Hypotheses, evidence and relationships: The HypER approach for representing scientific knowledge claims

Conference or Workshop Item

Hypotheses, Evidence and Relationships: The HypER Approach for Representing Scientific Knowledge Claims

Anita de Waard[1,2], Simon Buckingham Shum[3], Annamaria Carusi[7], Jack Park[3], Matthias Samwald[4,5], Ágnes Sándor[6]
[1]Elsevier Labs, Radarweg 29, Amsterdam, The Netherlands,
[2]UiL-OTS, Utrecht University, The Netherlands,
[3]Knowledge Media Institute, The Open University, UK,
[4]Digital Enterprise Research Institute, Nat'l University of Ireland Galway, Galway, Ireland,
[5]Konrad Lorenz Institute for Evolution and Cognition Research, Altenberg, Austria,
[6]Xerox Research Centre Europe, France
[7] Oxford e-Research Centre, University of Oxford

Abstract. Biological knowledge is increasingly represented as a collection of (entity-relationship-entity) triplets. These are queried, mined, appended to papers, and published. However, this representation ignores the argumentation contained within a paper and the relationships between hypotheses, claims and evidence put forth in the article. In this paper, we propose an alternate view of the research article as a network of 'hypotheses and evidence'. Our knowledge representation focuses on scientific discourse as a *rhetorical* activity, which leads to a different direction in the development of tools and processes for modeling this discourse. We propose to extract knowledge from the article to allow the construction of a system where a specific scientific claim is connected, through trails of meaningful relationships, to experimental evidence. We discuss some current efforts and future plans in this area.

Keywords: hypothesis identification, discourse analysis, pragmatic web, science publishing, argumentation tools, author intent.

1 Introduction

To populate biological databases, computational language processing tools are increasingly being utilized; see e.g. [Jensen et al. 2006] and [Hunter and Bretonnel Cohen, 2006] for an overview of this field. One of the main goals of this field, also referred to as (biomedical) fact extraction [Rebholz-Schuhmann, et al., 2005] is to produce a collection of triplets, consisting of entities, generally connected to an ontology instance, and related, via a verb, to other entities. These entity-relationship-entity (or subject-predicate-object) triplets are expressed in RDF or similar standards, and are used in different ways: as Structured Digital Abstracts, they are appended to scientific documents [Seringhaus and Gerstein, 2007] or in query interfaces, triplets are used as interface to query the life science literature.

An example of such a system is MEDIE [MEDIE] where questions can be asked of the nature 'what relationships to which other entities does entity X possess?' For

example, the answer to the question 'p53 <activate> X?' gives the following results, where the bold text is the **subject** of the triplet, the italicized is the *verb*, and the underlined text represents the object of the triplet:

(1) SRp55 is one of the most ubiquitous splicing factors and one that can be *up-regulated* **by DNA damage in the absence of p53** , …
(2) PIG3 or TP53I3 is the only known member of the medium chain dehydrogenase/reductase superfamily *induced* **by p53** …
(3) In the liver, DMBA induced strong onco/suppressor gene expression as early as 6 hours after the treatment, but MNU *increased* **the p53 gene expression** 12 hours after the treatment.
(4) Recently, we found that **nucleophosmin (NPM), a key factor involved in p53 signaling pathway**, interacts with HEXIM1 and *activates* P-TEFb-dependent transcription

There are two main problems with using this type of tool. The first issue is, of course, that current processing tools are not yet accurate enough. If we look at (2), 'the only known member of the medium chain superfamily that p53 induces', PIG3, is not recognized; in (3), the knowledge gleaned is proposed to be <the p53 gene expression> increased <the treatment>, which makes no sense at all. However, we can imagine that with more advanced Natural Language Processing tools, these issues might be solved, and great advances are made in this field. What remains problematic is that even if we were to have a perfect representation of phrases into triplets, this collection of sentences still do not answer the question 'what does p53 activate?' An important omission of this representation is that we get no grip on the validity or the *epistemic value* of each sentence: does it contain new experimental knowledge, created by the author; is it a citation of accepted knowledge, or is the statement purely hypothetical? In other words, what is the author intent behind the statement?

