Semantic Annotation, Analysis and Comparison: A Multilingual and Cross-lingual Text Analytics Toolkit

Lei Zhang
Institute AIFB
Karlsruhe Institute of Technology
76128 Karlsruhe, Germany
l.zhang@kit.edu

Achim Rettinger
Institute AIFB
Karlsruhe Institute of Technology
76128 Karlsruhe, Germany
rettinger@kit.edu

Abstract

Within the context of globalization, multilinguality and cross-linguality for information access have emerged as issues of major interest. In order to achieve the goal that users from all countries have access to the same information, there is an impending need for systems that can help in overcoming language barriers by facilitating multilingual and cross-lingual access to data. In this paper, we demonstrate such a toolkit, which supports both service-oriented and user-oriented interfaces for semantically annotating, analyzing and comparing multilingual texts across the boundaries of languages. We conducted an extensive user study that shows that our toolkit allows users to solve cross-lingual entity tracking and article matching tasks more efficiently and with higher accuracy compared to the baseline approach.

1 Introduction

Automatic text understanding has been an unsolved research problem for many years. This partially results from the dynamic and diverging nature of human languages, which results in many different varieties of natural language. These variations range from the individual level, to regional and social dialects, and up to seemingly separate languages and language families. In recent years there have been considerable achievements in approaches to computational linguistics exploiting the information across languages. This progress in multilingual and cross-lingual text analytics is largely due to the increased availability of multilingual knowledge bases such as Wikipedia, which helps at scaling the traditionally monolingual tasks to multilingual and cross-lingual applications. From the application side, there is a clear need for multilingual and cross-lingual text analytics technologies and services.

Text analytics in this work is defined as three tasks: (i) *semantic annotation* by linking entity mentions in the documents to their corresponding representations in the knowledge base; (ii) *semantic analysis* by linking the documents by topics to the relevant resources in the knowledge base; (iii) *semantic comparison* by measuring semantic relatedness between documents. While *multilingual* text analytics addresses these tasks for multiple languages, *cross-lingual* text analytics goes one step beyond, as it faces these tasks across the boundaries of languages, i.e., the text to be processed and the resources in the knowledge base, or the documents to be compared, are in different languages.

Due to the ever growing richness of its content, Wikipedia has been increasingly gaining attention as a precious knowledge base that contains an enormous number of entities and topics in diverse domains. In addition, Wikipedia pages that provide information about the same concept in different languages are connected through cross-language links. Therefore, we use Wikipedia as the central knowledge base.

With the goal of overcoming language barriers, we would like to demonstrate our multilingual and cross-lingual text analytics toolkit, which supports both service-oriented and user-oriented interfaces for semantically annotating, analyzing and comparing multilingual texts across the boundaries of languages.

2 Techniques

In this section, we first present the techniques behind our toolkit w.r.t. its three components: semantic annotation (Sec. 2.1), semantic analysis and semantic comparison (Sec. 2.2).

Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 13–16,
Gothenburg, Sweden, April 26-30 2014.

2.1 Wikipedia-based Annotation

The process of augmenting phrases in text with links to their corresponding Wikipedia articles (in the sense of Wikipedia-based annotation) is known as *wikification*. There is a large body of work that links phrases in unstructured text to relevant Wikipedia articles. While Mihalcea and Csomai (Mihalcea and Csomai, 2007) met the challenge of wikification by using link probabilities obtained from Wikipedia's articles and by a comparison of features extracted from the context of the phrases, Milne and Witten (Milne and Witten, 2008) could improve the wikification service significantly by viewing wikification even more as a supervised machine learning task: Wikipedia is used here not only as a source of information to point to, but also as training data to find always the appropriate link.

For multilingual semantic annotation, we adopted the wikification system in (Milne and Witten, 2008) and trained it for each language using the corresponding Wikipedia version. To perform cross-lingual semantic annotation, we extended the wikification system by making use of the cross-language links in Wikipedia to find the corresponding Wikipedia articles in the different target languages. More details can be found in our previous work (Zhang et al., 2013).

2.2 Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) has been proposed as an approach for semantic modeling of natural language text (Gabrilovich and Markovitch, 2006). Based on a given set of concepts with textual descriptions, ESA defines the concept-based representation of documents. Various sources for concept definitions have been used, such as Wikipedia and Reuters Corpus. Using the concept-based document representation, ESA has been successfully applied to compute semantic relatedness between texts (Gabrilovich and Markovitch, 2007). In the context of the cross-language information retrieval (CLIR) task, ESA has been extended to a cross-lingual setting (CL-ESA) by mapping the semantic document representation from a concept space of one language to an interlingual concept space (Sorg and Cimiano, 2008).

