESA based ranking. The rest of this paper is organized as follows: Section 2 describes our approach in detail; Section 3 explains our four different submitted runs for the semantic enrichment task; Section 4 shows the results; and finally we conclude in Section 5.

2 Approach

Our approach consists of two major steps; concept candidates generation from Wikipedia and DBpedia, and selection of K-best concepts.

2.1 Concept candidate generation

The concept candidates are the titles of Wikipedia articles, which are relevant to the given query. To retrieve these concept candidates, we search all of the Wikipedia articles with the given query, and sort them according to their TFIDF scores. Among the retrieved articles the N best articles are selected as concept candidates. Then, we find all the directly connected Wikipedia articles to the top ranked article from the N selected candidates, in the DBpedia graph. For example, for a given query "Hiroshima", we retrieve 260 different Wikipedia articles, such as "Atomic bombings of Hiroshima and Nagasaki", "Mazda Stadium", and "Nagasaki". With the intuition that the concepts containing similar strings may not provide the additional relevant documents, we exclude those concept candidates, which have a low edit distance to the query.

2.2 ESA ranking

ESA attempts to represent the semantics of the given term in the high distributional semantic space. These semantics are obtained by use of a high dimensional vector, where each dimension reflects a unique Wikipedia concept. This high dimensional vector is created by taking the TF-IDF weight of a given term in the corresponding Wikipedia articles. Semantic relatedness of two given terms can be obtained by calculating the correlation between two high dimensional vectors generated by ESA. We used the ESA implementation described in [1].

We calculate the ESA relatedness score between the query, and each of the concept candidates. Then we select the K-best concepts according to their ESA scores.

3 Experiment

We submitted four different runs in the semantic enrichment task at CHiC. All of these runs are based on the approach described in Section 2. They use different threshold of edit distance to eliminate the concept candidates. The values of N and K are taken as 20 and 10 respectively. Run1 excludes those concept candidates, that contain the query string as a substring. For example, for a given query "Hiroshima", we eliminate all the concept candidates such as "Hiroshima Prefecture", "Hiroshima Station", and "Hiroshima University", as they contain "Hiroshima" as a substring. Run2 and run3 exclude those concept candidates, that have token distance greater than 0.5, and 0.0 respectively. Run4 is the baseline, and does not perform the elimination step.

Run Type	Weak Precision	Strong Precision
DERI_SE1_CLEF-se(Run1)	0.8000	0.6800
DERI_SE2_CLEF-se(Run2)	0.7720	0.5860
DERI_SE3_CLEF-se(Run3)	0.7720	0.6560
DERI_SE4_CLEF-se(Run4)	0.7720	0.6560

Table 1. Weak and strong precision of manual relevance assessment

Run Type	Avg. Precision
DERI_SE1_CLEF-se(Run1)	0.3023
DERI_SE2_CLEF-se(Run2)	0.2395
DERI_SE3_CLEF-se(Run3)	0.2008
DERI_SE4_CLEF-se(Run4)	0.1504

Table 2. Average Mean Precision in Ad-hoc retrieval environment

4 Results

All of these runs are evaluated in two different phases: manually, and by using a query expansion experiment with a standard IR system. All of the suggested concepts are assessed manually for use in an interactive query expansion environment to check if these suggestions make sense with respect to the original query. These manual relevance assessment measures are on a three point scale: definitely relevant, maybe relevant, and not relevant. Table 1 shows the scores of weak precision and strong precision. Strong precision is the average precision over 25 queries of the "definitely relevant" suggestions, and weak precision is the average precision of the "definitely relevant", and "maybe relevant", over all

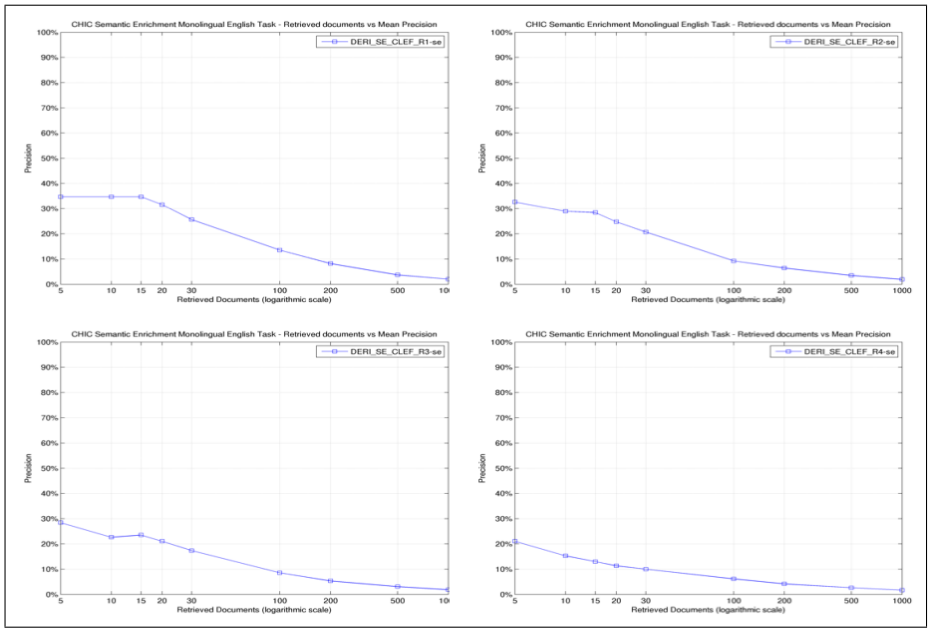

Fig. 1. Retrieved documents vs. mean precision for all of the runs

suggestions.

