Zones of conceptualisation in scientific papers: a window to negative and speculative statements

Maria Liakata
Department of Computing Science, Aberystwyth University
European Bioinformatics Institute, Cambridge
liakata@ebi.ac.uk

Abstract

In view of the increasing need to facilitate processing the content of scientific papers, we present an annotation scheme for annotating full papers with zones of conceptualisation, reflecting the information structure and knowledge types which constitute a scientific investigation. The latter are the Core Scientific Concepts (CoreSCs) and include Hypothesis, Motivation, Goal, Object, Background, Method, Experiment, Model, Observation, Result and Conclusion. The CoreSC scheme has been used to annotate a corpus of 265 full papers in physical chemistry and biochemistry and we are currently automating the recognition of CoreSCs in papers. We discuss how the CoreSC scheme relates to other views of scientific papers and indeed how the former could be used to help identify negation and speculation in scientific texts.

1 Introduction

The recent surge in the numbers of papers produced, especially in the biosciences, has highlighted the need for automatic processing methods. Work by [Lin (2009)] has shown that methods such as information retrieval are more effective if zones of interest are specified within the papers. Various corpora and annotation schemes have been proposed for designating a variety of linguistic phenomena permeating scientific papers, including negation, hedges, dependencies and semantic relations [Vincze et al. (2008); Pyysalo et al. (2007); Medlock and Briscoe (2007); McIntosh and Curran (2009)]. Other schemes follow the argumentation and citation flow within papers [Teufel et al. (2009); Teufel and Siddharthan (2007)] or indeed a combination of some of the above along multiple dimensions [Shatkay et al. (2008)].

In the following we present the CoreSC annotation scheme and a corpus with CoreSC annotations. The CoreSC scheme is used at the sentence level to identify the core components that constitute a scientific investigation. We discuss how the CoreSC scheme relates to other annotation schemes representing alternate views of scientific papers and how CoreSCs could be used to guide the identification of negation and speculation.

2 The CoreSC scheme

The CoreSC annotation scheme adopts the view that a scientific paper is the human-readable representation of a scientific investigation and therefore seeks to mark the components of a scientific investigation as expressed in the text. CoreSC is ontology-motivated and originates from the CISP meta-data [Soldatova and Liakata (2007)], a subset of classes from EXPO [Soldatova and King (2006)], an ontology for the description of scientific investigations. CISP consists of the concepts: `Motivation`, `Goal`, `Object`, `Method`, `Experiment`, `Observation`, `Result` and `Conclusion`, which were validated using an on-line survey as constituting the indispensable set of concepts necessary for the description of a scientific investigation. CoreSC implements these as well as `Hypothesis`, `Model` and `Background`, as a sentence-based annotation scheme for 3-layered annotation. The first layer pertains to the previously mentioned 11 categories, the second layer is for the annotation of properties of the concepts (e.g. "New", "Old") and the third layer caters for identifiers (conceptID), which link together instances of the same concept, e.g. all the sentences pertaining to the same method will be linked together with the same conceptID (e.g. "Met1").

If we combine the layers of annotation so as to

Table 1: The CoreSC Annotation scheme

Category	Description
Hypothesis	A statement not yet confirmed rather than a factual statement
Motivation	The reasons behind an investigation
Background	Generally accepted background knowledge and previous work
Goal	A target state of the investigation where intended discoveries are made
Object-New	An entity which is a product or main theme of the investigation
Object-New-Advantage	Advantage of an object
Object-New-Disadvantage	Disadvantage of an object
Method-New	Means by which authors seek to achieve a goal of the investigation
Method-New-Advantage	Advantage of a Method
Method-New-Disadvantage	Disadvantage of a Method
Method-Old	A method mentioned pertaining to previous work
Method-Old-Advantage	Advantage of a Method
Method-Old-Disadvantage	Disadvantage of a Method
Experiment	An experimental method
Model	A statement about a theoretical model or framework
Observation	the data/phenomena recorded in an investigation
Result	factual statements about the outputs of an investigation
Conclusion	statements inferred from observations & results relating to research hypothesis

give flat labels, we cater for the categories in table 1.

