Exploiting BabelNet for Multilingual Biomedical Synonym Expansion

Simon Clematide, Martin Davtyan, Fabio Rinaldi, and Dietrich Rebholz-Schuhmann

University of Zurich, Institute of Computational Linguistics, Binzmuehlestr. 14, 8050 Zurich, Switzerland
{siclemat,davtyan,rinaldi,rebholz}@cl.uzh.ch

Our challenge contribution for CLEF-ER consists in providing annotations for all three corpora of the challenge (Medline, EMEA, Patents) for the languages French and German. The objective of these experiments is to verify whether a general multilingual ontological resource as BabelNet (http://babelnet.org) can be used to substantially enrich the terminology provided by the challenge organizers. In order to reach this goal we also applied methods to create morphological variants of all terms found in the original terminology and the new ones derived from BabelNet. BabelNet contains connections between Wikipedia pages in different languages. For every English concept there are, (1) cross-lingual links to pages about the concept (e.g., English *Dodecanol* links to a French page *Dodécan-1-ol*), (2) mono-lingual redirections to alias pages (e.g., French *Alcool laurylique* redirects to *Dodécan-1-ol*), (3) translations of the Wikipedia concepts and WordNet translations of the concept (e.g., the English *Dodecanol* translates to the French concept *dodécanol*). All synonym candidates extracted from BabelNet via the English terms from the provided terminology of the challenge were post-processed. We removed Wikipedia topic markers of ambiguous names and lowercased the resulting terms. This resulted in additional 42,000 German and 38,000 French synonym candidates. For these (multi word) terms, also morphological variants were built: French variants by our grammar-based generator (see http://www.zora.uzh.ch/58113), German variants by applying the commercial morphological analyzer GERTWOL (http://www.lingsoft.fi/analysis) to the texts from the parallel corpora and terminology entries. If two different words have the same lemma according to GERTWOL, we treat these words as morphological variants. In order to reduce the noise introduced by the semantic and morphological term expansion, we applied the term candidates to all parallel corpora of the challenge using the Ontogeny term matcher (http://ontogene.org). We kept only those candidates that had at least 1 corresponding Concept Unique Identifier in the parallel sentence from the English silver standard corpus (SSC) provided by the challenge organizers. For the final named entity recognition step performed again by the OntoGene term matcher, we combined the newly created terminology and the originally provided one (including morphological variants produced as described above). The quality of our terminology enrichment depends crucially upon the quality of the BabelNet synonym relations.