

A Noun-Phrase Extraction System for Neuroscience Literature

Lei, Xiaojun¹, Yang, Xiaohong²

LangPower Computing, Inc., ¹Durham, NC, USA, and ²Tacoma, WA, USA

Conventional search engines employ keyword-based indexing in their database. However, words are not appropriate representations for texts which results in a low precision and high recall for searches. Although noun phrases represent concepts conveyed in texts, the difficulty in automatically extracting noun phrases with high coverage and accuracy from texts hampers the power of search engines to utilize phrase-based indexing. In specific domains such as neuroscience, documents are different from those of general fields in the following categories: 1) existence of multiple pre-modifiers of a head-noun; 2) complex syntactic structures such as multiple layers of embedded noun phrase as modifiers; 3) the majority of lexicon in phrases is domain-specific. For example, a noun phrase, "Solution conformation of the antibody-bound tyrosine phosphorylation site of the nicotinic acetylcholine receptor beta-subunit in its phosphorylated and nonphosphorylated states," has four-layers of post-modifiers that recursively modify the preceding components. To address the need of automatic indexing in the information retrieval for neuroscience research, we conducted a pilot study to automatically extract noun phrases from 100 publicly available neuroscience articles followed by parsing them into pre-modifier, head-noun, and post-modifier structure. This study was intended to test the hypothesis that adding domain-specific terms to the existing general English lexicon and adding domain-specific grammatical rules enable the noun-extraction system to process neuroscience texts. We first developed a noun phrase extraction system with a core proprietary parser for general English texts. It uses a rule-based rather than statistic-based parser. The parser consists of three thousands grammatical rules and fifty thousands common English words in a dictionary. Secondly, we extended the lexicon with additional domain-specific words using UMLS's SPECIALIST LEXICON as resource along with domain-specific grammatical rules that existed in the dataset. Third, this noun-phrase extraction system was tested on the dataset. The procedure of extracting noun phrases was: (1) input a document in plain text; (2) parse each sentence into its syntactic structure (a parse tree); (3) extract noun phrase nodes in the parse tree; (4) store noun phrases with their syntactic structures (parse sub-trees) in a file. (5) load the data into existing database tables for evaluation. The result showed that about 75% noun phrases were extracted correctly from the sample articles and 63% of these noun phrases were parsed to correct syntactic structures. Among the remaining 25% noun phrases that could not be parsed correctly due to the incompleteness of grammatical rules, 90% of them have been partially parsed, that is, some parts of noun phrases have been processed. These failed sentences and their parsing outputs were stored separately and could be manually processed by human. Among the 12% of 75% noun phrases that could be parsed to correct syntactic structures, the multiple-layer modifiers have been parsed to point to wrong components due to ambiguities in syntactic parsing. These ambiguities could be resolved by elaborating syntactic features in the lexicon. The study suggests that adding grammatical rules, adding biomedical terms, and elaborating syntactic features in the lexicon have the potential to enable the noun-phrase extraction system to provide phrase-based indexes for search engines in neuroscience.