Title: Using Freebase, an Automatically Generated Dictionary, and a Classifier to Identify a Person’s Profession in Tweets

Algorithms for classifying pre-tagged person entities in tweets into one of eight profession categories are presented. A classifier using a semi-supervised learning algorithm that takes into consideration the local context surrounding the entity in the tweet, hash tag information, and topic signature scores is described. In addition to the classifier, this research investigates two dictionaries containing the professions of persons. These two dictionaries are used in their own classification algorithms which are independent of the classifier. The method for creating the first dictionary dynamically from the web and the algorithm that accesses this dictionary to classify a person into one of the eight profession categories are explained next. The second dictionary is freebase, an openly available online database that is maintained by its online community. The algorithm that uses freebase for classifying a person into one of the eight professions is described. The results also show that classifications made using the automated constructed dictionary, freebase, or the classifier are all moderately successful. The results also show that classifications made with the automated constructed person dictionary are slightly more accurate than classifications made using freebase. Various hybrid methods, combining the classifier and the two dictionaries are also explained. The results of those hybrid methods show significant improvement over any of the individual methods.

Major: Computer Science

Educational Career:
Bachelor’s of Computer Science, BS, 2005, Brigham Young University

Committee in Charge:
Fernando Gomez, Chair, EECS
Damian Dechev, EECS
Marshall Tappen, EECS

Approved for distribution by Fernando Gomez, Committee Chair, on June 18, 2013.

The public is welcome to attend.