Artificial Intelligence in Medicine (2007) **39**, 127—136

ELSEVIER

ARTIFICIAL
INTELLIGENCE
IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach[☆,☆☆]

Fabio Rinaldi[a,*], Gerold Schneider[a], Kaarel Kaljurand[a], Michael Hess[a], Christos Andronis[b], Ourania Konstandi[b], Andreas Persidis[b]

[a] *Institute of Computational Linguistics, University of Zurich, Binzmühlestrasse 14, CH-8050 Zürich, Switzerland*
[b] *Biovista, 34 Rodopoleos Str., Ellinikon, GR-16777 Athens, Greece*

Received 16 January 2006; received in revised form 27 August 2006; accepted 28 August 2006

KEYWORDS
Information extraction;
Text mining;
Dependency parsing;
Biomedical literature;
Protein interactions

Summary

Objective: The amount of new discoveries (as published in the scientific literature) in the biomedical area is growing at an exponential rate. This growth makes it very difficult to filter the most relevant results, and thus the extraction of the core information becomes very expensive. Therefore, there is a growing interest in text processing approaches that can deliver selected information from scientific publications, which can limit the amount of human intervention normally needed to gather those results.
Materials and methods: This paper presents and evaluates an approach aimed at automating the process of extracting functional relations (e.g. interactions between genes and proteins) from scientific literature in the biomedical domain. The approach, using a novel dependency-based parser, is based on a complete syntactic analysis of the corpus.
Results: We have implemented a state-of-the-art text mining system for biomedical literature, based on a deep-linguistic, full-parsing approach. The results are validated on two different corpora: the manually annotated genomics information access

[☆] A preliminary version of the system described in this paper has been presented in [1]. Recent results obtained after the submission of this paper are described in [2]. All URLs mentioned in this paper have been accessed and verified on 10 May 2006.

[☆☆] The tools and resources used for the work described in this paper are freely available for research purposes. The DepGENIA corpus can be downloaded from the OntoGene web site (http://www.ontogene.org/). The Pro3Gres parser and the OntoGene text mining system can be obtained by contacting the authors of this paper.

[*] Corresponding author. Tel.: +41 44 635 6724; Fax: +41 44 635 6809.
E-mail addresses: rinaldi@ifi.unizh.ch (F. Rinaldi), gschneid@ifi.unizh.ch (G. Schneider), kalju@ifi.unizh.ch (K. Kaljurand), hess@ifi.unizh.ch (M. Hess), candronis@biovista.com (C. Andronis), okonst@biovista.com (O. Konstandi), andreasp@biovista.com (A. Persidis).

doi:10.1016/j.artmed.2006.08.005

(GENIA) corpus and the automatically annotated arabidopsis thaliana circadian rhythms (ATCR) corpus.

Conclusion: We show how a deep-linguistic approach (contrary to common belief) can be used in a real world text mining application, offering high-precision relation extraction, while at the same time retaining a sufficient recall.

1. Introduction

The amount of research results in the area of molecular biology is growing at such a pace that it is extremely difficult for individual researchers to keep track of them. As such results appear mainly in the form of scientific articles, it is necessary to process them in an efficient manner in order to be able to extract the relevant results.

In the context of the OntoGene project[1] we aim at developing and refining methods for discovery of interactions between biological entities (genes, proteins, pathways, etc.) from the scientific literature, based on a complete syntactic analysis of the articles, using a novel high-precision parsing approach. We consider that advanced parsing techniques combining statistics and human knowledge of linguistics have matured enough to be successfully applied in real settings.

In Section 2, we present the datasets upon which we have based our experiments. Section 3 describes the approach taken to analyse the data. Section 4 describes how the intermediate results of data analysis are used in the relation mining task. Section 5 describes the evaluation of our results. We conclude with a survey of related work in Section 6.

2. The datasets

The tools described in this paper have been applied to extract semantic relations from two distinct corpora, which are briefly described in this section.

The ATCR corpus (arabidopsis thaliana circadian rhythms) is a set of 147 MEDLINE abstracts (up to year 2004), extracted using the keywords: *arabidopsis thaliana* and *circadian rhythms*. It has been *automatically* annotated using the "Biolab Experiment Assistant (BEA)" ™. Circadian rhythms are near-24-h rhythms of biological processes that persist in the absence of environmental cues such as light and temperature. The existence of these endogenous rhythms is believed to confer an advantage to the organism by temporally orchestrating physiological and behavioral processes to better adapt to the predictable daily changes in the environment [4]. Circadian rhythms are widespread in nature and they have been the subject of intense studies in mammalian organisms (human, mouse, rat), drosophila, fungi, higher plants and photosynthetic cyanobacteria. The rhythms in all these systems are generated by rhythmically expressed genes that form highly interconnected positive and negative feedback loops. Among these systems, the study of the circadian system of the higher plant arabidopsis thaliana, has provided many examples of rhythmic outputs and photoreceptors, advancing our understanding of the molecular basis of clock function [5]. Given the significance of arabidopsis as a model system for the study of circadian biology, we chose a relevant corpus of recent papers as one of our datasets.

GENIA [6][2] is a corpus of 2000 MEDLINE abstracts which have been *manually* annotated (by domain experts) for various biological entities according to the GENIA ontology.[3] We use version G3.02 of the GENIA corpus, which includes 18546 sentences (average length 9.27 sentences per article) and 490 941 words (average of 26.47 words per sentence).

3. Corpus analysis

This section describes the approach taken in analyzing the two corpora described in the previous section. Both of them have been processed using a natural language processing pipeline (NLPPL), which consist of a number of tools described in Section 3.1. The core component of the pipeline is Pro3Gres: a fast, deep-linguistic statistical dependency parser, which is described in detail in Section 3.2.

Pro3Gres assumes that its input has already undergone sentence splitting, tokenization, lemmatization, and noun and verb group chunking, i.e. Pro3Gres focuses on finding the dependencies between the heads of the chunks. This leaves a lot of room for experimentation with different off-the-self part-of-speech taggers, chunkers, terminology extractors, and allows us to choose the preprocessing tools which are optimized for a certain domain (e.g. news articles or biomedical literature).

[1] http://www.ontogene.org/.

[2] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/.

[3] http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/genia-ontology.html.