BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **18**, No 1
Sofia • 2018 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2018-0008

Special Thematic Section on Semantic Models for Natural Language Processing

Preface

Kiril Simov, Petya Osenova

Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria

***Abstract*:** *With the availability of large language data online, cross-linked lexical resources* (*such as BabelNet, Predicate Matrix and UBY*) *and semantically annotated corpora* (*SemCor, OntoNotes, etc.*)*, more and more applications in Natural Language Processing* (*NLP*) *have started to exploit various semantic models. The semantic models have been created on the base of LSA, clustering, word embeddings, deep learning, neural networks, etc., and abstract logical forms, such as Minimal Recursion Semantics* (*MRS*) *or Abstract Meaning Representation* (*AMR*)*, etc.*

Additionally, the Linguistic Linked Open Data Cloud has been initiated (*LLOD Cloud*) *which interlinks linguistic data for improving the tasks of NLP. This cloud has been expanding enormously for the last four-five years. It includes corpora, lexicons, thesauri, knowledge bases of various kinds, organized around appropriate ontologies, such as LEMON. The semantic models behind the data organization as well as the representation of the semantic resources themselves are a challenge to the NLP community.*

The NLP applications that extensively rely on the above discussed models include Machine Translation, Information Extraction, Question Answering, Text Simplification, etc.

The idea behind this special topic was to gather contributions on the creation, maintenance and usage of semantic models for specific Natural Language Processing (NLP) tasks. We were interested in both – supervised and unsupervised approaches as well as in models that are predefined in structured resources or extracted on-the-fly from unstructured data. The semantic models include various topics of interest, such as: combining syntagmatic (corpus-based) and paradigmatic (lexicon-based) relations into semantic models; approaches to modeling semantically similarity and relatedness; design and application of distributional semantics models; graph-based semantic methods shallow and deep semantic architectures, based on neural networks; integrating non-semantic linguistic knowledge into semantic models; the applicability of the various logical semantic representations for NLP tasks; linking of linguistic resources through appropriate semantic models.

Six papers have been finally accepted as contributions to the topic of Semantic Models for NLP. They cover various related issues ranging from overviews and linking models to language-specific and task-specific applications.

The paper *Neural Network Models for Word Sense Disambiguation: An Overview* (Alexander Popov) focuses on the recent advances in neural language models that have led to methods for the effective distributed representation of linguistic data (word embeddings, context embeddings, sense embeddings) and architectures for Word Sense Disambiguation, including Recurrent neural networks with Long Short-Term Memory cells and Gated Recurrent Units.

The paper *Linking Datasets Using Semantic Textual Similarity* (John P. McCrae, Paul Buitelaar) presents an evaluation of some metrics (string similarity metrics, Smoothed Jaccard, Word Alignment metrics, structural similarity metrics) that have performed well in recent semantic textual similarity evaluations and then apply these to linking existing datasets.

The paper *A Semantic Multi-Field Clinical Search for Patient Medical Records* (E Umamaheswari Vasanthakumar, Francis Bond) describes the process of matching medical records of patients to the relevant clinical practice guidelines (CPGs). A basic ontology based search engine that supports multilevel Patient Medical Records search and rank was designed.

The paper *2L-APD: A Two-Level Plagiarism Detection System for Arabic Documents* (El Moatez Billah Nagoudi, Ahmed Khorsi, Hadda Cherroun, Didier Schwab) reports on a plagiarism detection system for Arabic based on two layers of assessment: fingerprinting and word embedding. The system was evaluated on three types of plagiarism: simple, shuffled words and phrases, and with more intelligent strategies.

The paper *Graph-Based Complex Representation in Inter-Sentence Relation Recognition in Polish Texts* (Arkadiusz Janz, Paweł K¸edzia, Maciej Piasecki) introduced a supervised approach for the recognition of Cross-Document Structure Theory (CST) relations in Polish texts. It relies on a graph-based representation of the sentences.

The paper *An Automatically Generated Annotated Corpus for Albanian Named Entity Recognition* (Klesti Hoxha, Artur Baxhaku) reports on the first automatically generated corpus with Named Entities for Albanian using newsmedia texts and the linking information from the Albanian Wikipedia.

Guest Editors:
Kiril Simov and Petya Osenova

***Reviewers*:** *Galia Angelova* (*IICT-BAS*); *Antonio Branco* (*University of Lisbon*); *Francis Bond* (*Nanyang Technological University*); *Gosse Bouma* (*University of Groningen*); *Aljoscha Burchardt* (*Deutsches Forschungszentrum für Künstliche Intelligenz*); *Nicoletta Calzolari* (*Institute for Computational Linguistics "A. Zampolli"*); *Ann Copestake* (*University of Cambridge*); *Thierry Declerck* (*Deutsches Forschungszentrum für Künstliche Intelligenz*); *Christiane D. Fellbaum* (*Princeton University*); *Sandra Kübler* (*Indiana University*); *Allesandro Lenci* (*University of Pisa*); *John Mccrae* (*National University of Ireland Galway*); *Preslav Nakov* (*Qatar Computing Research Institute, Qatar Foundation*); *Maciej Piasecki* (*Wroclaw University of Science and Technology*); *Piek Vossen* (*Vreje Universiteit, Amsterdam*).

Received 6.03.2018; *Accepted 6.03.2018*