The semantic analysis and semantic comparison components of our toolkit are based on CL-ESA in (Sorg and Cimiano, 2008). The semantic analysis component takes as input a document in a source language and maps it to a high-dimensional vector in the interlingual concept space, such that each dimension corresponds to an Wikipedia article in any target language acting as a concept. For semantic comparison, the documents in different languages are first translated into vectors in the interlingual concept space and then the cross-lingual semantic relatedness between the documents in different languages can be calculated using the standard similarity measure between the resulting vectors.

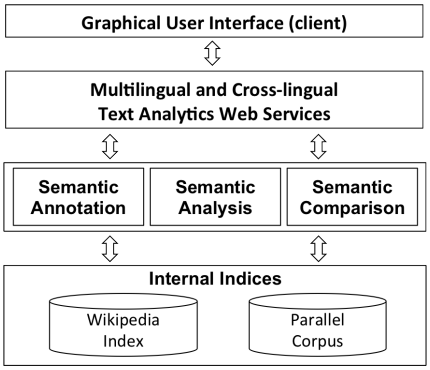


Figure 2: Architecture of our Toolkit.

3 Implementation

Our multilingual and cross-lingual toolkit is implemented using a client-server architecture with communication over HTTP using a XML schema defined in XLike project[1]. The server is a RESTful web service and the client user interface is implemented using Adobe Flex as both Desktop and Web Applications. The toolkit can easily be extended or adapted to switch out the server or client. In this way, it supports both service-oriented and user-oriented interfaces for semantically annotating, analyzing and comparing multilingual texts across the boundaries of languages. The architecture of our toolkit is shown in Figure 2.

For all three components, namely semantic annotation, analysis and comparison, we use Wikipedia as the central knowledge base. Table 1 shows the statistics of the Wikipedia articles in English, German, Spanish and French as well as the cross-language links between the them in these languages extracted from Wikipedia snapshots of May 2012[2], which are used to build our toolkit.

We now describe the user interfaces of these

[1] http://www.xlike.org/

[2] http://dumps.wikimedia.org/

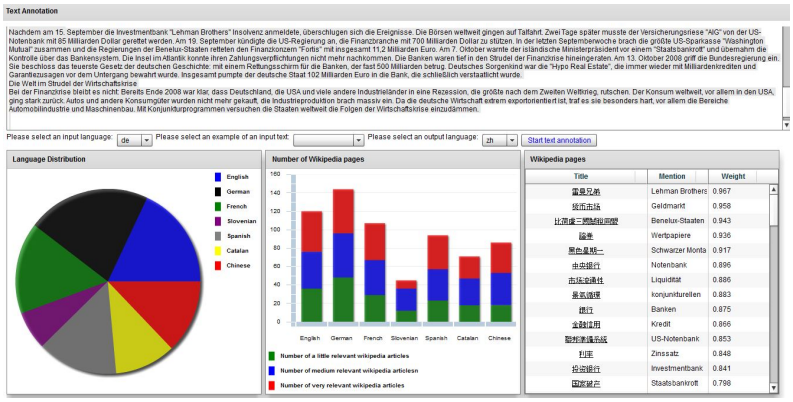


Figure 1: Screenshot of the Semantic Annotation Component of our Toolkit.

	English (EN)	German (DE)	Spanish (ES)	French (FR)
#Articles	4,014,643	1,438,325	896,691	1,234,567

(a) Number of articles.

	EN-DE	EN-ES	EN-FR	DE-ES	DE-FR	ES-FR
#Links (→)	721,878	568,210	779,363	295,415	455,829	378,052
#Links (←)	718,401	581,978	777,798	302,502	457,306	370,552
#Links (merged)	722,069	593,571	795,340	307,130	464,628	383,851

(b) Number of cross-language links.

Table 1: Statistics about Wikipedia.

