Multilingual Biomedical Dictionary

Philipp Daumke a Kornél Markó a Michael Poprat b Stefan Schulz a

Freiburg University Hospital, Department of Medical Informatics (http://www.imbi.uni-freiburg.de/medinf)

Jena University, Computational Linguistics Research Group (http://www.coling.uni-jena.de)

Abstract

We present a unique technique to create a multilingual biomedical dictionary, based on a methodology called Morpho-Semantic indexing. Our approach closes a gap caused by the absence of free available multilingual medical dictionaries and the lack of accuracy of non-medical electronic translation tools. We first explain the underlying technology followed by a description of the dictionary interface, which makes use of a multilingual subword thesaurus and of statistical information from a domain-specific, multilingual corpus.

Morpho-Semantic Indexing

To cope with various morphological processes in different languages we developed a term normalization methodology, called Morpho-Semantic Indexing (henceforth, MSI). MSI uses a special type of dictionary, with entries consisting of subwords, i.e. semantically minimal units. Subwords are grouped into equivalence classes (represented by Morpheme identifiers (MIDs)) which capture intralingual as well as interlingual synonymy. A morphosyntactic parser extracts subwords from texts and assigns MIDs in a three step procedure MSI (cf. Figure 1). The result is a morpho-semantically normalized expression in a language independent representation. Its usefulness for cross-language indexing and retrieval has already been proven.[1]

Biomedical Dictionary

Firstly, we use domain and language specific corpora to create a large list of surface words, word bigrams and trigrams (henceforth, target words (TW)) containing their frequencies within these corpora. All target words are subsequently translated to a set of MIDs (invoking the MSI procedure as described above). This data is encoded in a table Dict, each record being a quadrupel (TW, freq(TW), MSI(TW), lang). A user can send a query to the dictionary via a web interface1. Again, this query is firstly altered to a set of corresponding MIDs. This set is now compared to MSI(TW) in Dict at which all matching records are returned as an unordered output list. The concluding sorting algorithm takes into account the following criteria: 1. all records in the same language belong to one group 2. all records proving to be orthographic variants of the same word (by performing simple character comparisons) are subsumed in subgroups 3. The order of subgroups and single records within subgroups is determined by the frequency data of these records.

Conclusion, Future Work

Our approach provides a comprehensive amount of entries in six different languages. As the principle object of our MSI methodology aims for enhancement in CLIR we abstain from offering any additional lexical information at this point in time. Future work will focus to achieve interlingual coherence in our subword lexicon.

References