The CoreSC scheme was accompanied by a set of 45 page guidelines which contain a decision tree, detailed description of the semantics of the categories, 6 rules for pairwise distinction and examples from chemistry papers. These guidelines are available from http://ie-repository.jisc.ac.uk/88/.

3 The CoreSC corpus

We used the CoreSC annotation scheme and the semantic annotation tool SAPIENT [Liakata et al. (2009)] to construct a corpus of 265 annotated papers [Liakata and Soldatova (2009)] from physical chemistry and biochemistry. The CoreSC corpus was developed in two different phases. During phase I, fifteen Chemistry experts were split into five groups of three, each of which annotated eight different papers; A 16th expert annotated across groups as a consistency check. This resulted in a total of 41 papers being annotated, all of which received multiple annotations. We ranked annotators according to median success in terms of inter-annotator agreement (as measured by Cohen's kappa) both within their groups and for a paper common across groups. In phase II, the 9 best annotators of phase I each annotated 25 papers, amounting to a total of 225 papers.

The CoreSC corpus is now being used to train a classifier for the automation of Core Scientific concepts in papers.

4 Correlating CoreSCs to other zones of interest

Given the plethora of annotation schemes, it is interesting to investigate the correlation between different views of scientific papers and how different schemes map to each other. We recently looked at the correlation between the CoreSC scheme, which views papers as the humanly readable representation of scientific investigations and seeks to recover the investigation components within the paper, and AZ-II [Teufel et al. (2009)], which assumes a paper is the attempt of claiming ownership for a new piece of knowledge and aims to recover the rhetorical structure and the relevant stages in the argumentation.

By definition, the two schemes focus on different aspects of the papers, with CoreSCs providing more detail with respect to different types of methods and results and AZ-II looking mostly at the appropriation of knowledge claims. Based on a set of 36 papers annotated with both schemes, we were able to confirm that the two schemes are indeed complementary [Liakata et al. (2010)]. CoreSC categories provide a greater level of granularity when it comes to the content-related categories whereas AZ-II categories cover aspects of the knowledge claims that permeate across different CoreSC concepts.

In [Guo et al. (2010)] we followed a similar methodology for annotating abstracts with CoreSCs and an independently produced annotation scheme for abstract sections [Hirohata et al. (2008)]. We found a subsumption relation between the schemes, with CoreSCs providing the

finer granularity.

To obtain the mapping between annotation schemes, which allows annotation schemes to be defined in a wider context, we ideally require annotations from different schemes to be made available for the same set of papers. However, a first interpretation of the relation between schemes can be made by mapping between annotation guidelines.

5 Thoughts on using CoreSCs for Negation and Speculation

Current work of ours involves automating the recognition of CoreSCs and we plan to use them to produce extractive summaries for papers. We are also in the process of evaluating the usefulness of CoreSCs for Cancer Risk Assessment (CRA). An important aspect of the latter is being able to distinguish between positive and negative results and assess the confidence in any conclusions drawn. This naturally leads us to the need for exploring negation and speculation, both of which are prominent in scientific papers, as well as how these two phenomena correlate to CoreSCs.

While it seems that negation can be identified by means of certain linguistic patterns [Morante (2010)], different types of negation can appear throughout the paper, some pertaining to background work, problems serving as the motivation of the paper, others referring to intermediate results or conclusions. It is interesting to look at these different types of negation in the context of each of the different CoreSCs, the type of linguistic patterns used to express it and their distribution across CoreSCs. This can provide a more targetted approach to negation, while at the same time it can be used in combination with a CoreSC to infer the type of knowledge obtained (e.g. a positive or negative result). We plan to use automatic methods for recognising negation patterns in CoreSCs and relate them to specific CoreSC categories.

