Introducing LDL4HELTA: Linked data lexicography for high-end language technology application

Martin Kaltenböck and Ilan Kernerman

Martin Kaltenböck is Co-Founder, CFO and Managing Partner of Semantic Web Company. He leads the LDL4HELTA project, and leads and works in several national and international research, industry and public administration projects – mainly as regards project management, requirements engineering, and community and communication activities. He studied communication, psychology and marketing at the University of Vienna. https://linkedin.com/in/martinkaltenboeck/

Background

Business is becoming increasingly globalized, and enterprises as well as public organisations are increasingly acting in multiple areas, facing challenges of cross-lingual and inter-cultural barriers. English has been crowned a global language, yet regional identities flourish as well, and this trend correlates with a rise in human and machine solutions to facilitate and enhance communication across languages and cultures. Looking at the European market, for example, we see 24 working languages in EU28, which make cross-border services considerably complicated. This calls for powerful language technology (LT) and intense efforts to enable and materialize the vision of a multilingual digital single market (defined as a priority area of the European Commission, see http://ec.europa.eu/priorities/digital-single-market/).

As a result, we see a continuously growing LT market, primarily in Europe but also worldwide. The growth leads LT entrepreneurs to suggest solutions for data- and information-driven organisations to work internationally, to efficiently store, access, integrate and disseminate their data, and to allow for both inter- and intra-organisation communication across borders and language barriers by utilizing software tools.

Although the emerging LT industry is fairly young, it has rapidly changed the rules of the game and excluded major old-time players. It ranges as widely as machine translation (e.g. Google Translate, Translation Memory tools), speech technologies (e.g. Apple's Siri digital assistant), education (e.g. e-learning), text processing (e.g. MS Word / Office 365), text mining, content analysis, localization, etc. New major players consist of a range of world leading software and hardware corporations (Google, Apple, Microsoft, Facebook, SAP, IBM, Intel, Amazon, etc.) as well as newcomers offering innovative solutions.

Semantics and lexicography

To develop and provide satisfactory LT mechanisms and tools we need suitable software as well as high quality data and information in the form of corpora, dictionaries, word form lists and other fine lexical resources. Lexicography is a vital component in this ecosystem. It follows mainly a qualitative approach that tends to be both time-consuming and cost-intensive. Interoperating lexicography with emerging Semantic Web methods and technologies is already underway, mainly as part of academic research, but mainstream lexicographic applications still make little use of state-of-the-art linked data (LD) resources. LT, on the other hand, follows mainly a quantitative approach along statistical and machine learning technologies. Bridging the gap between these qualitative and quantitative approaches is a huge challenge, and new solutions that combine these fields successfully stand good chances to be useful for the market.

In this context, the emerging Semantic Web, with new LD and semantic technologies, offers innovative means for information and data retrieval, knowledge management systems, and other applications for lexical data exchange and integration. However, while existing methodologies are becoming mature, they still lack sufficient refinements for data quality mechanisms, provenance methods and security issues.

There are new RandD initiatives that aim to bridge the gap between these disciplines, but only a few commercial applications. Groundbreaking projects include LIDER (http://lider-project.eu/) and LOD2 – Creating Knowledge out of Interlinked Data, which included the development of 45+ LOD software components (http://lod2.eu). Several freely available sources (and semi-open lexicographic resources) have also been developed, such as the Linguistic Linked Open Data Cloud (http://linguistic-lod.org/llod-cloud), WordNet

Semantic Web Company is a leading provider of graph-based metadata, search and analytic solutions, based in Vienna. Its expert team provides consulting and integration services for semantic data and information management, and supports customers mainly in North America, Europe and Australia, including global 500 companies. It has recently been named on KMWorld's 2017 list of the 100 Companies That Matter in Knowledge Management (after being listed also in 2015 and 2016). https://semantic-web.com

PoolParty Semantic Suite is a world-class semantic technology tool that offers sharply focused solutions for knowledge organization and content business. As a semantic middleware, PoolParty enriches information with valuable metadata and links business and content assets automatically. http://poolparty.biz (https://wordnet.princeton.edu/), BabelNet (http://babelnet.org/), or DBpedia (http://dbpedia.org). Although these sources are comprehensive and useful, as well as available in machine readable formats (often providing an API) that allow relatively easy and efficient data integration, their main drawbacks still regard the content quality and (in)completeness.

