Invited Talk: Ontologies and Linked Open Data for Acquisition and Exploitation of Language Resources

Kiril Simov
Linguistic Modelling Deparment, IICT-BAS
Acad. G. Bonchev 25A, 1113 Sofia, Bulgaria
kivs@bultreebank.org

Recent developments in Natural Language Processing (NLP) are heading towards knowledge rich resources and technology. Integration of linguistically sound grammars, sophisticated machine learning settings and world knowledge background is possible given the availability of the appropriate resources: deep multilingual treebanks, representing detailed syntactic and semantic information; and vast quantities of world knowledge information encoded within ontologies and Linked Open Data datasets (LOD). Thus, the addition of world knowledge facts provides a substantial extension of the traditional semantic resources like WordNet, FrameNet and others. This extension comprises numerous types of Named Entities (Persons, Locations, Events, etc.), their properties (Person has a birthDate; birthPlace, etc.), relations between them (Person works for an Organization), events in which they participated (Person participated in war, etc.), and many other facts. This huge amount of structured knowledge can be considered the missing ingredient of the knowledge-based NLP of 80's and the beginning of 90's.

The integration of world knowledge within language technology is defined as an *ontology-to-text* relation comprising different language and world knowledge in a common model. We assume that the lexicon is based on the ontology, i.e. the word senses are represented by concepts, relations or instances. The problem of lexical gaps is solved by allowing the storage of not only lexica, but also free phrases. The gaps in the ontology (a missing concept for a word sense) are solved by appropriate extensions of the ontology. The mapping is partial in the sense that both elements (the lexicon and the ontology) are artefacts and thus — they are never complete. The integration of the interlinked ontology and lexicon with the grammar theory, on the other hand, requires some additional and non-trivial reasoning over the world knowledge. We will discuss phenomena like selectional constraints, metonymy, regular polysemy, bridging relations, which live in the intersective areas between world facts and their language reflection. Thus, the actual text annotation on the basis of ontology-to-text relation requires the explication of additional knowledge like co-occurrence of conceptual information, discourse structure, etc.

Such knowledge is mainly present in deeply processed language resources like HPSG-based (LFG-based) treebanks (RedWoods treebank, DeepBank, and others). The inherent characteristics of these language resources is their dynamic nature. They are constructed simultaneously with the development of a deep grammar in the corresponding linguistic formalism. The grammar is used to produce all potential analyses of the sentences within the treebank. The correct analyses are selected manually on the base of linguistic discriminators which would determine the correct linguistic production. The annotation process of the sentences provides feedback for the grammar writer to update the grammar. The life cycle of a dynamic language resource can be naturally supported by the semantic technology behind the ontology and LOD - modeling the grammatical knowledge as well as the annotation knowledge; supporting the annotation process; reclassification after changes within the grammar; querying the available resources; exploitation in real applications. The addition of a LOD component to the system would facilitate the exchange of language resources created in this way and would support the access to the existing resources on the web.

Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, page 1,
Sofia, Bulgaria, 8-9 August 2013.