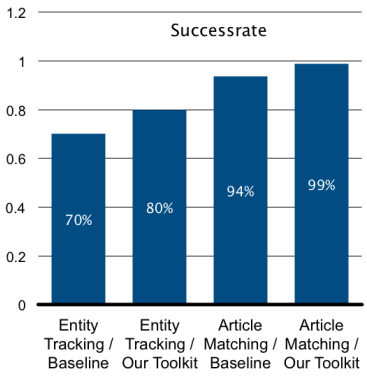
components. Due to the lack of space, we only show the screenshot of the semantic annotation component in Figure 1. The semantic annotation component allows the users to find the entities in Wikipedia mentioned in the input document. Given the input document in one language, the users can select the output language, namely the language of Wikipedia articles describing the mentioned entities. In the left pie chart, the users can see the percentage of Wikipedia articles in different languages as annotations of the input document. According to their weights, the Wikipedia articles in each language are organized in 3 relevance categories: high, medium and low. In the middle bar chart, the number of Wikipedia articles in each language and in each category is illustrated. The right data grid provides the Wikipedia article titles with their weights in the output language and the mentions in the input document. Clicking an individual title opens the corresponding Wikipedia article in the output language. The semantic analysis component has the similar user interface as the semantic annotation component. The difference is that the Wikipedia articles listed in the right data grid are topically relevant to the input documents instead of being mentioned as entities. Regarding the user interface of semantic comparison component, the main inputs are two documents that might be in different languages and the output is the semantic relatedness between them.

4 User Study

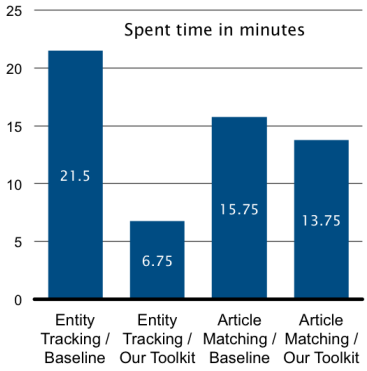
We conducted a task-based user study and the goal is to assess the effectiveness and usability of our multilingual and cross-lingual text analytics toolkit. We design two tasks reflecting the real-life information needs, namely *entity tracking* and *article matching*, to assess the functionality of our toolkit from different perspectives. The *entity tracking* task is to detect mentions of the given entities in the articles, where the descriptions of the entities and the articles are in different languages. Given articles in one language as context, the *article matching* task is to find the most similar articles in another language.

The participants of our user study are 16 volunteers and each of them got both tasks, which they had to solve in two ways: (1) using a major online machine translation service as baseline and (2) using our multilingual and cross-lingual text analytics toolkit with all the functionality. For each task, we randomly selected 10 parallel articles in English, French and Spanish from the JRC-Acquis parallel corpus[3]. After a survey,

[3] http://langtech.jrc.it/JRC-Acquis.html



(a) Avg. successrate per task / method



(b) Avg. time spent per task / method

Figure 3: Evaluation Results of the User Study.

we decided to provide the entity descriptions for entity tracking task and the context documents for article matching task in English, which all participants can speak. Regarding the articles to be processed, we set up the tasks using Spanish articles for the participants who do not know Spanish, and tasks with French articles for the participants who cannot speak French.

To measure the overall effectiveness of our toolkit, we have analysed the ratio of tasks that were completed successfully and correctly and the time the participants required for the tasks. The average success rate and time spent per task and per method are illustrated in Figure 3. For entity tracking task, we observe that a success rate of 80% was achieved using our toolkit in comparison with the success rate of 70% yielded by using the baseline. In addition, there is a significant gap between the time spent using different methods. While it took 21.5 minutes on average to solve the task using the baseline, only 6.75 minutes were needed when using our toolkit. Regarding the article matching task, both methods performed very well. Using our toolkit obtained a slightly higher success rate of 99% than 94% using the baseline. The time spent using both methods is not so different. The participants spent 15.75 minutes on average using the baseline while 2 minutes less were needed using our toolkit.

In terms of the user study, our toolkit is more effective than the baseline for both entity tracking and article matching tasks. Therefore, we conclude that our toolkit provides useful functionality to make searching entities, analyzing and comparing articles more easily and accurately in the multilingual and cross-lingual scenarios.

Acknowledgments

This work was supported by the European Community's Seventh Framework Programme FP7-ICT-2011-7 (XLike, Grant 288342).

References

[Gabrilovich and Markovitch2006] Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *AAAI*, pages 1301–1306.

[Gabrilovich and Markovitch2007] Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on artificial intelligence*, volume 6, page 12.

[Mihalcea and Csomai2007] Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *In CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM.

[Milne and Witten2008] David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA. ACM.

[Sorg and Cimiano2008] P. Sorg and P. Cimiano. 2008. Cross-lingual Information Retrieval with Explicit Semantic Analysis. In *Working Notes of the Annual CLEF Meeting*.

[Zhang et al.2013] Lei Zhang, Achim Rettinger, Michael Frber, and Marko Tadic. 2013. A comparative evaluation of cross-lingual text annotation techniques. In *CLEF*, pages 124–135.