Language Technologies Institute

Thesis Defense

GHC 6501| Wednesday, October 7, 2015| 11:00 am

Machine Translation 4 Microblogs

Wang Ling

Abstract

The emergence of social media caused a drastic change in the way information is published. In contrast to previous eras in which the written word was more dominated by formal registers, the possibility for people with different backgrounds to publish information has caused non-standard style, formality, content, genre and topic to be present in written documents. One source of such data are posts in microblogs and social networks, such as Twitter, Facebook and Sina Weibo. The people that publish these documents are not professionals, yet the information published can be leveraged for many ends. However, current NLP tasks, such as Part-of-Speech Tagging perform poorly in the presence of this type of data, since they are modelled using traditional assumptions and trained on existing edited data. One problem is the lack of annotated datasets in this domain. One such assumption is of spelling homogeneity, where we assume that there is only one way to spell tomorrow, whereas in microblogs, this word can be abbreviated to tmrw (among many other options) or spelled erroneously as tomorow.

In this thesis, we address the challenge of NLP on the domain of informal online texts, with emphasis on Machine Translation. This thesis makes the following contributions in this respect. (1) We present an automatic method to extract such data automatically from microblog posts, by exploring the fact that many bilingual users post translations of their own posts. (2) We propose a compositional model for word understanding based only on their character sequence, breaking the assumption that different word types are independent. This allows the model to generalize better on morphologically rich languages and the orthographically creative language used in microblogs. (3) Finally, we show improvements on several NLP tasks, both syntactically and semantically oriented, using both the crawled data and proposed character-based models. Ultimately, these are combined into a state-of-the-art MT system in this domain.

Bio: Wang Ling is a student of the dual Ph.D. program in Computer Science between Carnegie Mellon University and Instituto Superior Técnico, where he also received his master degree in 2009. His Ph.D. work focuses on Machine Translation and Deep Learning.

http://www.cs.cmu.edu/~lingwang/papers/thesis.pdf

COMMITTEE:

Alan W Black,
LTI, CMU (Chair)

Isabel Trancoso,
IST/INESC-ID

Luísa Coheur,
IST/INESC-ID

Chris Dyer,
LTI, CMU

Noah Smith,
LTI, CMU

Chris Callison-Burch
U Penn

Mário Figueiredo,
IST/IT