MWEs and Topic Modelling: Enhancing Machine Learning with Linguistics

Timothy Baldwin
University of Melbourne, Australia
tim@csse.unimelb.edu.au

Abstract

Topic modelling is a popular approach to joint clustering of documents and terms, e.g. via Latent Dirichlet Allocation. The standard document representation in topic modelling is a bag of unigrams, ignoring both macro-level document structure and micro-level constituent structure. In this talk, I will discuss recent work on consolidating the micro-level document representation with multiword expressions, and present experimental results which demonstrate that linguistically-richer document representations enhance topic modelling.

Biography

Tim Baldwin is an Associate Professor and Deputy Head of the Department of Computer Science and Software Engineering, University of Melbourne and a contributed research staff member of the NICTA Victoria Research Laboratories. He has previously held visiting positions at the University of Washington, University of Tokyo, University of Saarland, and NTT Communication Science Laboratories. His research interests cover topics including deep linguistic processing, multiword expressions, deep lexical acquisition, computer-assisted language learning, information extraction and web mining, with a particular interest in the interface between computational and theoretical linguistics. Current projects include web user forum mining, information personalisation in museum contexts, biomedical text mining, online linguistic exploration, and intelligent interfaces for Japanese language learners. He is President of the Australasian Language Technology Association in 2011-2012.

Tim completed a BSc(CS/Maths) and BA(Linguistics/Japanese) at the University of Melbourne in 1995, and an MEng(CS) and PhD(CS) at the Tokyo Institute of Technology in 1998 and 2001, respectively. Prior to commencing his current position at the University of Melbourne, he was a Senior Research Engineer at the Center for the Study of Language and Information, Stanford University (2001-2004).

Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), page 1,
Portland, Oregon, USA, 23 June 2011.