Identifying the Epistemic Value of Discourse Segments in Biology Texts

Anita de Waard[(1)] Paul Buitelaar[(2)] Thomas Eigner[(3)]
(1) Elsevier & Universiteit Utrecht, the Netherlands (anita@cs.uu.nl)
(2) DERI - NLP Unit, Galway, Ireland (paulb@deri.org)
(3) DFKI, Saarbrcken, Germany (teigner@dfki.de)

1 Introduction

To manage the flood of information that threatens to engulf (life-)scientists, an abundance of computer-aided tools are being developed. These tools aim to provide access to the knowledge conveyed within a collection of research papers, without actually having to read the papers. Many of these tools focus on text mining, by looking for specific named-entities that have scientific meaning, and relationships between these. An overview of the current state of the art is given in Rebholz-Schuhmann et al. (2005) and Couto et al. (2003). Typically, these tools identify a list of sentences containing relationships between two specific named-entities that can be found using rules or a thesaurus of synonyms. These sentences represent an overview of the interactions that are known with a specific entity, thus precluding the need for an exhaustive literature study. For example, the following are a few sentences that have been found using a typical text mining tool for the relationship 'p53 activates *':

1. *The p53 tumor suppressor protein exerts most of its anti-tumorigenic activity by transcriptionally activating several pro-apoptotic genes.*

2. *We found that p53 ... activates[,] the promoter of the myosin VI gene.*

However, in order to be able to use these statements and to draw conclusions about its subject ("Which entities does p53 activate?") we still need to read the article that they appeared in, identify the experimental context and the epistemic ('truth') value of each statement. For instance, 1. does not seem to represent an experimental finding that is arrived at in the paper

Proceedings of the 8th International Conference on Computational Semantics, pages 351–354,
Tilburg, January 2009.

that the sentence is taken from; instead, it seems to be a citation. So, to be able to evaluate its epistemic value (”How true is this?”) we need to read the paper that contained the sentence and paper(s) where the statement was first experimentally motivated. In the case of 2., a clear statement is given on what the authors of the paper have found. But “How did they find it?”, “What experimental setup and control experiments were used?”, “What were their assumptions?” Biologists will need to check these and other issues before accepting 2. as a fact. Our research therefore concerns the classification of sentences in biology texts by ’epistemic segment type’, with the purpose of enabling a better way to summarize, mine and compare statements within biology texts. The current paper describes a first venture into doing this in a computational way.

2 Epistemic Segment Types for Biology Texts

As motivated elsewhere we have identified seven epistemic segment types (De Waard, 2007):

- **Fact**: statement presumed to be accepted as true by the community, e.g. *Cellular transformation requires the expression of oncogenic RASV12.*
- **Hypothesis**: possible explanation for a set of phenomena, e.g. *This suggests possible roles for APC in G1 and G0 phases of the cell cycle.*
- **Implication**: interpretation of results, e.g. *These results indicate that our procedure is sensitive enough to detect mild growth differences.*
- **Method**: ways in which the experiment was performed, e.g. *We inserted 500 bp fragments ... in a modified pMSCV-Blasticidin vector.*
- **Problem**: discrepancies or unknown aspects of the known fact corpus, e.g. *The small number of miRNAs with a known function stresses the need for a systematic screening approach to identify more miRNA functions.*
- **Goal**: implicit hypothesis and problem, e.g. *identify miRNAs that can interfere with this process and ... contribute to the development of tumor cells*
- **Result**: a summary of the results of measurements, e.g. *we observed an approximately 4-fold increase in miR-311 signal*

For example, **Fact** segments are taken from another source of knowledge (explicitly referred to or presumed to be known) and therefore not experimentally ascertained in the article, whereas **Result** segments are obtained by measurements discussed in the paper itself. For the sentences in the previous section, we therefore see that 1. is a **Fact** and 2. is a typical **Result** segment. To classify the segments (manually first), we have used several linguistic clues, as well as an understanding of the context of a segment. Important linguistic clues are the verb tense of the segment and specific markers used to identify a segment transition, e.g. the transition between a Result and an Implication segment is usually indicated by a phrase such as 'These results suggest that'. The segment types and selected specific markers that we used in our research here are as follows (using regular expressions to shorten notation):

- **Hypothesis**: results indicate, suggest, suggesting that
- **Implication**: data||results demonstrate||suggest||indicate, data||results show
- **Method**: by cloning||using, using additionally, we activated||constructed
- **Goal**: to examine||identify||investigate||mimic||shed light||start to
- **Result**: as expected||predicted, resulting in, shows that, this confirms

3 Automatic Identification of Epistemic Segment Types

To investigate if we could use this set of markers for the automatic identification of segment type, we applied them to an independently developed data set of 1721 biomedical abstracts on 'mantle cell lymphoma' that we downloaded from PubMed. We randomly selected 100 sentences, in which a marker was identified, and to which one out of five segment types (**Hypothesis, Implication, Method, Goal, Result**) was assigned by a simple automatic procedure, i.e. we matched the markers to a part-of-speech enriched version of the PubMed corpus. One or more segment types were assigned in case of a match. The resulting assignments were then evaluated by the first author of this paper. Results were encouraging as only 30 out of 100 assignments were incorrect. Most of these (12) were between **Hypothesis, Implication**, which is not surprising as their markers are overlapping and therefore ambiguous. Others that were somewhat frequent were: **Hypothesis** instead of **Fact** (3), **Result** instead of **Fact** (3), **Result** instead of **Method** (2), **Goal** instead of **Method** (2), **Goal** instead of **Problem** (2). Of these however, **Fact** and **Problem** were not covered by our set of segment specific markers and could therefore not be recognized.

4 Conclusions and Future Work

As a first conclusion, results are encouraging enough to merit further research. We have identified several follow-up steps that can help improve our results. First, we plan to segment the sentences into smaller discourse units. For instance, sentences such as the following are quite clearly divided into two parts: a **Goal** and a **Method**:

- **Goal**: *To examine miRNA expression from the miR-Vec system,*
- **Method**: *a miR-24 minigene-containing virus was transduced into human cells.*

Such sentences are quite common, as are sentences containing **Method, Result** and **Result, Implication** segments; this clearly indicates that this move order is logical and occurs often. Secondly, there is a clear correlation between segment type and verb tense. **Method, Result** are overwhelmingly stated in the past tense, whereas **Fact, Implication** are given in the present tense. Using verb tense as a marker could further improve classification scores. Lastly, we are interested in applying our epistemic values to augment and improve bioinformatics tools, and investigating the value of these categories with users. We are actively pursuing collaborations in this area.

References

Rebholz-Schuhmann, D., H.Kirsch & F. Couto (2005) Facts from text - is text mining ready to deliver? *PLoS Biology* 3(2).

Couto, F., M.J. Silva & P. Coutinho (2003) Improving information extraction through biological correlation. In *Proc. European Workshop on Data Mining and Text Mining*, Dubrovnik.

Waard, A. de (2007) The pragmatic research article. In *Proc. 2nd International Conference on the Pragmatic Web*, Tilburg.