If we look at the epistemic value of sentences (1) – (4), it is clear that (1) – (2) contain a reformulation of existing knowledge, supported by references or presumed to be widely known; on the other hand, (3) and (4) are results found by the author in present and past work, respectively. To be able to accept these statements and add them to a knowledgebase, a user needs to be convinced that, first of all, the author intends a statement to be a plausible claim (as opposed, for instance, to a hypothetical claim, or a disputed citation), and secondly, that there is adequate backing for this claim. So two steps are needed: first, the assignment of epistemic status to a sentence (e.g. 'known fact' or 'experimental result' or 'hypothesis'), and secondly, a link to the evidence the author has to support her claim.That means that we need to know where new knowledge is presented in the text, and how this knowledge is supported by evidence, either through experiments, or through references. What we would like to have is a list of claims or *hypotheses*, made by specific authors, some presentation of *evidence* for the hypothesis, as well as *relationships* connecting them, concerning a) the nature of the evidence and b) the relationship to other hypotheses.

The shift to author intent means shifting our conceptualization of the text towards discourse: that is, a move from viewing the text as a collection of verbs and nouns, to a view of the contextualized pragmatic language used for science. We believe that utilizing this model of knowledge conveyed by biological discourse can increase the

value of existing text mining tools, and help improve access to collections of scientific papers represented as networks of collection of claims that have a defined epistemic value, with links to experimental evidence and argumentative relationships to other statements and evidence. We call this conceptual approach *'Hypotheses, Evidence and Relationships' (HypER)*. We argue that this representation adds essential knowledge to fact extraction, by taking into account how scientific hypotheses are argued, supported by experimental findings, and how they are interconnected. We are thus arguing for the need to add the dimension of *pragmatics* [Schoop, 2006] to existing semantic representations.

The basis of the HypER approach will be discussed in the next section. In section 3, we will discuss some related existing work, which underlies the HypER concept; in section 4, we describe some preliminary conclusions.

2 The HypER approach: Taking Scientists' Discourse Goals into Account

The primary move that we propose is changing the focus of textual analysis from the phenomenon studied (the object of the study), to the author's rhetorical/pragmatic intent. This view of knowledge representation stands on the metaphorical shoulders of a great collection of work in computational linguistics, discourse comprehension, and discourse analysis involved with identifying discourse goals and speaker intent. To paraphrase [Hovy, 1993]: 'As an initial assumption, we take it that scientific discourse is goal-oriented: scientists communicate for a reason.'

There is a wealth of literature in computational linguistics and discourse analysis [Schiffrin et al., 2003] that deals with the identification (and in some cases generation) of text in terms of discourse goals, focal shift, and pragmatic intent. This research is being used to analyze all manners of text, ranging from news [Van Attenveldt et al., 2008] and public sentiment towards government policies [Kwon, 2006] to conversations [Wooffitt, 2005], patient guideline author's communications [Boivin, 2009] and 'ex-gay rhetoric' [Stewart, 2008].

However, discourse goals are rarely analyzed for biological texts, which is our topic of study. So what intent do biologists have? We argue that primary research articles should be treated, primarily, as *persuasive* texts (see also [De Waard et al., 2006; De Waard and Kircz, 2008] and classic texts in scientific discourse studies, such as [Gross, 1996; Bazerman, 1988; Latour, 1997]). The author's main goal is to persuade the reader of the validity of her claims. There are two aspects to this: the value for the author(s) and the value for the reader(s). The author puts a claim forward as having a certain value, but readers are not constrained to accept it that. The persuasiveness of the discourse lies in the authors' attempt to persuade their readers to accept the epistemic values they put on claims. The predominant goal of scientific authors is to convince their peers of their claims, and share the epistemic values they have assigned to statements. To do this, they use rhetoric, typical to the narrative form, and supported by references and (experimental) data [Latour et al., 1997]. So, to represent scientific articles, we should identify the critical rhetorical elements inside the text.

