Review of the 2015 EUROLAN in Computational Linguistics

Jelena Mitrović
jmitrovic@gmail.com,
University of Belgrade
Faculty of Philology

PAPER SUBMITTED: 10 February 2016
PAPER ACCEPTED: 14 February 2016

EUROLAN Summer School (13–25th July, 2015) was the twelfth in the series of summer schools that take place every two years in Romania. Topics of these summer schools are always very "hot", as was the case with this year's topic – an area of Computational Linguistics and data and knowledge management in general – Linguistic Linked Open Data – LLOD. This year's summer school took place in the beautiful town of Sibiu, situated in the heart of Transylvania, surrounded by mountains and wonderful nature.

Twenty distinguished lecturers who are some of the biggest names in the area of Computational Linguistics, and certainly in the LLOD movement, and who significantly contributed to its development, held intensive courses during two weeks of the summer school. Mornings were mainly reserved for theoretical parts of acquiring new concepts, while the afternoons were filled with practical work, tutorials and practice. Participants came from all over Europe, but also from China and Australia. One of the benefits of this kind of learning and enhancement of professional skills is also the possibility to share experiences and to connect with colleagues from all over the world – which is definitely very important! This was the first summer school I attended and it was a great pleasure to spend two weeks with like-minded people who share my passion for linguistics and using computer technologies to process natural language.

The concept of Semantic Web was in the centre of all lectures and tutorials at this summer school. Semantic Web is a project of developing a universal medium for exchange of information using documents with meaning that computers can process on the web. The main goal of this concept is semantic interoperability of web resources and existence of an infrastructure for machine interpretation and reasoning about the contents on the web.

RDF (Resource Description Framework) is a concept that originated from the search for efficient solutions in information retrieval tasks and is one of the standards of Semantic Web. It represents a general method for conceptual description of information – semantic links between electronic resources. It consists of ordered triplets: Subject – Predicate – Object, where the Subject is the RDF URI reference of the resource we are describing, the Predicate is the RDF URI reference, i.e. the semantic relation, and the Object is the RDF URI reference, the meta-datum itself.

Key technology of the Semantic Web is also SPARQL (pronounced as the word "sparkle", a recursive acronym for SPARQL Protocol and RDF Query Language). This language was developed specifically for searching RDF databases and is a W3C standard. Other important elements of Semantic Web are XML (eXtensible Markup Language), which defines the data structure (RDF/XML), ontologies, i.e. models that represent knowledge or sets of concept definitions and relations that exist between them, OWL (Web Ontology Language) which is used for ontology publishing and sharing. All of these technologies are crucial for functioning of the LLOD paradigm.

Linked Open Data – LOD, the basis for LLOD, according to the principles set in 2006 by Tim Berners-Lee, are data that 1) Use URIs (Uniform Resource Identifiers) as names of things; 2) Use HTTP URIs links so that those names can be found 3) Provide useful information using RDF and SPARQL standards; 4) Contain links to other URIs for discovering as many things as possible; 5) Data should be open for usage via open licences.

Graphical representation of the LLOD cloud was developed following the initiative of the Open Linguistics Working Group [1] and this group is responsible for its development and maintenance, in the scope of the Open Knowledge Foundation Network (OKFN) [2]. Figure 1 shows the LLOD cloud diagram containing corpora, databases, terminological bases, dictionaries, linguistic data categories, typological databases.

DBpedia is a very important part of both LLOD and LOD paradigms. It transforms data from Wikipedia pages into RDF. It contains URIs and other metadata for each page, starting with infobox parts of Wikipedia pages. BabelNet is also a very important element of the LLOD cloud. It is a semantic network that automatically extracts data from WordNet, Open Multilingual WordNet (set of all open wordnets), Wikipedia (the largest collaborative encyclopaedia), Wikidata (the largest collaborative knowledge base), Wiktionary (the largest collaborative dictionary), OmegaWiki (medium-sized multilingual dictionary).

A lexical resource can be included in the LLOD cloud if the following requirements are fulfilled: 1) the resource has to be available through resolvable http:// (or https://) URI links; 2) the data have to be resolvable into RDF data in one of the mostly used RDF formats (RDFa, RDF/XML, Turtle, N-Triples); 3) It has to contain at least 1000 ordered triples; 4) It has to be connected through RDF links

[1] OWLG http://linguistics.okfn.org/2011/05/20/the-open-linguistics-working-group/

[2] OKFN https://okfn.org/

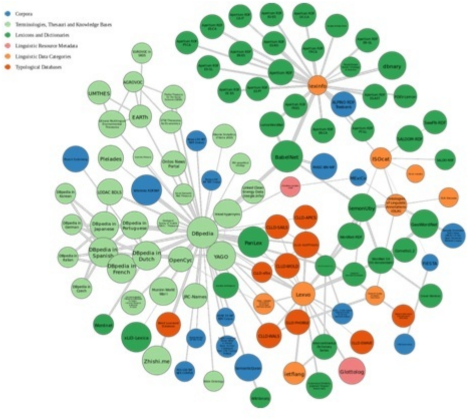

Figure 1. LLOD cloud

with a resource that is already included in the LLOD diagram, or it has to have at least 50 links to other resources; 5) It can be accessed through RDF crawling, through an RDF dump, or through a SPARQL endpoint.

At this summer school, I also had a chance to find out that the idea of accepting the LLOD technologies has been accepted in the WordNet community, that is to say, LLOD as the basic mechanism for creating links between wordnets in different languages, through the interlingual index (ILI). Accepting open licences and shared formats has led to more available data from world's wordnets. One of the projects that stemmed from that notion is the Universal wordnet whose goal is to resolve the problems of polysemy and synonymy through connecting the same or similar concepts in different languages.

The conclusion of the story about Linguistic Linked Open Data is that they are a useful solution for many purposes because they: 1) allow for integration of information – it is possible to find and combine information from various resources in an efficient way; 2) enable multilingual usage for many tools; 3) enforce dynamic publishing – data on the web are not static, different versions can be seen and errors corrected; 4) use graph based models which allow for representing any kind of a linguistic resource; 5) information retrieval is structured, e.g. we can get an answer to the question “What are the names of all Nobel Prize winners originating from France?”.

Besides gaining invaluable knowledge about the way in which many Semantic Web technologies function and witnessing examples of their usage in hands-on sessions, I also received a lot of useful suggestions as to how we could include Serbian linguistic resources and tools into the LLOD cloud, and all of that from experts in this field, some of which have invented or significantly enhanced the technologies we used. The EUROLAN School left a very positive impression on me, especially because the organizers insisted that all participants should spend as much time as possible together, which is why we had a chance to better acquaint ourselves with the lecturers and to get valuable advice. The next one in the series of these summer schools will be organized in 2017 and everyone who is interested in Computational Linguistics should attend it, maybe it will even be me again!