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 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 the structured information contained in the infoboxes and to convert it in triples.

In fact, every DBpedia resource can be mapped into types in the onthology (Person, 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. One of the important aspects that can be improved by this approach is the types coverage in a certain language.

questions/problems: - how to identify/map entities from each language? - Freebase types are enough for this process? - Is there any drawback?

hypotheses: - an inter-language approach provides better accuracy

2 Related Work

As a rich and free resource, Wikipedia contains very large amount of articles written in different languages and various types of link information showing the relations between articles. It has been used as external resource in many natural language processing tasks successfully ([2], [4], [3]). Among the link information in Wikipedia, the inter-language link, which is created by article authors, connects large amount of articles that describe the same term but are written in different languages

Besides of article titles, there still exists large amount of information that could be used for dictionary construction, such as the text inside the linked articles.

//Conferir - It has been demonstrated that these inter-language links are useful resources for bilingual dictionary construction ([3]).

Papers:

- Bilingual Dictionary Extraction from Wikipedia

- Automatic Expansion of DBpedia Exploiting Wikipedia Cross-Language Information

- Cross-Language Information Retrieval with Latent Topic Models Trained on a Comparable Corpus

3 Methodology

At the time of starting the experiments, there are 360 mappings available for the English DBpedia, covering around 1.7M entities, against almost 4M articles in Wikipedia. The remaining pages are automatically mapped to the trivial top-level class owl:Thing. Hereafter, when we speak about coverage, we will always refer to classes different from owl:Thing. (tirei de um artigo, talvez atualizar e usar no nosso?)

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. Figure 1 shows a portion of the entity matrix.

(interessante eu acabei modificando o arquivo do interlanguage igual esse autor, podemos justificar a escolha pela matriz dele)

(COLOCAR MATRIZ)

3.2 Datasets

We will be using the English and Portuguese Wikipedias, and corresponding DBPedias. The process is described here:

https://docs.google.com/document/d/1jyfbK5wYL9L5ljMaRDVm4gCacLPM-sz-S5AsqQoOuQM/edit

4 Experiments

...

5 Conclusion

Potential Conclusions / Impact
- Is it necessary a “transitive closure” method?
- Drawbacks?
- How complete are the Freebase types?

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. 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)
3. 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
4. Mihalcea, R.: Using Wikipedia for Automatic Word Sense Disambiguation. In: North American Chapter of the Association for Computational Linguistics (NAACL 2007) (2007)