Enabling Technology for Code-Switching Data
Thamar Solorio, The University of Houston

Most of the natural language processing (NLP) technology being developed today assumes a one-input, one-language setting. This is also true for recent efforts into multilingual approaches: while this technology can handle multiple languages without needing major adaptation changes, these models still expect to process one language at a time. Performance of state-of-the-art systems degrades rapidly when the input contains mixed-language data. Over the last decade or so, the RiTUAL (Research in Text Understanding and Analysis of Language) Lab, which I direct, has been working on enabling technology for mixed-language data, with the goal of bridging the performance gap between the automated processing of monolingual text and that of mixed-language data. To achieve this goal, we have devoted efforts into developing linguistic resources, organizing international shared tasks and workshops to raise awareness of the many interesting NLP challenges in the computational processing of mixed language data, and developing our own technology for low level syntactic analysis tasks, as well as other downstream tasks, such as named entity recognition. During this talk I will discuss the progress we and others have achieved in recent years, I will present the approaches we have found to be more successful in handling mixed-language data, and briefly discuss empirical results. I will end the presentation with an open discussion of the outstanding challenges in this line of research.