Cross-language context extraction for disambiguation

Renan Dembogurski

No Institute Given

Abstract.

1 Introduction

Among the open data communities responsible for the Semantic Web development, one of the most important is DBpedia. As their main site states, DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. In fact, the DBpedia project releases an extraction framework used to extract information from the infoboxes and to convert it to a N-Triples format. The infobox can be interpreted as a summary of the related Wikipedia article.

Most DBpedia resources can be classified according to their corresponding types in the onthology, for instance, a Person, or a Location, etc. This association is done through mappings (between onthology and infoboxes) created manually that are incomplete for various reasons. Among the various mappings provided by DBpedia, one in specific is the dataset linking a DBpedia resource to the same resource in other languages, called inter-language links.

With this dataset it is possible to create an information extraction approach that can be later used to enrich a language in various aspects. In other words, we can search for information related to a certain resource in one language in other languages, using the inter-language mapping. If we do find extra information that was missing from the original language, we can use it to improve the information retrieval process.

We show in this paper that this Cross-language approach can increase the recall value when retrieving information. We also propose a methodology to filter, process and append new attributes for resources. We have used the Portuguese language as our main language, and the English, Italian, Spanish, French and German languages as complement languages. Another contribution of our research is the Portuguese Corpus we developed using 300 news articles that were manually annotated covering various topics.

2 Related Work

Wikipedia contains a large amount of articles written in different languages composing a rich and free content internet encyclopedia. It has been used as external resource in the Natural Language Processing ([3], [9], [5]) and Information Retrieval [4], [10] research fields successfully. Among the tasks that can benefit from a resource pool such as Wikipedia, one is the the Cross-Language Information Retrieval (CLIR) task.

In the Cross-Language scenario a user submits a query in one language to retrieve documents in a different language. Since Wikipedia has inter-language links connecting articles in various languages, it has been used for CLIR extensively [13], [11], [12], [2]. Many approaches and techniques have been proposed for the CLIR task, such as dictionary approach and the disambiguation techniques; probability/statistic-based methods; transitive and triangulation methods; and Web-based approaches.

Regarding disambiguation techniques, there are researches using Wikipedia directly [9], [7] and Linked Open Data sources such as DBpedia [14], [6], [8]. DBpedia also provides the inter-language links from Wikipedia as a set of resources in different files for each language. The idea of increasing the DBpedia coverage using these files was used in [1], where the authors propose an automatic methodology to do so. In this paper we propose a different methodology to increase DBpedia coverage and show the results using our approach in the disambiguation task.

3 Methodology

The methodology adopted for the Cross-Language enrichment is describe in Figure (colocar figura). First we get our two groups of files as input, then we proceed with the filtering, conforming, search and appending stages.

3.1 Entity Representation

Our approach exploits the Wikipedia cross-language links to represent each entity. We use an approach similar to [1] where we represent entities using an entity matrix.

(aqui inicia o texto desse autor, podemos nos basear) Formally, we proceed as follows to automatically derive the set of sentities ϵ, also used to build the training set. Let $\mathcal{L}$ be the set of languages available in Wikipedia, we first build a matrix E where the i-th row represents an entity $e_i \in E$ and j-th column refers to the corresponding language $l_j \in \mathcal{L}$. The crosslanguage links are used to automatically align on the same row all Wikipedia articles that describe the same entity. The element $E_{i,j}$ of this matrix is null if a Wikipedia article describing the entity e_i does not exist in l_j. An instance in our machine learning problem is therefore represented as a row vector e_i where each j-th element is a Wikipedia article in language l_j. Table 1 shows a portion of the entity matrix.

In our case we want to find the first group of types for a resource. We search its main language for types and, if we don't find types for it, we also search additional languages.

en	de	it	. . .
Xolile Yawa	Xolile Yawa	null	. . .
The Locket	null	Il segreto del medaglione	. . .
Barack Obama	Barack Obama	Barack Obama	. . .
null	null	Giorgio Dendi	. . .
. . .	. . .	. . .	. . .

Table 1: A portion of the entity matrix

3.2 Datasets

We will be using several files to fulfill our task of Cross-Language enrichment. These files are provided by DBpedia and can be described in two groups of files:

- The first group has a single file, the interlanguage links triples file. This file is responsible of linking a resource in one language to another resource in another language using the predicate *sameAs*. From now on we will refer to this file as **ILF**. An example of this relation can be seen in 2.
- The second group is composed of triples files with types for each resource in a certain language. From now on we will reference these files as **instance types files**. The relation between a resource and its types is described by the predicate *rdf:type*. The user determines the number of languages to be used.

Subject	Predicate	Object
http://pt.dbpedia.org/resource/Alice	sameAs	http://dbpedia.org/resource/Alice
http://pt.dbpedia.org/resource/Alice	sameAs	http://es.dbpedia.org/resource/Alice
http://pt.dbpedia.org/resource/Alice	sameAs	http://it.dbpedia.org/resource/Alice

Table 2: The structure of the ILF.

