Boundary Correction of Protein Names Adapting Heuristic Rules

Tomohiro Mitsumori[1], Sevrani Fation[1], Masaki Murata[2]
Kouichi Doi[1] and Hirohumi Doi[1]

[1] Graduate School of Information Science,
Nara Institute Science and Technology,
8916-5 Takayama-cho Ikoma 630-0101, Japan
{mitsumor, fation, doy}@is.aist-nara.ac.jp
doi@cl-sciences.co.jp
[2] Keihanna Human Info-Communication Research Center,
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
murata@crl.go.jp

Abstract. In this study, we made some heuristic rules related to the boundary of protein names for automated extraction of protein names from biomedical literatures. The automated extraction of protein names was carried out based on Support Vector Machine (SVM). ¿From the analysis of the results, we found whether some words of modifier words set were included or not as part of protein names. It is critical whether the modifier words set is or not included in a protein name. Adapting some heuristic rules to the corpus, the F-score was improved about 1.3% (from 76.10% to 77.41%) compared with the case without adapting proposed rules.

1 Introduction

The goal of our study is the automated information extraction related to the protein-protein interactions from biomedical literature. In this study, we carry out the extraction of protein names using SVM.

Recently, some studies have been reported regarding this issue. One difficulty is to exactly recognize the boundary of protein names because some of protein names are represented as compound words. Yamamoto et al. [1] reported a 74.9% F-score in case of exact boundary matching, and 85.0% in case of partial matching. The partial matching means that a word obtained after learning matches some of the words of the real word. Based on hand written rules, Franzén et al. [2] reported a F-score of 67.1% in case of exact boundary matching, and 82.9% in partial matching. In this paper, we make use of some heuristic rules concerning the boundary of protein names to improve precision and/or recall.

2 Experimental conditions

Some studies reported about the protein name recognition. Some are based on the machine learning[1][3] and the other are based on the hand written