Cross Lingual Semantic Search by Improving Semantic Similarity and Relatedness Measures

Nitish Aggarwal
Supervisor: Dr. Paul Buitelaar

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

Abstract. Since 2001, the semantic web community has been working hard towards creating standards which will increase the accessibility of available information on the web. Yahoo research recently reported that 30% of all HTML pages contain structured data such as microdata, RDFa, or microformat. Although multilinguality of the web is a hurdle in information access, the rapid growth of the semantic web enables us to retrieve fine grained information across the language barrier. In this thesis, firstly, we focus on developing a methodology to perform cross-lingual semantic search over structured data (knowledge base), by transforming natural language queries into SPARQL. Secondly, we focus on improving the semantic similarity and relatedness measures, to overcome the semantic gap between the vocabulary in the knowledge base and the terms appearing in the query. The preliminary results are evaluated against the QALD-2 test dataset, which achieved a F1 score of 0.46, an average precision of 0.44, and an average recall of 0.48.

1 Introduction

The rapid growth of the semantic web offers a wealth of semantic knowledge for facilitating an interactive way to access the information, by providing structured metadata[1] in a standard format such as microdata, RDFa or microformat. This structured data facilitates the possibility of automatic reasoning and inferencing. Thus, by embedding such knowledge within web documents, additional key information about the semantic relations among data objects can be captured.

People desire to access the multilingual information available on the web, while querying in their native language. To address this issue, we present cross-lingual semantic search, which aims to retrieve all the relevant information even if it is available in languages different from the query language. Translating search queries ([17], [10]) into the corresponding languages of the documents is the current approach for cross-lingual information retrieval. However, the poor accuracy of translation of short texts like queries, poses a certain problem to

[1] http://events.linkeddata.org/ldow2012/slides/Bizer-LDOW2012-Panel-Background-Statistics.pdf

this method. Hence, using large knowledge bases as an interlingua [23] may prove beneficial.

The approach discussed here considers DBpedia [3] as the structured knowledge base. DBpedia contains a large ontology describing more than 3.5 millions instances extracted from Wikipedia info-boxes, forming a good and general structured knowledge source. Also, it is very well-connected to several other linked data repositories in the Semantic Web. DBpedia contains a huge number of instances in many languages, however, the ontology (properties & classes) is mainly covered in English. Thus, querying this knowledge base is not possible in other languages even if the instances are multilingual. Cross-lingual search is required to query this structured knowledge base, which is the major goal of this work.

In order to query a structured knowledge base, one requires a structured query to start with. Therefore, the conversion of a natural language query (NL-query) to a structured query is required. There are several efforts ([6], [15], [14]) to convert a NL-query to SPARQL[2] in the monolingual scenario. In particular, Freitas et al. [6] proposed an approach based on the combination of entity search, a Wikipedia-based semantic relatedness (using the Explicit Semantic Analysis measure), and spreading activation. Our approach takes inspiration from Freitas et al. to perform search across different languages. We focus on better interpreting NL-queries in different languages, driven by traversal over the large structured knowledge base, and constructing a corresponding SPARQL query. However, the gap between the vocabularies used in NL-queries and the structured knowledge base makes this task challenging. This gap can be filled by calculating cross-lingual similarity and relatedness between these vocabularies, which is the key to our proposed approach. In particular, we present our approach for cross-lingual semantic search, which includes three components: entity search, linguistic analysis, and semantic similarity and relatedness.

Following this approach, cross-lingual document retrieval can also be performed if the documents are already marked-up with the knowledge base, for instance, Wikipedia articles are annotated with DBpedia.

2 Proposed Approach

The key to our approach for cross-lingual semantic search is the interpretation of NL-queries in different languages, driven by the traversal over the large structured knowledge base, and construction of the corresponding SPARQL query. Semantic and linguistic variations of natural language text can create a gap between terms appearing in NL-queries and the vocabulary of the knowledge base. A well-interpreted SPARQL query, which is formed from a given NL-query can overcome this gap, by referring to the knowledge base. Figure 1 shows the three components of our approach along with an example of a NL-query in German[3].

[2] http://www.w3.org/TR/rdf-sparql-query/

[3] Translated from the QALD-2 challenge dataset, which has 100 NL-queries in English, over DBpedia

2.1 Entity Search

The query interpretation process starts by identifying the potential entities, i.e. the ontology concepts (classes and instances), present in the NL-query. A baseline entity search can be defined as the identification of an exact match between the label of an ontology concept and the query text segment by using a simple string similarity, for example, DBpedia: Bill_Clinton shown in Figure 1. However, more sophisticated identification is needed to handle a rich semantic and linguistic analysis of NL-queries, for example, the English NL-query "Give me the capitals of all countries in Africa" has multiple possible entities for the same text segment, "DBpedia: Africa", "DBpedia: Country" and "YAGO: African_Country". "DBpedia: Africa" and "DBpedia: Country" can be identified by the baseline, but these are not the most appropriate entities to link for this NL-query. Therefore, semantic and linguistic analysis are required to identify "YAGO: African_Country" for the text segment "countries in Africa". Semantic analysis provides that "Africa" and "African" are the same and linguistic analysis interprets that "countries in Africa" is equivalent to "African country" as will be explained in Section 2.2.