At the basic level, and as a first approach, these rhetorical elements can be represented by the main hypotheses the author posits, and supporting evidence in the form of experiments and references. For example, we would like to see a summary of the abstract referred to in sentence (1) [Yan, 2008] to look something like this:

Hypotheses: "knockdown of mutant p53 markedly inhibits cell proliferation"
"one mechanism by which mutant p53 acquires its gain-of-function is through the inhibition of Id2 expression"
Evidence (experimental):

- knockdown of mutant p53 markedly inhibits cell proliferation <link to method + figure>
- knockdown of mutant p53 sensitizes tumor cells to growth suppression by various chemotherapeutic drugs <link to method + figure>
- knockdown of Id2 can restore the proliferative potential of tumor cells inhibited by withdrawal of mutant p53 <link to method + figure>

Evidence (supported by other hypotheses):

- Overexpression of mutant p53 is a common theme in human tumors <'supports' link to claim in other text>
- knockdown of mutant p53 sensitizes tumor cells to growth suppression by various chemotherapeutic drugs <'supports' link to claim in other text>

What is critical here is the identification of *new* knowledge, claimed by the authors, vs. the elements on which this knowledge is *based*, in terms of experimental results and references to other work, and the underlying relationships. Adding this evaluation to even the sentences in the abstract (which the previous examples are based on) can lead to a much more usable representation of the scientific text. Ideally, each item in the 'evidence' list should be augmented by, a) a description of the method used to obtain the result, b) a figure representing the result, and c) links to the references that support the or detract from the main hypothesis. Such a representation would allow us to construct a semantic network of linked hypotheses and evidence.

There are many ways to identify and represent this persuasive cluster of scientific findings, both in terms of visualization and in terms of XML/RDF-based knowledge representations. It is our goal to further coordinate and stabilize such formats for modeling, exchanging, and facilitating access to knowledge expressed in this way. Some plans are described in section 4, but first, we identify some key elements needed to construct such a system, and current work on realizing these.

3. Elements of a System for Creating HypER-based Knowledge

To work with knowledge that is represented in this way, various systems need to be developed and interlinked, to allow a user to search for, view and browse the heritage of a specific claim, evaluate the evidence supporting it, link it to other claims, and follow the trail of hypotheses and evidence across the literature. We do not have either the space or the experience to describe a full-fledged system that could deliver

a HypER – driven knowledge representation, but want to start to make a list of the types of elements which such a system could include:

A. *Hypothesis Creation/Identification* tools – to manually or automatically create and/or extract hypotheses and relationships
B. *Argumentation Representation* tools – to allow user interaction with the knowledge presented, and discussions between the authors/users
C. *Discourse Representations* – for representing documents containing hypotheses and evidence
D. *Rhetorical relationships/argumentational schema's* – for relations between hypotheses, and hypotheses and evidence
E. *Peer review* tools – to validate the hypotheses, experimental descriptions and data
F. *System for Methodological modeling* tools – to model and compare experimental methods
G. *Intellectual property rights management* – for this disconnected set of content

Combined, these elements could form the building blocks of a system that allows a user to explore the provenance of a specific claim, evaluate the data supporting it, and follow the trail of claims derived from or leading to the current claim. In this paper we cannot elaborate all of the above, but will focus here on current work in two key areas: first, we will discuss (3.1) argumentation interfaces and then (3.2) hypothesis extraction methods. C and D, concerning discourse and relationship representation, are discussed in a paper also submitted to this workshop [Groza et al, 2009].

3.1 Argumentation representation

To better allow the exploration of related arguments and interaction in a community, and build hypothesis-based knowledge 'gardens' [Park, 2008] we need appropriate interactive argumentation tools. Argumentation visualization tools are created to analyze the discourse and (dis)agreement between collections of documents. Their goal is to present the user with a distillation of the key discourse moves within and between documents, without having to read each one, and see argumentation and claims and counterclaims represented at a higher level of abstraction.

