Improving named entity recognition accuracy for gene and protein in biomedical text literature

ABSTRACT

The task of recognising biomedical named entities in natural language documents called biomedical Named Entity Recognition (NER) is the focus of many researchers due to complex nature of such texts. This complexity includes the issues of character-level, word-level and word order variations. In this study, an approach for recognising gene and protein names that handles the above issues is proposed. Similar to the previous related works, our approach is based on the assumption that a named entity occurs within a noun group. The strength of our proposed approach lies on a Statistical Character-based Syntax Similarity (SCSS) algorithm which measures similarity between the extracted candidates and the well-known biomedical named entities from the GENIA V3.0 corpus. The proposed approach is evaluated and results are satisfied. For recognitions of both gene and protein names, we achieved 97.2% for precision (P), 95.2% for recall (R), and 96.1 for F-measure. While for protein names recognition we gained 98.1% for P, 97.5% for R and 97.7 for F-measure.

Keyword: Biomedical; Information extraction; Named entity recognition; Natural language processing; NER