Handling and Mining Linguistic Variation in UGC (invited talk)

Leon Derczynski
Department of Computer Science
The University of Sheffield
leon@dcs.shef.ac.uk

1 Abstract

Across its many forms, user-generated content (UGC) acts as a sample of all human discourse. It has been harder to process this type of text with traditional tools. People have even looked at normalising this text to look like the data that traditional tools are used to. This talk examines the kind of variation we see in user-generated content, and contrary to the trend of normalisation, not only presents methods for coping with the noise without changing it, but also goes on to explain the many kinds of latent information expressed by the stable, consistent linguistic variation seen across society and the internet.

2 Biography

Dr. Leon Derczynski is a Research Associate in the University of Sheffield's NLP group. He has worked on the forefront of social media language processing since 2012, developing and releasing tools to the community and in GATE, and is applying this by current work on multiple EU projects centred around detecting and predicting rumours and false claims on the web.

Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects, page 10,
Hissar, Bulgaria, September 10, 2015.