Augmenting concept definition in gloss vector semantic relatedness measure using Wikipedia articles

ABSTRACT

Semantic relatedness measures are widely used in text mining and information retrieval applications. Considering these automated measures, in this research paper we attempt to improve Gloss Vector relatedness measure for more accurate estimation of relatedness between two given concepts. Generally, this measure, by constructing concepts definitions (Glosses) from a thesaurus, tries to find the angle between the conceptsô gloss vectors for the calculation of relatedness. Nonetheless, this definition construction task is challenging as thesauruses do not provide full coverage of expressive definitions for the particularly specialized concepts. By employing Wikipedia articles and other external resources, we aim at augmenting these conceptsô definitions. Applying both definition types to the biomedical domain, using MEDLINE as corpus, UMLS as the default thesaurus, and a reference standard of 68 concept pairs manually rated for relatedness, we show exploiting available resources on the Web would have positive impact on final measurement of semantic relatedness.

Keyword: Semantic relatedness; Biomedical text mining; Web mining; Bioinformatics; UMLS; MEDLINE; Wikipedia; Natural language processing