A well-established body of work is concerned with argumentation schematisation in the legal and news domains. Van Den Braak et al. [2006] review various argumentation visualization tools and find some support that these tools do support improved reasoning abilities. In Bex et al [2007], an example is given of how legal 'stories' are converted into a set of statements, connected by legal (argumentational) relations, to allow an overview of an (eye-witness) account. In a different genre, that of news Van Atteveldt [Van Atteveldt, 2008] marks up a corpus of newspaper articles with the Relational Content Analysis method [Popping, 2000; Roberts, 1997], to construct a detailed picture of the relations between different 'actors' and nodes, which is then modeled in RDF and accessed with semantic technologies.

There have been several efforts to model scientific argumentation to an existing schema. The Open University developed a thoroughly founded ontology of

argumentation relations [Mancini and Buckingham Shum, 2006] to provide a network of argumentation on a specific issue. The ClaiMaker (now: Cohere) tools [Buckingham Shum, 2008] enable users to annotate significant ideas and claims on a document, linked by a user-extensible set of semantic relationships. The SALT initiative [Groza et al., 2008] provides a LaTeX-based tool to computer science that allows authors the ability to identify their main claims, and mark up relationships to supporting statements using RST relations [Mann and Thomson, 1987].

In the MachineProse proposal, which bridges the argumentation visualisation and structured abstract approaches, [Dinakarpandian et al., 2006], science is represented as a set of assertions, which can be 'represented in its simplest form as a pair of entities'. A paper can affirm, negate or be inconclusive about an assertion. Here, curators identify a set of assertions and evaluations with a paper; the paper proposes submission of structured opinions together with article submission.

Several argumentation visualization tools have been developed for the life sciences, as well. In NeuroScholar [Burns and Cheng, 2006] a model is made of the argumentation within an article; the system uses these claims and places them within in the context of related claims. SWAN [Ciccarese et al., 2008] focuses on identifying hypotheses in papers on Alzheimer's disorder, and uses these as the starting point for a discussion forum. Currently, the identification of claims and hypotheses from the underlying texts is a manual process, but initiatives are underway to help automate this process [Das et al., 2008].

The point of these developments is that when a claim-evidence structure has been populated as suggested here and published online, a benefit becomes available to communities of practice in the research, clinical, and educational spaces. We believe that a concise collection of claims and relations is suited to the social gestures available at hypothesis discussion sites such as Cohere or SWAN. In this scenario, elements of HypER structures become information resources that support annotations that identify claims, questions, and arguments. These annotations are addressable information resources separate from the HypER documents, but linked to them. In such systems, web conversations are started when annotations are connected to other annotations with coherence relations. For example, Issue-based Information System (IBIS) conversations relying on dialogue mapping [Conklin, 2005] begin when some of those coherence relations are chosen to answer or ask questions, and to offer arguments in support or refutation of claims made. The benefit is this: conversations external to but anchored in scholarly presentations of scientific research facilitate a wider participation in the research itself and create opportunities where discoveries are made. These conversations can occur within the context of a particular research project as well as engaging comparison among several research projects.

3.2 Identifying hypotheses from papers

There are different approaches used to identify hypotheses from text, either manually, automatically, or semi-automatically; very often, these require discourse parsing as a first step. We discuss a few examples of discourse parsing relevant to our case. Marcu [Marcu, 1999] automatically identifies Rhetorical Structure Theory (RST) relations

[Mann and Thomson, 1987] between elementary discourse units (edu's). The work of Teufel [1999] focuses on finding so-called argumentative zones, which are defined as a [group of] sentences that have the same rhetorical goal. Teufel et al. [1999] identify six such zones, such as those defining 'own' vs. 'other' work; stating the background of a piece of work or its results. Mizuta and Collier [2007] identified similar, but smaller-grained zones in biological texts. Biber and Jones [2005] define a collection of biological Discourse Units, and the respective Discourse Unit Type by various linguistic markers.

