Learning Relations from Biomedical Corpora Using Dependency Trees

Sophia Katrenko and Pieter Adriaans

Human-Computer Studies Laboratory,
University of Amsterdam,
Kruislaan 419, 1098VA, Amsterdam, The Netherlands
katrenko@science.uva.nl, pietera@science.uva.nl

Abstract. In this paper we address the relation learning problem in the biomedical domain. We propose a representation which takes into account the syntactic information and allows for using different machine learning methods. To carry out the syntactic analysis, three parsers, LinkParser, Minipar and Charniak parser were used. The results we have obtained are comparable to the performance of relation learning systems in the biomedical domain and in some cases out-perform them. In addition, we have studied the impact of ensemble methods on learning relations using the representation we proposed. Given that recall is very important for the relation learning, we explored the ways of improving it. It has been shown that ensemble methods provide higher recall and precision than individual classifiers alone.

1 Introduction

Not only the number of publications in the biomedical domain grows rapidly every year, there are also many approaches proposed to how to handle such amount of data.

These approaches primarily consider such tasks as text mining, information extraction and information retrieval. Information retrieval focuses on the retrieval of the full documents, while the goal of information extraction is to find text fragments relevant to the user need. However, it is often useful to get more fine-grained information, for instance, the list of biomedical instances or relations. Such information might be especially important for the curation of existing resources, such as databases of interactions (Albert et al., 2003).

The paper is organized as follows. We start with the discussion of the related work and problem statement. In Section 3, we present our approach and provide motivation for it. Further, we test our approach on two data sets for the interaction extraction. We report on our results and conclude with the discussion and outlook for the future work.

2 Problem Statement and Related Work

The biggest collection of medical documents is Medline, with 2,000 citations added every week. The large size of this collection makes it impossible to annotate it all by

K. Tuyls et al. (Eds.): KDECB 2006, LNBI 4366, pp. 61–80, 2007.

humans. Consequently, there have been several attempts to create smaller annotated corpora based on Medline, such as Genetag used for the gene/protein named entity recognition (NER) (Tanabe et al., 2005), or MedTag, the corpus comprising Genetag, MedPost and ABGene (Smith et al., 2005). There have also been corpora created with a special purpose to be used by the various challenges, e.g. corpus of the annotated gene-protein relations for the "Genic Interaction Extraction Challenge" (Nédellec, 2005).

In general, the relation learning problem can be seen as a two-step process. First, the relation arguments have to be identified. Further, it is necessary to check whether the relation holds. This setting has also been used for the relation discovery in other domains (Zelenko et al., 2003), moreover, it is often assumed that the arguments have already been found. In this case, the relation learning is reduced to the second step which involves procedures enabling such verification. It has been shown by Bunescu et al. (2005) that provided the correct names of proteins are given, the accuracy of relation discovery is much higher.

The relation learning task can be formulated in the following way:

Definition 1 (Relation learning). *Given a data set* D [1] *and an n-ary relation Rel with the arguments* $X, Y \ldots Z$ *find all instances* $x \in X, y \in Y, \ldots, z \in Z$ $(x, y, z \in D)$*, such that* $Rel(x, y, \ldots, z)$ *holds.*

An example of the relation learning task is given below. In the typical scenario, one starts with the preprocessing (which includes such steps as tokenization and might require some additional analysis depending on the method used). The first step consists of named entity recognition, where all proteins occurring in the sentence are identified. There are three of them, *retinoblastoma*, *RIZ*, and *E1A*. The next step is to detect if there are any relations among them. The correct answer is an interaction between *retinoblastoma* and *RIZ*, while *E1A* does not participate in any interaction.

Input: The retinoblastoma protein binds to RIZ, a zing-finger protein that shares an epitope with the adenovirus E1A protein.
Preprocessing: The| *retinoblastoma* | *protein* | *binds* | *to* | *RIZ* | , | *a* | *zing* − *finger* | *protein* | *that* | *shares* | *an* | *epitope* | *with* | *the* | *adenovirus* | *E1A* | *protein* | . |
Step1:*The*⟨prot⟩ retinoblastoma ⟨/prot⟩ protein binds to ⟨prot⟩ RIZ ⟨/prot⟩ , a zing-finger protein that shares an epitope with the adenovirus ⟨prot⟩ E1A ⟨/prot⟩ protein .
Step2: The ⟨p1 pair="1"⟩⟨prot⟩retinoblastoma⟨/prot⟩⟨/p1⟩ protein binds to ⟨p1 pair="1"⟩⟨prot⟩RIZ⟨/prot⟩⟨/p1⟩, a zing-finger protein that shares an epitope with the adenovirus ⟨prot⟩E1A⟨/prot⟩ protein.
Output: interaction(retinoblasma, RIZ)

[1] Where a data set D can be text, semi-structured data, etc.