Automatic Recognition of Topic-Classified Relations between Prostate Cancer and Genes from Medline Abstracts

Hong-Woo Chun[1] **Yoshimasa Tsuruoka**[2] **Jin-Dong Kim**[1]
1. Department of Computer Science, Graduate School of
Information Science and Technology, The University of Tokyo
2. School of Informatics, University of Manchester
{chun,tsuruoka,jdkim}@is.s.u-tokyo.ac.jp

Rie Shiba[3] **Naoki Nagata**[4] **Teruyoshi Hishiki**[4]
3. Japan Biological Information Research Center,
Japan Biological Informatics Consortium
4. Biological Information Research Center,
National Institute of Advanced Industrial Science and Technology
{rshiba,nnagata,t-hishiki}@jbirc.aist.go.jp

Jun'ichi Tsujii[1,2,5]
5. SORST, Japan Science and Technology Corporation
tsujii@is.s.u-tokyo.ac.jp

Abstract

To recognize instances of medical information concerning prostate cancer and its relevant genes, we developed a machine learning-based relation recognizer using rich contextual features. We collected prostate cancer-related abstracts from Medline. We then constructed an annotated corpus of prostate cancer and gene relations, which consisted of six $topic-classified$ categories, with more detailed information describing the type of prostate cancer and gene relation. The corpus was made with the help of biologists and a disease and gene dictionary-based name recognition technique. The process of dictionary-based name recognition generates disease-gene pairs that become $candidates$ for biomedically related pairs. Since dictionary matching tends to over-generate candidates, we used a machine learning-based named entity recognition method (1) to provide a feature for each candidate, and (2) to filter out over-generated candidates.

Experimental results showed that using a maximum entropy-based relation recognition method and a maximum entropy-based named entity recognition method together greatly improves precision at the cost of a small reduction in recall for the topic-classified relations.