Instead of characterizing the discourse structure of research articles, Sándor [2007] aims at detecting rhetorical metadiscourse functions that are attached to propositions in biology articles. Rhetorical metadiscourse functions are recurring comments that authors formulate in order to indicate the epistemic value of the propositions, i.e. their status with respect to the state of the art. The status of a proposition may be for example that of a substantially new finding; the author may want to state that a particular solution is not known; a statement may serve as background knowledge; it may be a contradiction, or a new research tendency. The analysis is carried out with the XIP dependency parser [Aït-Mokhtar et al., 2002].

Another approach to discourse parsing is the creation of discourse annotation tools that allow manual discourse parsing. For instance, in the Cohere tool [Buckingham Shum, 2008], authors (or users of the system) manually create their claims, and link them by hand. In SALT, authors identify their own claims, as well; the more recent KonneX platform uses Latent Semantic Indexing to identify relevant conclusions [Groza, 2008]. The SWAN project uses annotators to assign pertinent hypotheses [Ciccarese et al., 2008] and allows discussions based on these hypotheses. Some thoughts on the effects of these developments for the identification of epistemic value are discussed elsewhere [Carusi & De Waard, submitted].

In [De Waard, 2007], a model for structuring the rhetoric with marked-up discourse units is proposed, that aims to support future processing of rhetorically-structured biology texts; preliminary experiments to expand this model to add epistemic value to sentences in PubMed abstracts have been promising [De Waard et al., 2009]. Other recent research and development can contribute to the reduction in hand curation. IBM introduced the open source Unstructured Information Management Architecture (UIMA) and applied it to biomedical documents [Uramoto et al, 2004]. Etzioni et al. have, through their KnowItAll and TextRunner projects, published numerous papers related to the harvesting of relational information from web documents [Etzioni et al., 2005]. The evolution of Direct Memory Access Parsing (DMAP) [Livingston & Reisbeck, 2007] and its OpenDMAP open source product [Hunter et al., 2008], are aimed at accelerating biomedical discovery. With NeuroScholar [Burns and Cheng, 2006] and a range of other open source bioinformatics tools available, we see opportunities for direct application to the HypER project.

Concerning hypothesis-centric data formats, various standards are emerging for representing such discourse, including the standards used for SWAN, Cohere, and the RDFa-based aTags. aTags [aTags] are a convention for using Semantic Web technologies and standards for simple representation of annotated assertions in web environments. They are based on the RDFa syntax [RDFa] and the Semantically

Interlinked Online Communities (SIOC) vocabulary [SIOC], which makes it possible to embed aTag annotations into normal web pages. The statements, their links to evidence and annotations with ontology terms can be processed by Semantic Web tools, enabling the rapid integration of statements from different sources. In a preliminary trial, we generated aTags for a small corpus of biomedical abstracts through manual curation [Samwald and Stenzhorn, 2009]. Furthermore, aTags were extracted from conclusion sections of PubMed abstracts and were made available for faceted browsing [Samwald, 2009]. In future work, we will further explore the practicability and expressivity of this simple representation of statements, and will compare this approach with other systems.

4 Conclusion

We are firmly convinced that it is time for information technologies to push beyond what we can do (extracting triplets) towards what we should do: create and extract a knowledge model which works for humans to make sense of the vast information environment they are engulfed by. We believe the most promising way forward is to use and combine elements from the tools described above, and hope that a combined effort can help overcome the objections to each individual technique. First of all, we plan to build on our discourse elements model, and try to identify linguistic markers that might enable automatic identification of rhetorical elements within biological text. If automatically defined, rhetorical elements might also help prepopulate argumentation visualization tools with claims and assertions. We are planning a multi-disciplinary collaboration, to develop a common framework for identifying, defining, and relating hypotheses in scientific text.

