Poster K-6

Construction and Evaluation of an Abbreviation Dictionary for Normalizing Protein Names

Authors:
Joon-Ho Lim (*ETRI*)
Jaesoo Lim (*ETRI*)
Hyunchul Jang (*ETRI*)
Soo-Jun Park (*ETRI*)

Short Abstract: To normalize protein names to a database, it is necessary to disambiguate an abbreviation to definition. To construct an abbreviation dictionary, we collect 2920541 PubMed abstracts, and extract abbreviation-definition pairs using Schwartz' algorithm. The constructed dictionary has 25844 pairs, and 13509 abbreviation names. An experimental result shows 82% accuracy.

Long Abstract:

Recently, many researchers try to construct protein-protein interaction networks using text mining techniques such as named entity recognition and relational event extraction. However, to construct an accurate network from texts, recognized named entities should be normalized. For example, named entities such as "Amyloid beta", "Amyloid-beta peptide", and "Abeta" are semantically identical, so they should be represented as a single node in a constructed network. For normalization, many researches have been tried to map protein names into a protein database such as GeneDB, Swiss-Prot, DIP, and etc. [1]. However, in the text, there are many and various abbreviations, which INTOt exist in synonym list of protein databases. Therefore, to normalize abbreviated names to database, it is necessary to disambiguate an abbreviation (short-form) to its definition (long-form). In this abstract, we construct an abbreviation dictionary to disambiguate abbreviations, and evaluate it.

To construct an abbreviation dictionary, we collect PubMed abstracts, and then, extract abbreviation-definition pairs. (1) We collect 2,920,541 PubMed articles published between 2001 and 2005. This large size of article set can represent various orthographic variants of abbreviations such as "Ang-II", "Ang-2", "AngII", etc. (2) From the collected abstracts, we extract abbreviation-definition pairs using a simple and accurate algorithm [2]. It definitionpatterns from text, and determines correct boundaries of definition strings. Because extracted definition strings contain various inflectional and morphological variants, we combine similar definition strings to one entry using the Lexical Variants Generation (LVG) (see http://0-lexsrv3.nlm.nih.gov.csulib.ctstateu.edu/LexSysGroup/Projects/lvg/current/). (3) To exclude rare abbreviation-definition pairs, we cut off pairs which occur less than 10 times in the collection. The result abbreviation dictionary has 25,844 abbreviation-definition pairs, and 13,509 abbreviation names. The average number of definition per abbreviation is 1.19.

To evaluate the dictionary, we try to apply the dictionary to the test corpus, which are 6,000 articles of INTOs disease and Diabetes Mellitus. Biologists manually annotated named entities and abbreviations to each article. There are 8,669 abbreviation-definition pairs in the test corpus. We assign the most frequent definition to every abbreviation in the corpus, the

accuracy is 82.33%. As correct cases, 20.9% is exact matching with raw corpus, 42.5% is matching via LVG normalization, and 18.8% is correct matching but not different string such as "amyloid beta" and "beta amyloid". As incorrect cases, 9.05% of abbreviation INTOt exist in the dictionary, and incorrect abbreviation disambiguation is 8.60%.

[1] Lynette Hirschman, Marc Colosimo, Alexander Morgan and Alexander Yeh. (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. BMC Bioinformatics 6 S11.
[2] A. Schwartz and M. Hearst. (2003). A simple algorithm for identifying abbreviation definitions in biomedical texts. In Proceedings of the Pacific Symposium on Biocomputing (PSB).