IU BLOOMINGTON

EMERGING AREAS OF RESEARCH

Abstract Template -- Due June 30, 2017

Title of initiative to be proposed:

Human-in-the-loop Natural Language Processing for Low Resource Languages Spoken in Indiana

Name of lead PI, with title, department/school:

Professor Sandra Kübler, Department of Linguistics, COAS

Key team member names and departments/schools (up to 10 names):

Matthew Anderson, School of Informatics and Computing
Kelly Harper Berkson, Department of Linguistics, COAS
Timur Gilmanov, School of Informatics and Computing
Donald Williamson, School of Informatics and Computing

Description of area to be proposed. What constitutes this area of research or creative activity as emerging? (Word limit=500)

Indiana now has one of the largest Burmese refugee populations in the United States, with more than 13,000 people living in Indianapolis alone and nearly 20,000 in total. These refugees come from poverty abroad into poverty in Indiana, and access to basic services is complicated by language barriers. The language barrier is daily manifested in Indianapolis-area hospitals, which have near-constant needs for Laiholh (Chin) and Burmese translation capability. We will develop novel methodologies for automatic speech recognition (ASR) systems and machine translation (MT) (potentially speech synthesis) for low-resource languages and develop deployable systems for medical professionals and first responders supporting key low-resource languages spoken in Indiana including Burmese, Laiholh, Karen, Karenni, Kachin, and Mon.

Standard approaches to ASR and MT, based on machine learning techniques, require large sets of language examples accompanied by correct transcriptions or translations. Thus for every language for which we need to develop these technologies we are forced to adjust analytical techniques based on language characteristics and to invest considerable effort (in the range of hundreds of thousands of hours) into creating data sets with which to train the machine learners. Acoustic models typically need thousands of hours of transcribed recordings, and machine translation requires hundreds of thousands of translated sentences. A trained linguist can need between 100 and 1000 seconds to transcribe 10 seconds of speaker recordings. Translating is only marginally faster. The problem is compounded if we need speech tools for specific language domains, such as the medical field, which require annotated data for domain-specific examples. Such efforts are deeply valuable, but are only feasible if there is a commercial interest in a language. For a wide range of languages, then—even those with millions of speakers—it is simply impossible to develop ASR and MT resources given current technologies and resources.

We propose to create a center for developing human-in-the-loop approaches to natural language processing applications for low-resource languages, with a specific focus on languages spoken by refugees in Indiana. We will create approaches that combine novel machine learning techniques for small data sets (e.g., one-shot learning, active learning) with novel methods for graph computing that allow us to integrate syntactic and semantic graphs (i.e., knowledge about sentence structure and meaning), along with knowledge- or grammar-driven approaches that extract knowledge/rules about a language from linguistically naïve speakers. We have access to native speakers, including 30 IU students, who can provide information about how the language works if we ask the right questions, and we have access to linguists who can provide knowledge about how to exploit similarities between languages. We will develop novel methods for partially automating such linguistic fieldwork so that the involvement of trained specialists (linguists and machine learning experts) is kept at a minimum.

The team consists of professionals with expertise in audio processing, speech recognition, fieldwork and linguistics, computational linguistics, machine learning, high performance computing graphs, and in Burmese and Laiholh. Potential funding can come from NSF, ONR, iARPA, DARPA, and from companies such as Amazon, Google, Microsoft, or Nuance.

Please submit to earprogram@indiana.edu