Extracting Higher Order Relations From Biomedical Text

Syeed Ibn Faiz
Department of Computer Science
The University of Western Ontario
syeedibnfaiz@gmail.com

Robert E. Mercer
Department of Computer Science
The University of Western Ontario
mercer@csd.uwo.ca

Abstract

Argumentation in a scientific article is composed of unexpressed and explicit statements of old and new knowledge combined into a logically coherent textual argument. Discourse relations, linguistic coherence relations that connect discourse segments, help to communicate an argument's logical steps. A biomedical relation exhibits a relationship between biomedical entities. In this paper, we are primarily concerned with the extraction of connections between biomedical relations, a connection that we call a higher order relation. We combine two methods, namely biomedical relation extraction and discourse relation parsing, to extract such higher order relations from biomedical research articles. Finding and extracting these relations can assist in automatically understanding the scientific arguments expressed by the author in the text.

1 Introduction

We use the term ***higher order relation*** to denote a relation that relates two biomedical relations. Consider, for example, the following sentence:

(1) Aspirin appeared to prevent VCAM-1 transcription, since it dose-dependently inhibited induction of VCAM-1 mRNA by TNF.

We can find two biomedical relations involving Aspirin: Aspirin–*prevents*–VCAM-1 transcription and Aspirin–*inhibits*–induction of VCAM-1 mRNA. These two relations are connected by the word *since*. The higher order relation conveys a causal sense, which indicates that the latter relation causes the earlier one. In genetic transcription mRNA is generated (a process known by the reader, so not expressed in the argument). This piece of the author's argument is that by observing inhibition of mRNA induction (the genetic process that activates transcription) by different doses of aspirin, the inference that aspirin prevents the transcription can be made. This inference is textually signalled by the discourse connective *since*.

Formally, we define a higher order relation as a binary relation that relates one biomedical relation with another biomedical relation. In this paper we propose a method for these extracting higher order relations using discourse relation parsing and biomedical relation extraction.

2 Extracting Higher Order Relations

There are two stages in our method for extracting higher order relations from text. In the first stage we use a discourse relation parser to extract the explicit discourse relations from text. In the second stage we analyze each extracted explicit discourse relation to determine whether it can produce a higher order relation. We use a biomedical relation extraction system in this process. For each argument of an explicit discourse relation we find all occurrences of biomedical relations in it. Higher order relations are then constructed by pairing the biomedical relations or observations found in the discourse arguments. The sense of the explicit discourse relation is used to interpret all the higher order relations derived from it.

Parsing an explicit discourse relation involves three steps: identifying the explicit discourse connective, the arguments and the sense. In (Faiz and Mercer, 2013) we showed how to use syntactic and surface level context to achieve a state-of-the-art result for identifying discourse connectives from text. Our work on a complete explicit discourse relation parser is presented in (Faiz, 2012). For identifying the arguments of discourse connectives we use the head-based representation proposed by Wellner and Pustejovsky (Wellner and Pustejovsky, 2007). We found that this head-based

Proceedings of the First Workshop on Argumentation Mining, pages 100–101,
Baltimore, Maryland USA, June 26, 2014.

representation is very suitable for the task of extracting higher order relations. The head of an argument plays an important role in selecting a biomedical relation as an argument to a higher order relation.

This observation regarding the heads of the discourse arguments has another useful implication. Since the biomedical relations that we have to consider need to involve the argument head, we only have to extract the portion of the argument that is influenced or dominated by the head. One simple way to do this is to consider the dependents of the head in the dependency representation. Wellner (2009) reported that finding the dependents of the syntactic head of an argument often gives a good approximation of the argument extent .

3 Evaluation

Our algorithm for extracting higher order relations depends on discourse parsing and biomedical relation extraction. We have discussed our implementation of these components and evaluated their performance in previous work (Faiz, 2012; Faiz and Mercer, 2013). We have evaluated the algorithm we present in this paper in terms of how accurately it can use those components in order to find higher order relations. More specifically, we will measure how accurately it can determine the part of the full argument extent that contains the biomedical entities in it.

For this evaluation we used the AIMed corpus (Bunescu et al., 2005). This corpus contains an annotation for protein-protein interactions. From this corpus we collected 69 discourse relations.

For both ARG1 and ARG2 we performed two tests. We measured from the argument heads how many protein mentions occurring within the argument extent (the *True Positives*) are found and how many protein mentions that lie beyond the argument extent (the *False Positives*) are found. For ARG1, we found that our algorithm missed only one protein mention and incorrectly found three proteins from outside the argument extent, a precision of 98% and a recall of 99.32%. For ARG2, we obtained a 100% precision and a 99% recall.

We conducted another experiment, which is similar to the previous one except that now instead of counting only the protein mentions, we counted all the words that can be reached from an argument head. In other words, this experiment evaluates our algorithm in terms of how accurately it can identify the full argument extent (i.e., the words in it). For ARG1 and ARG2 we got an F-score of 91.98% and 92.98% respectively.

4 Discussion

Extraction of many higher order relations is dependent on coreference resolution. For example, in (1), Aspirin is anaphorically referred to in ARG2. In our current implementation we lack coreference resolution. Therefore, augmenting a coreference resolution module in our pipeline would be an immediate improvement.

In our implementation, we used a simple but imperfect method to determine whether a biomedical relation involves the head of a discourse argument. We checked whether the head appears between the biomedical entities or within a short distance from either one in the sentence. However, this simple rule may produce spurious higher order relations. One way to improve this method would be to consider the rules we presented for rule-based biomedical relation extraction. Most of the rules give a dependency path corresponding to the relation they can extract. That path can then be analyzed to determine whether the relation depends on the head.

Acknowledgments

This work was partially funded by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to R. Mercer.

References

Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk W. Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155, February.

Syeed Ibn Faiz and Robert E. Mercer. 2013. Identifying explicit discourse connectives in text. In *Canadian Conference on AI*, pages 64–76.

Syeed Ibn Faiz. 2012. Discovering higher order relations from biomedical text. Master's thesis, University of Western Ontario, London, ON, Canada.

Ben Wellner and James Pustejovsky. 2007. Automatically identifying the arguments of discourse connectives. In *EMNLP-CoNLL*, pages 92–101. ACL.

Ben Wellner. 2009. *Sequence models and ranking methods for discourse parsing*. Ph.D. thesis, Brandeis University, Waltham, MA, USA. AAI3339383.