The need remains to combine such openly available sources with quality lexicographic resources, including monolingual, bilingual or multilingual dictionaries that offer comprehensive data such as precise definitions, examples of usage, and other grammatical and semantic information, among others.

The LDL4HELTA project

Linked Data Lexicography for High-End Language Technology Application (LDL4HELTA, https://ldl4.com/) attempts to deal with the issues described above by combining lexicography with LD and integrating closed data sources with open ones to develop new LT methods and tools. This project is part of the EUREKA bilateral Austria-Israel RandD framework (http://eurekanetwork.org/project/id/9898), endorsed and supported by the Austrian Research Promotion Agency and the Israeli Chief Scientist Office (Israel Innovation Authority). It is led by Semantic Web Company (SWC) and K Dictionaries (KD), with scholarly cooperation of the Austrian Academy of Sciences and Polytechnic University of Madrid.

The project brings together lexicographic resources of KD with SWC's expertise in semantic technologies for the development of new products and services, to help the international LT market meet the fast-growing demands for dedicated language-independent, language-specific and cross-language solutions. These, in turn, will enhance cross-lingual search and usage for multilingual data management and integration. This entails:

- Enhancing knowledge and technology transfer between the partners in lexical methodologies and LD and semantic technologies;
- Combining state-of-the-art lexicographic and LT resources with Semantic Web and LD mechanisms to bridge the gap between them and generate new cross-language lexical tools and services;
- Integrating existing and new tools of the partners to give way to improved enterprise-ready software and data solutions for a wider market;
- Developing new software components for to upgrade data quality.

In order to provide the above-mentioned solutions, an integrated multilingual metadata and data management approach is needed, and this is where SWC's PoolParty Semantic Suite plays a crucial role. As PoolParty follows W3C Semantic Web standards (http://w3.org/standards/semanticweb/), such as SKOS (http://w3.org/2004/02/skos/), it already incorporates language-independent-based technologies. However, as regards text analysis and extraction, the ability to process multilingual information and data is key for success – which means that such systems need to speak as many languages as possible.

The cooperation with KD in the course of LDL4HELTA will enable PoolParty Semantic Suite to continuously "learn to speak" more and more languages, and do so more precisely, by making use of KD's rich multi-language lexical content and its know-how in lexicography as a base for improved text analysis and processing.

The first goal of the LDL4HELTA project is to model and convert KD data into RDF format, make it enrichable by third-party sources, by applying the Linked Data Design Principles proposed by Tim Berners Lee (https://w3.org/DesignIssues/LinkedData.html), and to make use of a SPARQL endpoint as an API to enable complex and flexile data querying.

The second goal is to improve word sense disambiguation as regards entity extraction and semantic annotation. Several methods are combined to attain this purpose, including (i) using dictionary data (ii) using thesauri and knowledge models, (iii) making use of corpora and freely available lexical resources, and (iv) integrating users' first-choice mechanisms.

The project started in July 2015 and ends in September 2017. It is supported by an advisory board including Prof. Christian Chiarcos (Goethe University, Frankfurt), Mr. Orri Erling (Google, San Francisco), Prof. Asunción Gómez Pérez (Universidad Politécnica de Madrid), Dr. Sebastian Hellmann (Leipzig University), Prof. Alon Itai (Technion, Haifa), and Ms. Eveline Wandl-Vogt (Austrian Academy of Sciences).

Ilan Kernerman is CEO of K Dictionaries, leading strategic development and cooperation. He edits and publishes *Kernerman Dictionary News,* has co-edited conference papers and other collections, and been involved in international lexicography associations and projects, most recently Asialex president (2015-2017) and the Globalex initiative. His interests include lexicography and the interoperability with NLP and knowledge systems. http://kdictionaries.com/ilank_2015.pdf

K Dictionaries creates cross-lexical resources for 50 languages and cooperates with industry and academia partners worldwide. Based in Tel Aviv and incorporating cutting edge pedagogical and multilingual lexicography methodologies, it develops manually crafted and automatically generated linguistic data serving natural language processing technologies for human and machine use. http://kdictionaries.com