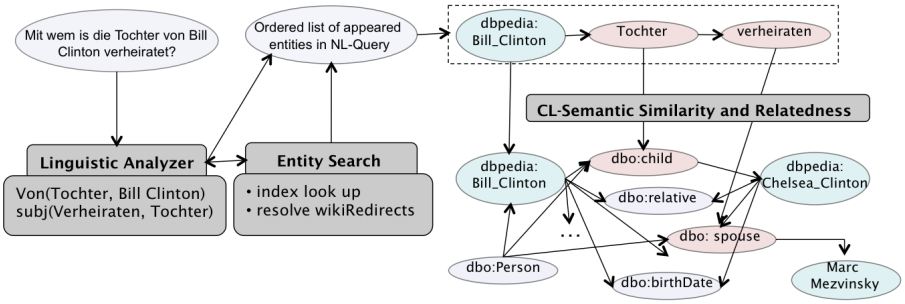

Fig. 1. Query interpretation pipeline for an example German NL-Query "Mit wem is die Tochter von Bill Clinton verheiratet?" which is "Who is the daughter of Bill Clinton married to?" in English

For languages other than English, entity search becomes more challenging as they may include richer linguistic variations such as compound words and gender specific articles.

2.2 Linguistic Analysis

A deep linguistic analysis of the NL-query is performed by generating a parse tree and typed dependencies, by using the Stanford parser.[4] The generated parse tree provides key phrase extraction for identifying potential ontology concepts. For instance, in the query "Who wrote the book The pillars of the Earth?", the

[4] http://nlp.stanford.edu/software/lex-parser.shtml

phrase "The pillars of the Earth" is identified as a noun phrase. This suggests that we should use the whole phrase to find an ontology concept, rather than separated search for each of the tokens. Linguistic analysis also provides entity recognition with linguistic variations. For instance, in the above discussed example, linguistic analysis interprets "Countries in Africa" as PP_in(countries, Africa), which means it is equivalent to YAGO: African_Country.

We convert the given NL-query into an ordered list of potential terms by using generated typed dependencies. To create this ordered list, first we select a central term among all the identified terms, where the central term is the most plausible term to start matching a NL-query to the vocabulary appearing in the knowledge base. This selection is performed by prioritising the ontology instances over classes. Then, we retrieve the directly dependent terms of the central term by following the generated typed dependencies, and add them into the ordered list. Similarly, we perform this action for all the other terms in the list. For instance, in our example NL-query shown in Figure 1, firstly, the system identifies "Bill Clinton" as a central term,[5] and then "Tochter" as direct dependent of "Bill Clinton" followed by "verheiratet" as direct dependent of "Tochter".

2.3 Knowledge base graph traversing using semantic similarity and relatedness

A knowledge base graph can be defined as the structured data of well-connected entities and their properties. Therefore, the next step is the traversal of the obtained ordered list of potential terms from the linguistic analysis step, over this knowledge base. For instance in Figure 1, the ordered list obtained from our example query "Mit wem is die Tochter von Bil Clinton verheiratet?" is <Bill Clinton, Tochter, verheiratet>. Firstly, we search for the Entity "Bill Clinton" in DBpedia as our approach takes DBpedia as knowledge base, and retrieve all of its properties. Then, we find the most semantically similar or related property of direct dependent term "Tochter" by calculating cross-lingual similarity between all the properties of Bill Clinton and the term "Tochter". After obtaining relevant property, i.e. child, we find the entity DBpedia:Chelsea_Clinton, connected with entity Bill Clinton by property child. We perform the same steps with the retrieved entity for directly dependent term "verheiratet" of "Tochter", and so on till end of the ordered list. Finally, we retrieved the relevant entity and also all the linked documents in different languages containing the description about this entity.

