Invited Talk: Domain-adaptation of Natural Language Processing Tools for RE

Tejaswini Deoskar
Institute for Logic, Language and Computation (ILLC), University of Amsterdam
t.deoskar@uva.nl

Natural language processing tools like part-of-speech taggers and parsers are being used in a variety of applications involving natural language, including RE. Such tools, based on statistical models of language, are learnt via supervised machine learning algorithms from human-annotated data. Due to their dependence on annotated data, which is limited in size and genre, these models have a fall in performance for words or constructions not encountered in the annotated data, as well as for genres or domains of language different from the supervised training data. This talk will present Tejaswini Deoskar's work on semi-supervised learning, where a model initially trained on supervised data is further improved by using unannotated data, available in much larger quantities. Such semi-supervised training improves performance over low-frequency words and constructions, i.e. those in the long tail of language use, and may also be used to adapt supervised NLP models to perform better over new domains of text such as those used in RE documents.

Biography of Tejaswini Deoskar Tejaswini Deoskar is assistant professor at the Institute for Logic, Language and Computation (ILLC) at the University of Amsterdam. Her research focuses on probabilistic learning techniques and probabilitistic models for natural language. She is interested in the general problem of learning the syntax and semantics of natural language using "data-driven" methods. This "data" usually consists of large collections of language usage (most commonly, text). In particular, she worked on "semi-supervised" learning techniques for language, where the data is a combination of bare text, plus text that is annotated with extra syntactic or semantic information. She is also especially interested in the class of grammars called "strongly-lexicalised" grammars. She worked on semi-supervised learning for such grammars, especially the grammar formalism Combinatory Categorial Grammar (CCG).