A large corpus automatically annotated with semantic role information

Mittwoch
08.03.2017
15:45 – 16:30
B4 1, Foyer

Asad Sayeed
Universität des Saarlandes
asayeed@coli.uni-saarland.de

Pavel Shkadzko
Universität des Saarlandes
p.shkadzko@gmail.com

Vera Demberg
Universität des Saarlandes
vera@coli.uni-saarland.de

We present Rollenwechsel-English (RW-eng), large, automatically labeled corpus based on the ukWaC web crawl corpus and the British National Corpus (BNC). Our automatic annotations contain predicate-argument relation information at a sentence level. The basic annonation is performed by a combination of semantic role labelling using the SENNA SRL tool and MALT dependency parsing. SENNA provides PropBank-style role labelling over whole phrases through a sequence-labelling model. We use MALT parses to identify predicate-head relations within the text spans found by SENNA.

The large size of this corpus (approx. 78 million sentences) makes it useful for distributional semantic modeling, in which semantic relations are required, but human annotation at scale is unrealistic to obtain. It has already been successfully used in large-scale models of thematic fit/selectional preferences, both count-based and neural. The presence of both whole phrases and identified heads allows for richer semantic modelling applications.

The XML-formatted, UTF8-compliant corpus lists every automatically-identified predicate per sentence that is found in the source corpora and with these predicates, each role-filling phrase, including the automatically-discoved head. Head-finding is performed by a cascading series of heuristics, and the found heads are listed with the heuristic used to identify them. RW-EN is available at `http://rollen.mmci.uni-saarland.de/RW-eng/`.

References: • Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P. (2011): Natural language processing (almost) from scratch. *JMLR*. • Sayeed, A., Demberg, V., Shkadzko, P. (2015): An exploration of semantic features in an unsupervised thematic fit evaluation framework. *IJCoL*: 1(1), 25–40. • Tilk, O., Demberg, V., Sayeed, A., Klakow, D., Thater, S. (2016): Event participant modelling with neural networks. *EMNLP*.