Our approach relies on semantic matching between recognised potential terms and properties in the knowledge base. Therefore, to find the most appropriate properties, a good cross-lingual semantic similarity and relatedness measure is required. We cannot rely solely on semantic similarity measures, as relatedness can better map the term "verheiratet" to the retrieved property "spouse", because they are semantically related but not semantically similar. Therefore, to

[5] The term to start the search around in whole DBpedia graph

investigate different similarity and relatedness measures, we are building a Java library which will include many structure-based and corpus-based similarity and relatedness measures. We are performing the experiments with several structure-based measures ([20], [25], [24], [12], [19]) and corpus-based measures ([11], [9], [7], [22]). However, corpus-based approaches rely on the assumption that related words would co-exist in the same document, which is normally not the case with the similar words, e.g. synonymy. Hence, towards the initial step for tuning the corpus-based relatedness to similarity [1], we combine the Explicit Semantic Analysis (ESA) [7] based relatedness score with the WordNet-based Lin [12] similarity scores calculated for the words falling under the corresponding syntactic role category, in both of the short phrases to be compared. We are further working on ESA and its variants (association strength, relevancy function and vector correlations) [22] to improve the corpus-based relatedness, and are planning to submit it in WWW-2013.

3 Evaluation

For the preliminary evaluation of our proposed approach, we examine it in the monolingual scenario. In this experiment, we used the WordNet-based similarity and relatedness proposed by Pirro [19], as it is computationally efficient in comparison to ESA. We performed the experiments [2] against the QALD-2 test dataset, which includes 100 NL-queries in English and their corresponding SPARQL, to retrieve the relevant entities from DBpedia. We calculated the average precision, average recall, and F1 score of the results obtained by our approach. Our approach does not completely explore all of the types of queries appearing in the dataset, as some of them are more challenging complex NL-queries, which would require SPARQL aggregation, and ask type queries. The results are shown in Table 1.

For testing our approach in a cross-lingual setting, we are preparing the benchmark by manually translating the English NL-queries of the QALD-2 test dataset into German.

Total	**Answered**	**Right**	**Partially right**	**Avg. Precision**	**Avg. Recall**	**F1**
100	80	32	7	0.44	0.48	0.46

Table 1. Evaluation on QALD-2 test dataset of 100 NL-queries over DBpedia

4 State of the Art

Most of the proposed approaches to address the task of Cross-Lingual Information Retrieval (CLIR), reduce the problem into the monolingual scenario, by translating the search query or documents in the corresponding language. Many of them perform query translation ([16], [18], [17], [10])) into the language of the documents. However, all of these approaches suffer from the poor performance of the machine translation on short texts (query). Jones et al. [10] performed

query translation by restricting the translation for the cultural heritage domain, while [17] makes use of the Wikipedia cross-lingual links structure.

Without relying on machine translation, some of the approaches ([13], [26], [21]) make use of distributional semantics. They calculate the cross lingual semantic relatedness measures between query and the documents. However, none of these approaches take any linguistic information into account, and do not make use of large available structured knowledge base. With an assumption that documents of different languages are already marked-up with the knowledge base (for instance, Wikipedia articles are annotated with the DBpedia), the problem of CLIR can be converted into the query over structured data. There is still a language barrier, as queries can be in different languages, while most of the structured data are only available in English. Qall-Me [5] performs NL-query over the structured information, by using the textual entailment to convert a natural language question into SPARQL. This system relies on availability of multilingual structured data. It can only retrieve the information which is available in the query language. Therefore, this system is not able to perform CLIR. Freitas et al. [6] proposed an approach for natural language querying over linked data, based on the combination of entity search, a Wikipedia-based semantic relatedness (using ESA) measure, and spreading activation. Our approach takes inspiration from the same.

Since our proposed approach mainly relies on good cross-lingual similarity and relatedness measures, we are working on improving the existing measures to reflect better similarity and relatedness. There are several structure-based methods ([20], [25], [24], [12], [19]), and corpus-based methods ([13], [26], [22]), to calculate similarity and relatedness. Although, structure-based methods require a structure predefined by experts, which is not a trivial task for a large number of language pairs. Corpus-based methods represent the semantics of a term by its distribution in large multilingual corpus, and calculate relatedness by taking correlation between distribution of terms to be compared. These approaches only require comparable multilingual corpus like Wikipedia. However, the corpus-based methods perform well for document similarity, but need to improve for short text or phrases. Therefore, we are working on improving these measures to reflect better similarity and relatedness scores.

5 Conclusion and Future Work

We presented our proposed approach for cross-lingual semantic search, which includes entity search, deep linguistic analysis, and cross-lingual semantic similarity and relatedness. With this approach, cross-lingual information retrieval at document level can also be performed, if the documents are already marked up with the structured knowledge base.

The next main steps are to develop the different components of our proposed approach for cross-lingual semantic search. All of these components mainly rely on better cross-lingual similarity and relatedness measures. Therefore, we are mainly concerned in improving the existing semantic relatedness measures to re-

flect higher accuracy in semantic matching for multiple languages. As discussed in Section 2.3, we are working on ESA and its variants to improve the similarity and relatedness measures. Hence, we are evaluating it with different association strengths such as Latent Semantic Analysis [11] and Latent Dirichlet Allocation [4]. Thomas et al. [8] report significant improvement by taking probabilistic weighted association strength into account. However, one other major issue in corpus-based relatedness is that all the measures do not take the mutual relatedness of documents into account. Hence, we are planning to investigate the current ESA model by fusing it with other existing measures.