As some of the groups involved in the tools and technologies described in the previous section are connecting to each other, we are interested in exploring a a platform that can support this richer, more argumentation-focused approach to representing scientific knowledge. This approach could help align research in argumentation, computational linguistics, sociology of science, hypermedia, semiotics, and semantic and pragmatic web sciences. At our website, http://hyp-er.wik.is, first steps made towards, for instance, bringing various discourse representations inline, making attempts to coordinating efforts for the automatic identification of hypotheses, and developing a model for a 'hypothesis-centric' conference paper. We look forward to continuing these efforts, and invite members of the Semantic Web and Computational Linguistic communities to join forces with us.

References

1. Aït-Mokhtar, S., Chanod, J.P. and Roux, C. (2002). Robustness beyond shallowness: incremental dependency parsing. Nat. Lang. Eng., 8(2/3):121-144.
2. Atags: http://hcls.deri.org/atag/data/microRNA_atags.html
3. Bazerman, C. 1988. Shaping written knowledge: the genre and activity of the experimental article in science, Madison, Wisconsin: Univ. of Wisconsin Press, 1988.

4. Bex, F., Prakken, H., Reed, C., & Walton, D. (2001). Towards a Formal Account of Reasoning about Evidence: Argumentation Schemes and Generalisations, 1-28.
5. Biber, D., & Jones, J. K. (2005). Merging corpus linguistic and discourse analytic research goals: Discourse units in biology research articles. Corpus Linguistics and Linguistic Theory, 2(2005).
6. Boivin A, Green J, van der Meulen J, Légaré F, Nolte E (2009). Why consider patients' preferences? A discourse analysis of clinical practice guideline developers. Med Care. 2009 Aug; 47(8):908-15.
7. Buckingham Shum, S. (2008). Cohere: Towards Web 2.0 Argumentation. Proc. 2nd Int. Conf. Computational Models of Argument, 28-30 May 2008, Toulouse: IOS Press.
8. Buckingham Shum, S.J., Uren, V., Li, G., Sereno, B. and Mancini, C. (2007). Modeling Naturalistic Argumentation in Research Literatures: Representation and Interaction Design Issues. Int. Jnl. of Intelligent Systems, (Spec. Issue on Comp. Models of Natural Argument, Eds: C. Reed and F. Grasso, 22, (1), pp.17-47).
9. Burns, GAPC, Cheng, W-C, Thompson, RH and Swanson, The NeuARt II system: a viewing tool for neuroanatomical data based on published neuroanatomical atlases, BMC Bioinformatics. 2006; 7: 531.
10. Carusi, A., De Waard, A., (submitted). Changing modes of scientific discourse analysis, changing perceptions of science, Submitted to IEEE E-Science Conf., Workshop, Web Semantics in Action, December 11-12, 2009.
11. Chi, Ed H., Pirolli, Chen, and Pitkow (2001). Using Information Scent to Model User Information Needs and Actions on the Web. In Proc. of ACM CHI 2001 Conf. on Human Factors in Computing Systems, pp. 490-497. ACM Press, April 2001. Seattle, WA.
12. Ciccarese, P., Wu, E., Wong,G. Ocana, M. Kinoshita, J., Ruttenberg, A. Clark, T. (2008) The SWAN biomedical discourse ontology, JBiomed Inf. 2008 Oct;41(5):739-51.
13. Conklin, J. (2005) Dialogue Mapping: Building Shared Understanding of Wicked Problems. Wiley.
14. Corbett, Edward P.J. & Connors, Robert J. (1999) Classical Rhetoric for the Modern Student, 4th Edition. New York & Oxford: Oxford University Press.
15. Das, S. Green, T. Weitzman, L. Lewis-Bowen, A. & Clark, T. 2008. Linked Data in a Scientific Collaboration Framework. 17th Int'l WWW Conf. (WWW2008), Beijing, China
16. Dinakarpandian, D., Lee, Y, Vishwanath,K., Lingambhotla, R. 2006. MachineProse: An Ontological Framework for Scientific Assertions, J Am Med Inf. Assoc. 2006;13:220-232.
17. Stephan Gross, The Rhetoric of Science, (Harvard University Press, 1996)
18. Groza, T., Handschuh, S. Möller, K. and Decker, S. (2008) KonneXSALT: First Steps Towards a Semantic Claim Federation Infrastructure. The Semantic Web: Research and Applications, LNCS 5021, 80-94. Springer
19. Groza, T., Handschuh, S. Clark,T., Buckingham Shum, S. and De Waard, A. (submitted). A Short Survey of Discourse Representation Models, Submitted to the ISWC Workshop on Scientific Discourse Representation, 2009.
20. Etzioni, O., Cafarella, Downey, M., Popescu, D., Tal, A.-M., Weld, S.W, and Yates, A. (2005) Unsupervised named-entity extraction from the Web: An experimental study Artificial Intelligence, 165(1):91-134, 2005
21. Hirschman, L., Yeh, A. Blaschke, C. and Valencia, A., Overview of BioCreAtIvE: critical assessment of information extraction for biology, BMC Bioinformatics 2005, 6 (Suppl. 1):
22. Hovy, E. H. (1993). Automated discourse generation using discourse structure relations. Artificial Intelligence, 63, 341-385.
23. Hunter, L. and Bretonnel Cohen, K. 2006. Biomedical Language Processing Perspective: What's Beyond PubMed? Molecular Cell 21, 589-594, March 3, 2006.
24. Hunter L, Lu Z, Firby J, Baumgartner WA Jr, Johnson HL, Ogren PV, Cohen KB. (008). OpenDMAP: an open source, ontology-driven concept analysis engine. BMC Bioinformatics. 2008 Jan 31;9:78