The ILF is modified twice to become a file in the matrix representation presented in 1 as a pre-processing step (**ILFC**). Our experiments have shown that this ensures less processing when using our greedy algorithm as will be seen in later sections.

- Explain the Globo corpus

4 Methodology

Since we want to complement the types of new resources in the main language instance types file, we look for new resources in the ILFC triples file. For each line in the latter file we will remove entries we already know, in other words, resources that have types already in their language instance types file. This reduces redundancy and prevents duplication of resources. One interesting factor of this filtering process is that if a user selects few languages as complement, the input file can drop 10 times in size.

After the filtering stage, we proceed with the conforming stage. This second stage transforms the ILF from a lines to a columns format file (ILFC), with every column representing a language as shown in 1. This format have a few advantages:

- We do not need to allocate all the instance types files into memory, we can process them by language (column). In the initial lines format we do not know what languages have types for a certain resource, so we need to allocate multiple instance types files into memory in case we need to query them.
- We can easily apply sequential filters to the ILFC file to reduce processing. As we end querying a language instance types file for types we can remove lines/columns from the ILFC file that won be used when querying types in the next language. An example of the sequential filtering can be seen in Figure (colocare figura).

5 Results

Without Cross-language approach			With Cross-language approach		
Corpus	Accuracy	Global MRR	Corpus	Accuracy	Global MRR
Default	0.778	0.4139	Default	0.888	0.4475

Table 3: A disambiguation performance comparison. The results using the Cross-language approach to the right and not using it to the left.

- Is it worth doing for more languages?

6 Conclusion

Potential Conclusions / Impact
- Is it necessary a “transitive closure” method?
- Drawbacks?

References

1. Aprosio, A.P., Giuliano, C., Lavelli, A.: Automatic expansion of dbpedia exploiting wikipedia cross-language information. In: Cimiano, P., Corcho, s., Presutti, V.,

Hollink, L., Rudolph, S. (eds.) ESWC. Lecture Notes in Computer Science, vol. 7882, pp. 397–411. Springer (2013), http://dblp.uni-trier.de/db/conf/esws/eswc2013.html#AprosioGL13

2. Braschler, M., Schäuble, P., Peters, C.: Cross-language information retrieval (clir) track overview. In: TREC (1999), http://dblp.uni-trier.de/db/conf/trec/trec1999.html#BraschlerSP99
3. Buscaldi, D., Rosso, P.: Mining knowledge from wikipedia for the question answering task. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC). pp. 727–730. Genoa, Italy (2006)
4. 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
5. Erdmann, M., Nakayama, K., Hara, T., Nishio, S.: An approach for extracting bilingual terminology from wikipedia. In: Proceedings of the 13th international conference on Database systems for advanced applications. pp. 380–392. DASFAA'08, Springer-Verlag, Berlin, Heidelberg (2008), http://dl.acm.org/citation.cfm?id=1802514.1802552
6. Hakimov, S., Oto, S.A., Dogdu, E.: Named entity recognition and disambiguation using linked data and graph-based centrality scoring. In: Proceedings of the 4th International Workshop on Semantic Web Information Management. pp. 4:1–4:7. SWIM '12, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2237867.2237871
7. Li, C., Sun, A., Datta, A.: A generalized method for word sense disambiguation based on Wikipedia. In: Proceedings of the 33rd European conference on Advances in information retrieval. Lecture Notes in Computer Science, vol. 6611, pp. 653–664. Springer, Berlin/Heidelberg (2011), http://portal.acm.org/citation.cfm?id=1996889.1996972
8. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: Shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems. pp. 1–8. I-Semantics '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2063518.2063519
9. Mihalcea, R.: Using Wikipedia for Automatic Word Sense Disambiguation. In: North American Chapter of the Association for Computational Linguistics (NAACL 2007) (2007)
10. Milne, D.N., Witten, I.H., Nichols, D.M.: A knowledge-based search engine powered by wikipedia. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. pp. 445–454. CIKM '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1321440.1321504
11. 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 Cross-language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access. pp. 58–65. CLEF'08, Springer-Verlag, Berlin, Heidelberg (2009), http://dl.acm.org/citation.cfm?id=1813809.1813818
12. Nie, J.Y.: Cross-Language Information Retrieval. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers (2010)
13. Sorg, P., Cimiano, P.: Cross-lingual Information Retrieval with Explicit Semantic Analysis. In: Working Notes for the CLEF 2008 Workshop (2008), http://www.clef-campaign.org/2008/working_notes/sorg_paperCLEF2008.pdf

14. Villar Rodríguez, E., Torre-Bastida, A.I., García-Serrano, A., González, M.: Using linked open data sources for entity disambiguation. In: Online Reputation Management - RepLab. Valencia (09/2013 2013)