There is a consensus that identifying speculation is a harder task than identifying negation. Part of the problem is that "speculative assertions are to be identified on the basis of the judgements about the author's intended meaning, rather than on the presence of certain designated hedge terms" [Medlock and Briscoe (2007); Light et al. (2004)]. When annotating papers with CoreSCs, annotators are required to understand the paper content rather than base category assignments entirely on linguistic patterns. This is why we have chosen experts as annotators for the creation of the CoreSC corpus. So both speculation and CoreSC annotation appear to be higher level annotation tasks requiring comprehension of the intended meaning. Looking at the annotation guidelines for hedges [Medlock and Briscoe (2007)], it would seem that cases of hedge type 1 correspond to to CoreSC `Conclusion`, hedge type 2 pertains to `Background`, hedge type 3 would mainly be cases of `Motivation`, hedge type 4 maps to `Motivation` or `Hypothesis`, hedge type 5 maps to `Goal` and hedge type 6 maps to `Conclusion`. One can look at speculation in the zones/windows identified by the previously mentioned CoreSCs. Indeed, two of the categories, `Hypothesis` and `Motivation` are speculative by definition. We intend to port the issue of identifying speculation in our papers to that of identifying the corresponding CoreSCs. We also plan to annotate the hedge classification data of [Medlock and Briscoe (2007)] with CoreSCs to confirm the mapping between the two schemes.

References

Y. Guo, A. Korhonen, M. Liakata, I Silins, L. LiSun, and U. Stenius. Identifying the information structure of scientific abstracts: An investigation of three different schemes. In *Proceedings of BioNLP 2010. To appear.*, Uppsala, Sweden, 2010.

K. Hirohata, N. Okazaki, S. Ananiadou, and M. Ishizuka. Identifying sections in scientific abstracts using conditional random fields. In *Proc. of the IJCNLP 2008*, 2008.

M. Liakata and L.N. Soldatova. The art corpus. Technical report, Aberystwyth University, 2009. URL `http://www.aber.ac.uk/en/cs/research/cb/projects/art/art-corpus/`.

M. Liakata, Claire Q, and S. Soldatova. Semantic annotation of papers: Interface & enrichment tool (sapient). In *Proceedings of BioNLP-09*, pages 193–200, Boulder, Colorado, 2009.

M. Liakata, S. Teufel, A. Siddharthan, and C. Batchelor. Corpora for the conceptualisation and zoning of scientific papers. 2010.

M. Light, X.Y. Qiu, and P. Srinivasan. The language of bioscience: Facts, speculations and statements in between. In *Proceedings of BioLink 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users*, Boston, 2004.

J. Lin. Is searching full text more effective than searching abstracts? *BMC Bioinformatics*, 10:46, 2009.

T. McIntosh and J.R. Curran. Challenges for automatically extracting molecular interactions from full-text articles. *BMC Bioinformatics*, 10(311), 2009.

B. Medlock and T. Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *45th Annual Meeting of the ACL*, pages 23–30, Prague, Czech Republic, 2007.

R. Morante. Descriptive analysis of negation cues in biomedical texts. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*, pages 1429–1436, Valletta, Malta, 2010.

S. Pyysalo, F. Ginter, J. Heimonen, J. Bjorne, J. Boberg, J. Jarvinen, and T. Salakoski. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1), 2007.

H. Shatkay, F. Pan, A. Rzhetsky, and W.J. Wilbur. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Journal of Bioinformatics*, 24:18:2086–2093, 2008.

L.N. Soldatova and R.D. King. An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3:795–803, 2006.

L.N. Soldatova and M. Liakata. An ontology methodology and cisp-the proposed core information about scientific papers. Technical Report JISC Project Report, Aberystwyth University, 2007. URL `http://ie-repository.jisc.ac.uk/137/`.

S. Teufel and A. Siddharthan. Whose idea was this, and why does it matter? attributing sicentific work to citations. In *Proceedings of NAACL-HLT-07*, 2007.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP-09*, Singapore, 2009.

V. Vincze, G. Szarvas, R. Farkas, G. Mra, and J. Csirik. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9, 2008.