In order to evaluate the approach in a query expansion environment, all of the suggestions are used as additional terms to the query. With these enriched queries, the results are assessed according to the ad-hoc retrieval standards. Then, the average precision and recall are calculated. In Table 2, we report the Mean Average Precision (MAP) of all the runs. Run1 performs the best, suggesting that concepts may not improve the results over the original query if they contain the query as a substring. For instance, for a given query "Hiroshima", the suggestion "Hiroshima University" may not help to find the relevant documents, that cannot be found by the query "Hiroshima". Figure 2 shows the mean interpolated precision scores against different recall values, and Figure 1 shows the retrieved documents vs. mean precision, for all of the submitted runs.

5 Conclusion and Future Work

We presented our approach for query expansion, which includes concept candidates generation from Wikipedia and DBpedia, and the selection of K-best concepts according to the ESA scores. The approach reached high precision according to the manual relevance assessment evaluation, meaning that most of the suggestions make sense in query expansion. Also, it raises the future direction of

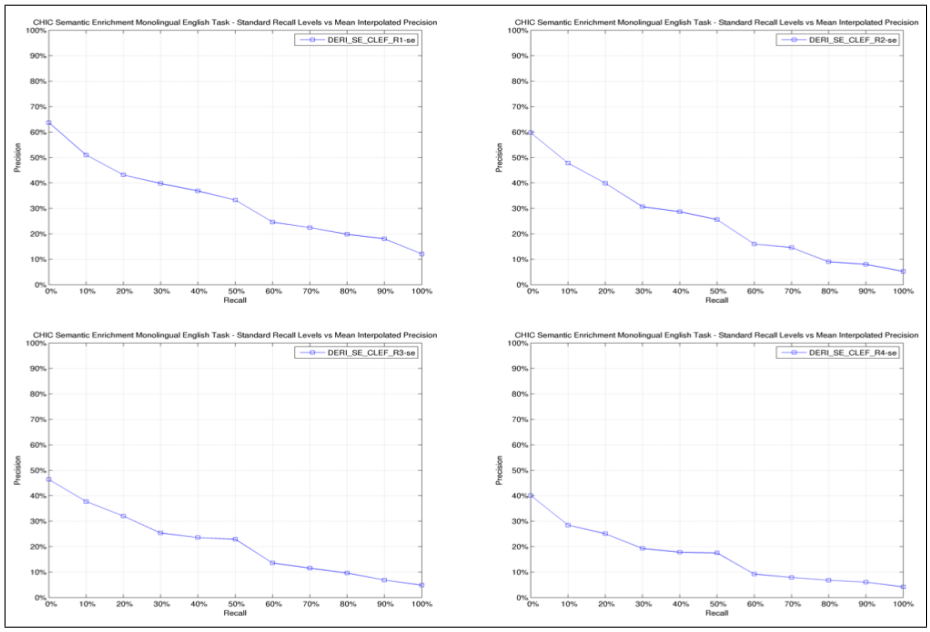

Fig. 2. Standard recall vs. mean interpolated precision for all of the runs

query expansion by using Wikipedia and DBpedia. Therefore, we are planning to investigate this approach with different ranking methods, and by taking more Wikipedia features, such as Wikipedia link and category structure, into account.

References

1. Aggarwal, N., Asooja, K., Buitelaar, P.: DERI&UPM: Pushing corpus based relatedness to similarity: Shared task system description. In: SemEval-2012, SEM, First Joint Conference on Lexical and Computational Semantics, and co-located with NAACL, Montreal, Canada (6 2012)
2. Bhogal, J., Macfarlane, A., Smith, P.: A review of ontology based query expansion. Inf. Process. Manage. 43(4), 866–886 (Jul 2007), http://dx.doi.org/10.1016/j.ipm.2006.09.003
3. Bizer, C., Cyganiak, R., Auer, S., Kobilarov, G.: Dbpedia.org - querying Wikipedia like a Database. In: Developers track at 16th International World Wide Web Conference (WWW2007), Banff, Canada, May 2007 (May 2007)
4. Buscaldi D., Rosso P., S.E.: A WordNet-based query expansion method for geographical information retrieval. In: CLEF 2005 Working Notes. (http://www.clef-campaign.org/2005/working_notes/CLEF2005WN-Contents1.htm) 21-23 September, Vienna, Austria C. Peters (Ed.) (2005)
5. Egozi, O., Markovitch, S., Gabrilovich, E.: Concept-based information retrieval using explicit semantic analysis. ACM Trans. Inf. Syst. 29(2), 8:1–8:34 (Apr 2011), http://doi.acm.org/10.1145/1961209.1961211

6. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: In Proceedings of the 20th International Joint Conference on Artificial Intelligence. pp. 1606–1611 (2007)

Acknowledgements

This work is supported in part by the European Union under Grant No. 248458 for the Monnet project and by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).