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. Aggarwal, N., Buitelaar, P.: A system description of natural language query over dbpedia. In: Interacting with Linked Data (ILD 2012), 9th Extended Semantic Web Conference (5 2012)
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 (2007)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (Mar 2003), http://dl.acm.org/citation.cfm?id=944919.944937
5. scar Ferrndez, Spurk, C., Kouylekov, M., Dornescu, I., Ferrndez, S., Negri, M., Izquierdo, R., Toms, D., Orasan, C., Neumann, G., Magnini, B., Vicedo, J.L.: The qall-me framework: A specifiable-domain multilingual question answering architecture. Web Semantics 9, 137 – 145 (2011)
6. Freitas, A., Oliveira, J.a.G., O'Riain, S., Curry, E., Da Silva, J.a.C.P.: Querying linked data using semantic relatedness: a vocabulary independent approach. In: Proceedings of the 16th NLDB. pp. 40–51 (2011)
7. 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)
8. Gottron, T., Anderka, M., Stein, B.: Insights into explicit semantic analysis. In: Proceedings of the 20th ACM international conference on Information and knowledge management. CIKM '11 (2011)
9. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 50–57. SIGIR '99 (1999)
10. Jones, G., Fantino, F., Newman, E., Zhang, Y.: Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from Wikipedia. CLIA 2008 p. 34 (2008)
11. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Processes 25, 259–284 (1998)
12. Lin, D.: An information-theoretic definition of similarity. In: Proc. of the 15th Int'l. Conf. on Machine Learning. pp. 296–304 (1998), http://portal.acm.org/citation.cfm?id=657297

13. Littman, M., Dumais, S.T., Landauer, T.K.: Automatic cross-language information retrieval using latent semantic indexing. In: Cross-Language Information Retrieval, chapter 5. pp. 51–62. Kluwer Academic Publishers (1998)
14. Lopez, V., Motta, E., Uren, V.S.: Poweraqua: Fishing the semantic web. In: ESWC. pp. 393–410 (2006)
15. Lopez, V., Sabou, M., Motta, E.: Powermap: Mapping the real semantic web on the fly. In: International Semantic Web Conference. pp. 414–427 (2006)
16. Lu, C., Xu, Y., Geva, S.: Web-based query translation for english-chinese CLIR. Computational Linguistics and Chinese Language Processing (CLCLP) 13(1), 61–90 (March 2008)
17. Nguyen, D., Overwijk, A., Hauff, C., Trieschnigg, D.R.B., Hiemstra, D., De Jong, F.: Wikitranslate: query translation for cross-lingual information retrieval using only wikipedia. In: Proceedings of the 9th CLEF (2009)
18. Pirkola, A., Hedlund, T., Keskustalo, H., Jrvelin, K.: Dictionary-based cross-language information retrieval: Problems, methods, and research findings. Information Retrieval 4, 209–230 (2001)
19. Pirró, G., Euzenat, J.: A feature and information theoretic framework for semantic similarity and relatedness. In: Proceedings of the 9th international semantic web conference on The semantic web - Volume Part I. pp. 615–630. ISWC'10 (2010)
20. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. In: IEEE Transactions on Systems, Man and Cybernetics. pp. 17–30 (1989)
21. Sorg, P., Braun, M., Nicolay, D., Cimiano, P.: Cross-lingual information retrieval based on multiple indexes. In: Working Notes for the CLEF 2009 Workshop. Cross-lingual Evaluation Forum, Corfu, Greece (September 2009)
22. Sorg, P., Cimiano, P.: An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. In: Helmut Horacek, Elisabeth Métais, R.M.M.W. (ed.) Proceedings of the International Conference on Applications of Natural Language to Information Systems (NLDB). pp. 36–48. Springer (Juni 2009)
23. Steinberger, R., Pouliquen, B., Ignat, C.: Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. In: In Proc. of the 4th Slovenian Language Technology Conf., Information Society (2004)
24. Sun, P.R.: Using information content to evaluate semantic similarity in a taxonomy. In: In Proceedings of the 14th International Joint Conference on Artificial Intelligence. pp. 448–453
25. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics. pp. 133–138. ACL '94, Association for Computational Linguistics, Stroudsburg, PA, USA (1994), http://dx.doi.org/10.3115/981732.981751
26. Zhang, D., Mei, Q., Zhai, C.: Cross-lingual latent topic extraction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 1128–1137. ACL '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1858681.1858796

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).