Mittwoch
08.03.2017
16:30 – 17:00
B4 1, 0.24

Exploring the impact of transparency and productivity of multiword term constituents on single-word term identification

Anna Hätty
Robert Bosch GmbH
anna.haetty@de.bosch.com

Michael Dorna
Robert Bosch GmbH
michael.dorna@de.bosch.com

Terms are expressions that characterize specialized domains. They comprise both single- (SWT) and multiword terms (MWT). For the scoring of multiword expressions as terms their components are often taken into account (e.g. Zhang, 2012), and transparency plays a role for translation and synonym extraction of MWTs. Our goal is to exploit the relation of MWTs to their constituents in order to identify SWTs among the constituents. SWTs are often less specific than MWTs and it is harder to score them for termhood. We hypothesize that SWTs which are frequently found in diverse complex terms and which contribute to the meaning of the MWTs are more likely to be terms as well. Therefore we approach this problem by taking the constituents of multiword terms as term candidates. We then investigate the influence of their productivity and transparency within these MWTs on the prediction of termhood. We use the ACL RD-TEC (Zadeh and Handschuh, 2014), a corpus for the evaluation of term extraction in the field of Computational Linguistics. We address transparency with a vector space model by computing the similarity of compound and constituent vectors. We show that transparency variance for highly productive heads influences their prediction as terms. We investigate the interplay of productivity, transparency, variance of transparency in constituent families and frequency by training a classification model with those features. Finally, we compare this approach with the modified C-value for SWTs by Barrón-Cedeño et al. (2009).

References: • Barrón-Cedeño, A., Sierra, G., Drouin, P., & Ananiadou, S (2009). An Improved Automatic Term Recognition Method for Spanish. CICLing '09 • Zadeh, B. Q., & Handschuh, S. (2014). The ACL RD-TEC: a dataset for benchmarking terminology extraction and classification in computational linguistics. CompuTerm '14 • Zhang, C., Niu, Z., Jiang, P., & Fu, H. (2012). Domain-specific term extraction from free texts. FSKD '12, 1290-1293, IEEE.