Multilingual Disambiguation of Named Entities Using Linked Data

Ricardo Usbeck♠♢, Axel-Cyrille Ngonga Ngomo♢, Wencan Luo♡, and Lars Wesemann♠

♢ University of Leipzig, Germany ,
♠ R & D, Unister GmbH, Leipzig, Germany,
♡ University of Pittsburgh, United States of America
email: {usbeck|ngonga}@informatik.uni-leipzig.de

Abstract. One key step towards extracting structured data from unstructured data sources is the disambiguation of entities. With AGDISTIS, we provide a time-efficient, state-of-the-art, knowledge-base-agnostic and multilingual framework for the disambiguation of RDF resources. The aim of this demo is to present the English, German and Chinese version of our framework based on DBpedia. We show the results of the framework on texts pertaining to manifold domains including news, sports, automobiles and e-commerce. We also summarize the results of the evaluation of AGDISTIS on several languages.

1 Introduction

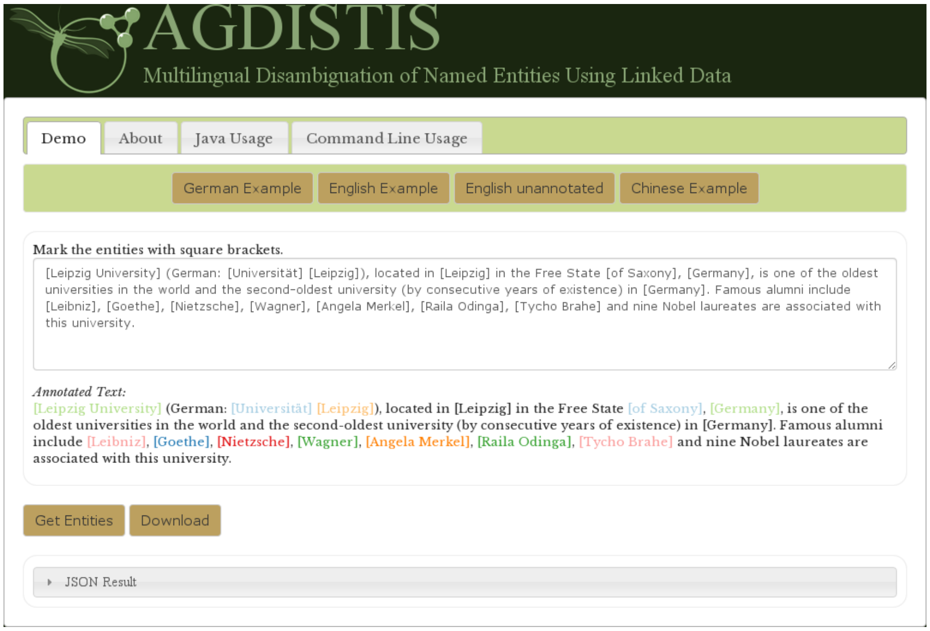
A significant portion of the information on the Web is still only available in textual format. Addressing this information gap between the Document Web and the Data Web requires amongst others the extraction of entities and relations between these entities from text. One key step during this processing is the disambiguation of entities (also known as entity linking). The AGDISTIS framework [7] (which will also be presented at this conference) addresses two of the major drawbacks of current entity linking frameworks [1,2,3]: time complexity and accuracy. With AGDISTIS, we have developed a framework that achieves polynomial time complexity and outperforms the state of the art w.r.t. accuracy. The framework is knowledge-base-agnostic (i.e., it can be deployed on any knowledge base) and is also language-independent. In this demo, we will present AGDISTIS deployed on three different languages (English, German and Chinese) and three different knowledge bases (DBpedia, the German DBpedia and the Chinese DBpedia). To the best of our knowledge, we therewith provide the first Chinese instantiation of entity linking to DBpedia. We will also demonstrate the AGDISTIS web services endpoints for German, English and Chinese disambiguation and show how data can be sent to the endpoints. Moreover, the output format of AGDISTIS will be explained. An online version of the demo is available at `http://agdistis.aksw.org/demo`.

2 Demonstration

Within our demonstration, we aim to show how AGDISTIS can be used by non-expert as well as expert users. For non-experts, we provide a graphical user interface (GUI). Experts can choose to use the REST interfaces provided by the tool or use a Java snippet to call the REST interface. The whole of this functionality, which will be described in more details in the following sections, will also be demonstrated at the conference.

2.1 AGDISTIS for non-expert users

A screenshot of the AGDISTIS GUI is shown in Figure 1. This GUI supports the following workflow.



Mark the entities with square brackets.

[Leipzig University] (German: [Universität] [Leipzig]), located in [Leipzig] in the Free State [of Saxony], [Germany], is one of the oldest universities in the world and the second-oldest university (by consecutive years of existence) in [Germany]. Famous alumni include [Leibniz], [Goethe], [Nietzsche], [Wagner], [Angela Merkel], [Raila Odinga], [Tycho Brahe] and nine Nobel laureates are associated with this university.

Annotated Text:

[Leipzig University] (German: [Universität] [Leipzig]), located in [Leipzig] in the Free State [of Saxony], [Germany], is one of the oldest universities in the world and the second-oldest university (by consecutive years of existence) in [Germany]. Famous alumni include [Leibniz], [Goethe], [Nietzsche], [Wagner], [Angela Merkel], [Raila Odinga], [Tycho Brahe] and nine Nobel laureates are associated with this university.