25. Jensen, L. J., Saric, J., & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. Group, 7 (Feb), 119-129.
26. Kwon, N., Hovy, E. H., Zhou, L., & Shulman, S. W. (2006). Identifying and Classifying Subjective Claims. The 7th Annual International Conference on Digital Government Research 06, May 21-24, 2006, San Diego, CA, USA.
27. Latour, B., and Woolgar, S. (1979). Laboratory Life: The Social Construction of Scientific Facts. Beverly Hills: Sage, 1979.
28. Livingston, K., and Riesbeck, C.K. (2007). Using Episodic Memory in a Memory Based Parser to Assist Machine Reading, AAAI Spring Symp. on Mach. Reading, AAAI Press.
29. Mancini, C. and Buckingham Shum, S.. (2006). Modeling discourse in contested domains: A semiotic and cognitive framework. Int. Jnl. Hum-Comp.Studies, 64(11), pp. 1154-1171.
30. Mann, W. C. and Thompson, S. A. 1987. Rhetorical Structure Theory: A Theory of Text Organization, University of Southern California, Information Sciences Institute (ISI), Report Number ISI/RS-87-190, June 1987.
31. Marcu, D. (1999) A decision-based approach to rhetorical parsing, Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, p.365-372, June 20-26, 1999, College Park, Maryland.
32. MEDIE: http://www-tsujii.is.s.u-tokyo.ac.jp/medie/search.cgi
33. Mizuta, Y. and Collier, N. (2004) An Annotation Scheme for a Rhetorical Analysis of Biology Articles, in Proceedings of the Fourth Intl. Conference on Language Resources and Evaluation (LREC 2004).
34. Park, J. (2008). Knowledge Gardening as Knowledge Federation. Proceedings Knowledge Federation 2008, Dubrovnik, Croatia, October, 2008
35. Popping, R. (2000). Computer-assisted Text Analysis. Newbury Park / London: Sage. Quillan, R. M. (1968). Semantic memory. In M. Minsky (Ed.), Semantic Information Processing, pp. 216–270. Cambridge, MA: MIT Press.
36. RDF-A: http://www.w3.org/TR/rdfa-syntax/
37. Rebholz-Schuhmann, D., Kirsch, H., & Couto, F. (2005). Facts from Text: Is Text Mining Ready to Deliver? PLoS Biology, 3(2), 188-191.
38. Roberts, C. W. (Ed.) (1997). Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcript. Mahwah, NJ: Lawrence Erlbaum.
39. Roos, M., Marshall, S., Gibson, A.P., Schuemie, M., Meij, E. Katrenko, S., van Hage, W., Krommydas, K. and Adriaans, P. (2009). Structuring and extracting knowledge for the support of hypothesis generation in molecular biology, BMC Bioinf. 10 (S10), Oct 2009.
40. Samwald, M. (2009) Extracting conclusion sections from PubMed abstracts for rapid key assertion integration http://precedings.nature.com/documents/3775/version/1
41. Samwald, M. and Stenzhorn, H. (2009) "Simple, ontology-based representation of biomedical statements through fine-granular entity tagging and new web standards" The 12th Annual Bio-Ontologies Meeting, June 28, 2009, at ISMB 2009, Stockholm, Sweden.
42. Sándor, Á. 2007. Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts. Revue Française de Linguistique Appliquée 200(2):97-109.
43. Schiffrin, D., Tannen D. and Hamilton, H.E. (eds). The Handbook of Discourse Analysis. Blackwell Publishing, 2003.
44. Schoop, M., de Moor, A. and Dietz J.L.G. (2006) The pragmatic web: a manifesto. Communications of the ACM, 49(5) 75 – 76, May 2006.
45. Seringhaus, M. R., & Gerstein, M. B. (2007). Publishing perishing? Towards tomorrow's information architecture, BMC Bioinformatics. BMC Bioinformatics, 5, 1-5, 2007.
46. SIOC: http://sioc-project.org/
47. Stewart, C. O. (2008) Social cognition and discourse processing goals in the analysis of `ex-gay' rhetoric, Discourse & Society, Vol. 19, No. 1, 63-83 (2008)
48. Teufel, S. (1999) Argumentative Zoning: Information Extraction from Scientific Text, PhD Thesis, University of Edinburgh, 1999

