Poster K-17

Effects of Head Noun Features on Statistical Named Entity Recognition in Biomedical Literatures

Authors:
Jaesoo Lim (*Electronics and Telecommunications Research Institute*)
Hyunchul Jang (*Electronics and Telecommunications Research Institute*)
Joon-Ho Lim (*Electronics and Telecommunications Research Institute*)
Soo-Jun Park (*Electronics and Telecommunications Research Institute*)

Short Abstract: We suggest head noun features with positional information instead of using chunk tag features for biomedical named entity recognition (NER). We evaluated our statistical NER system on the GENIA corpus. The system using head nouns performed 1.7%p of f-scores better than the system using chunk tags.

Long Abstract:

A text mining from biomedical literatures consists of many subtasks, such as named entity recognition (NER), relational event extraction, semantic annotation, knowledge representation and more. The NER is one of the most necessary steps for further text mining tasks. The shared task from the GENIA project showed that statistical methods are widely used for NER tasks [1]. In this abstract, we will show the effects of head noun features for statistical NER in biomedical literatures.

A noun phrase (NP) is a phrase whose head is a noun or a pronoun, optionally accompanied by a set of modifiers, and a head noun is the modified word. In the noun phrases, we can find the head nouns by simple rules using Collins' head-rules [2]. A biomedical named entity is a kind of noun phrase. So the head nouns within NPs provide rich syntactic information for identifying biomedical named entities. After chunking sentences, we find head nouns in NPs, and take them as features for statistical NER models. In addition to the lexical term of head noun, we discriminate features by adding three types of positional information; before, after and itself. For example, in noun phrase "recombinant HBxAg protein", "protein" is the head noun of the phrase. So, both "recombinant" and "HBxAg" get "proteinbefore" feature, and "protein" gets "proteinitself" feature. If there exists another word after "protein", it will get "proteinafter" feature.

For the baseline system with Maximum Entropy modeling, we choose six contextual features. Part-of-speeches, previously predicted semantic categories, words, prefixes, suffixes and word shapes are they. These contextual features are extracted from left two word tokens, right two word tokens, and the target word, except for the previously predicted semantic category features.

Table 1. Comparison of chunk tag features and head noun features

Model	Recall	Precision	F-Score
baseline + chunk tag	0.6474	0.6375	0.6424
baseline + head noun	0.6696 (+0.0222)	0.6494 (+0.0119)	0.6594 (+0.0170)

Table 1 shows the effects of head noun features in GENIA corpus. In-domain tagger and chunker are used for this experiment. But, instead of chunk tag features, we append head noun features. Both recall and precision rates are slightly increased compared to the results of the system using chunk tag features.

In this abstract, we used head noun features with positional information instead of using chunk tag features. They were ascertained as good features to recognize biomedical named entities in statistical approaches.

[1] Kim, J.-D. et al. Introduction to the Bio-Entity Recognition Task at JNLPBA. In: Proceedings of JNLPBA-04, (2004) 70-75.
[2] Collins, M. Head-Driven Statistical Models for Natural Language Parsing. Ph.D. Dissertation. (1999) University of Pennsylvania.