Fig. 1: Screenshot of the demo with an English example which is already annotated.

Entity Recognition After typing or pasting text into the input field, users can choose between either annotating the entities manually or having the entities detected automatically. In the first case, the labels of the entities are to be marked by using square brackets (see central panel of Figure 1). In the case of an automatic annotation, we send the text to the FOX framework, which has been shown to outperform the state of the art in [6]. We will demonstrate this feature by using both manually pre-annotated text and text without annotations in our examples (see upper bar of Figure 1). Moreover, we will allow the crowd to enter arbitrary texts that pertain to their domain of interest.

Automatic Language Detection Once the user has set which entities are to be disambiguated, the marked-up text is sent to the language detection module based on [5]. We chose this library because it is both precise (> 99% precision) and time-efficient. If the input is detected to belong to one of the languages we support (i.e., German, Chinese, English), then we forward the input to a dedicated AGDISTIS instance for this given language. In all other cases, an error message is shown to the user, pointing towards the language at hand not being supported. The main advantage of this approach is that the user does not need to select the language in which the text is explicated manually, thus leading to an improved user experience. We will demonstrate this feature by entering text in different languages (German, English, French, Chinese, etc.) and presenting the output of the framework for each of these test cases.

Entity Linking This is the most important step of the whole workflow. The annotated text is forwarded to the corresponding language-specific deployment of AGDISTIS, of which each relies on a language-specific version of DBpedia 3.9. The approach underlying AGDISTIS [7] is language-independent and combines breadth-first search and the well-known HITS algorithm. In addition, string similarity measures and label expansion heuristics are used to account for typos and morphological variations in naming. Moreover, Wikipedia-specific surface forms for resources can be used.

Output Within the demo the annotated text is shown below the input field where disambiguated entities are colored to highlight them. While hovering a highlighted entity the disambiguated URI is shown. We will demonstrate the output of the entity linking by using the examples shown in the upper part of Figure 1. The output of the system will be shown both in a HTML version and made available as a download in JSON. Moreover, we will allow interested participants to enter their own examples and view the output of the tool.

2.2 AGDISTIS for expert users

To support different languages we set up a REST URI for each of the language versions. Each of these endpoints understands two mandatory parameters: (1) `text` which is an UTF-8 and URL encoded string with entities annotated with XML-tag `<entity>` and (2) `type='agdistis'` to disambiguate with the AGDISTIS algorithm. In the future, several wrappers will be implemented to use different entity linking algorithms for comparison. Following, a CURL[1] snippet shows how to address the web service, see also `http://agdistis.aksw.org`:

```
curl --data-urlencode "text='<entity>Barack Obama</entity> arrives
in <entity>Washington, D.C.</entity>.'" -d type='agdistis'
{AGDISTIS URL}/AGDISTIS
```

[1] http://curl.haxx.se/

3 Evaluation

English and German Evaluation. AGDISTIS has been evaluated on 8 different datasets from diverse domains such as news, sports or buisiness reports. For English datasets AGDISTIS is able to outperform the currently best disambiguation framework, TagMe2, on three out of four datasets by up to 29.5% F-measure. Considering the only German dataset available for named entity disambiguation, i.e., `news.de` [4], we are able to outperform the only competitor DBpedia Spotlight by 3% F-measure.
Chinese Evaluation. We evaluated the Chinese version of AGDISTIS within a question answering setting. To this end, we used the multilingual benchmark provided in QALD-4[2]. Since the Chinese language is not supported, we extended the QALD-4 benchmark by translating the English questions to Chinese and inserted the named entity links manually. The accuracies achieved by AGDISTIS for the train and test datasets are 65% and 70% respectively.

4 Conclusion

We presented the demo of AGDISTIS for three different languages on three different DBpedia-based knowledge bases. In future work, we aim to create a single-server multilingual version of the framework that will intrinsically support several languages at the same time. To this end, we will use a graph merging algorithm to combine the different versions of DBpedia to a single graph. The disambiguation steps will then be carried out on this unique graph.

Acknowledgments This work has been supported by the ESF and the Free State of Saxony and the FP7 project GeoKnow (GA No. 318159).

References

1. Paolo Ferragina and Ugo Scaiella. Fast and accurate annotation of short texts with wikipedia pages. *IEEE software*, 29(1), 2012.
2. Pablo N. Mendes, Max Jakob, Andres Garcia-Silva, and Christian Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, 2011.
3. Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation: a unified approach. *TACL*, 2:231–244, 2014.
4. Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both. N3 - a collection of datasets for named entity recognition and disambiguation in the nlp interchange format. In *LREC*, 2014.
5. Nakatani Shuyo. Language detection library for java, 2010.
6. René Speck and Ngonga Ngomo. Ensemble learning for named entity recognition. In *International Semantic Web Conference*. 2014.
7. Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Sören Auer, Daniel Gerber, and Andreas Both. Agdistis - agnostic disambiguation of named entities using linked open data. In *International Semantic Web Conference*. 2014.

[2] `http://greententacle.techfak.uni-bielefeld.de/~cunger/qald`