49. Teufel, S., J. Carletta and Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles, In: Proceedings of EACL 1999.
50. Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H. and Takeda, K. (2004) A text-mining system for knowledge discovery from biomedical documents, IBM Systems Journal, Vol.43, No.3, pp.516-533, 2004
51. Van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N. and Schlobach, S. (2008), Good News or Bad News? Conducting sentiment analysis on Dutch text to distinguish between positive and negative relations, in: C. Cardie and J. Wilkerson (eds.), Spec. Iss. J. of Inf. Technology and Politics, vol. 5 (1), pp. 73-94.
52. Van den Braak, S.W., Van Oostendorp, H., Prakken, H., & Vreeswijk, G.A.W. A critical review of argument visualization tools: Do users become better reasoners? Workshop notes of the ECAI-06 Workshop on Computational Models of Natural Argument, Riva del Garda (Italy), August 28-29, 2006.
53. Waard, A. de, Buitelaar, P. Eigner, T. 2009. Identifying the Epistemic Value of Discourse Segments in Biology Texts. Int. Workshop on Computational Semantics, Tilburg, February, 2009.
54. Waard, A. de. (2007). A pragmatic structure for research articles. In Proceedings of the 2nd international Conference on Pragmatic Web (Tilburg, The Netherlands, October 22 - 23, 2007). ICPW '07, vol. 280. ACM, New York, NY, 83-89.
55. Waard, A. de; Breure, L.; Kircz, J.G.; Oostendorp, H. van (2006). Modeling Rhetoric in Scientific Publications. Current Res. in Inf. Sci. and Techn. pp. 352-356, 2006.
56. Wooffitt, R. (2005) Conversation Analysis and Discourse Analysis : a Comparative and Critical Introduction, Sage Publications, London, 2005
57. Yan W, Liu G, Scoumanne A, Chen X. (2008). Suppression of inhibitor of differentiation 2, a target of mutant p53, is required for gain-of-function mutations. Canc Res 68: 6789-96