Xlike - Cross-lingual Knowledge Extraction (FP7-ICT-2011-7)

Marko Grobelnik
Jozef Stefan Institute
Ljubljana, Slovenia
marko.grobelnik@ijs.si

Blaž Fortuna
Jozef Stefan Institute
Ljubljana, Slovenia
blaz.fortuna@ijs.si

Gregor Leban
Jozef Stefan Institute
Ljubljana, Slovenia
gregor.leban@ijs.si

Jan Rupnik
Jozef Stefan Institute
Ljubljana, Slovenia
jan.rupnik@ijs.si

Andrej Muhič
Jozef Stefan Institute
Ljubljana, Slovenia
andrej.muhic@ijs.si

Aljaž Košmerlj
Jozef Stefan Institute
Ljubljana, Slovenia
aljaz.kosmerlj@ijs.si

ABSTRACT

XLike is a European research project funded under FP7 that lasted between 2012 and 2015. The goal of the project was to develop technology to monitor, aggregate and extract knowledge that is spread across global mainstream and social media and to enable cross-lingual services for publishers, media monitoring and business intelligence.

KEYWORDS

Cross-linguality, knowledge extraction, text mining

1 INTRODUCTION

The goal of the X-LIKE project was to develop technology to monitor and aggregate knowledge that is currently spread across global mainstream and social media, and to enable cross-lingual services for publishers, media monitoring and business intelligence.

In terms of research contributions, the aim was to combine scientific insights from several scientific areas to contribute in the area of cross-lingual text understanding. By combining modern computational linguistics, machine learning, text mining and semantic technologies we dealt with the following two key open research problems:

- extraction and integration of formal knowledge from multilingual texts with cross-lingual knowledge bases, and
- adaptation of linguistic techniques and crowdsourcing to deal with irregularities in informal language used primarily in social media.

As an interlingua, knowledge resources from Linked Open Data cloud (http://linkeddata.org/) were used with special focus on general common sense knowledge base CycKB (http://www.cyc.com/). For the languages where no required linguistic resources were available, we used a probabilistic interlingua representation trained from a comparable corpus drawn from the Wikipedia.

The developed solutions were be applied on two case studies, both from the area of news. For the Bloomberg case study the domain was financial news, while for the Slovenian Press Agency we dealt with general news. The technology developed in the project was used to introduce cross-lingual knowledge and information from social media in services for publishers and end-users in the area of summarization, contextualization, personalization, and plagiarism detection. Special attention was paid to analyzing news reporting bias from multilingual sources.

The developed technology is language-agnostic, while within the project we specifically addressed English, German, Spanish, and Chinese as major world languages and Catalan and Slovenian as minority languages.

The XLike project was coordinated by Jozef Stefan Institute (Slovenia). The other technology partners were also Karlsruher Institute of Technology (Germany), Polytechnic University of Catalonia (Spain), University of Zagreb (Croatia), Tsinghua Univeristy (China) and Intelligent Software Components S.A. (Spain). The use case partners were Bloomberg L.P. (United States) and Slovenian Press Agency (Slovenia).

Beside the research outcomes, there were also several technologies and services that were developed that are still actively used and developed. These services will be described in the next section.

2 MOST RELEVANT DEVELOPED SERVICES

2.1 Xling

In cases when we need to compare similarity between two documents in different languages, translating a document first can be an unnecessarily expensive operation. As an alternative we developed a service that can take on input two documents in and of 100 most popular languages and compute a similarity score between the documents. The service is based on the Canonical Correlation Analysis and uses aligned Wikipedia pages to train a common semantic space into which the original documents are projected. The service is available at http://aidemo.ijs.si/xling/wikipedia.html

2.2 Wikifier

Wikifier is an entity linking service for 100 most popular world languages. It is able to take the document and identify in it mentioned people, locations, organizations and things and disambiguate them using Wikipedia as the knowledge base. In order to disambiguate among hundreds of possible candidates for a given phrase, the service takes into account the context of the whole document. The service is available at http://wikifier.org/

2.3 Event Registry

Event Registry is a service for global media monitoring. It collects and analyzed news in over 30 languages in real time. All news are semantically annotated and categorized. By analyzing the similarity of the news content, the articles are also grouped into events based on their relatedness. Each group of articles contain all different news content that all discuss about the same thing that happened. Due to cross-lingual services, each event can contain articles in several languages. For each event, semantic information such as what happened, where, when, who was involved, etc. is also automatically extracted. All news content as well as extracted events are automatically stored in the system, which currently stores over 250 million news articles and 10 million events extracted since 2014. Event Registry provides extensive search options where you can find articles or events based on topics, categories, news sources, locations, dates, etc. Results can not only be listed but also visualized in numerous ways in order to display the timeline, top news sources, top concepts, keywords, categories, and other properties that summarize a large number of results. Event Registry is available at http://eventregistry.org/

REFERENCES

[1] Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014a). *Cross-lingual detection of world events from news articles*. In Proceedings of the 13th International Semantic Web Conference, pp. 21-24.

[2] Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014b). *Event Registry: Learning About World Events from News*. In Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, WWW Companion '14, pp. 107-110. International World Wide Web Conferences Steering Committee.

[3] J. Brank, G. Leban, and M. Grobelnik. *A high-performance multithreaded approach for clustering a stream of documents*. In Proceedings of the 17th International Multiconference Information Society, 2014.

[4] J. Brank, G. Leban, and M. Grobelnik. *Annotating documents with relevant wikipedia concepts*. Proceedings of SiKDD 2017, forthcoming, 2017.

[5] J. Rupnik, A. Muhic, G. Leban, P. Skraba, B. Fortuna, and M. Grobelnik. *News across languages-cross-lingual document similarity and event tracking*. Journal of Artificial Intelligence Research, 55:283–316, 2016.

[6] M. Trampus and B. Novak. *The internals of an aggregated web news feed.* Proceedings of 15th Multiconference on Information Society IS-2012, 2012.

[7] A. Muhic, J. Rupnik, and P. Skraba. *Cross-lingual document similarity.* Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces (ITI), IEEE, pages 387–392, 2012.