Scientific Authoring Support: A Tool to Navigate in Typed Citation Graphs

Ulrich Schäfer
Language Technology Lab
German Research Center for
Artificial Intelligence (DFKI)
D-66123 Saarbrücken, Germany
ulrich.schaefer@dfki.de

Uwe Kasterka
Computer Science Department
Saarland University
Campus
D-66123 Saarbrücken, Germany
uwe.kasterka@dfki.de

Abstract

Scientific authors urgently need help in managing the fast increasing number of publications. We describe and demonstrate a tool that supports authors in browsing graphically through electronically available publications, thus allowing them to quickly adapt to new domains and publish faster. Navigation is assisted by means of typed citation graphs, i.e. we use methods and resources from computational linguistics to compute the kind of citation that is made from one paper to another (refutation, use, confirmation etc.). To verify the computed citation type, the user can inspect the highlighted citation sentence in the original PDF document. While our classification methods used to generate a realistic test data set are relatively simple and could be combined with other proposed approaches, we put a strong focus on usability and quick navigation in the potentially huge graphs. In the outlook, we argue that our tool could be made part of a community approach to overcome the sparseness and correctness dilemma in citation classification.

1 Introduction and Motivation

According to different studies, the number of scientific works is doubled every 5-10 years. Important issues to be addressed by the scientific community are finding relevant information and avoiding redundancy and duplication of work. The organization and preservation of scientific knowledge in scientific publications, vulgo text documents, thwarts these efforts. From a viewpoint of a computer scientist, scientific papers are just 'unstructured information'.

One specific, but very important aspect of the content of scientific papers is their relation to previous work and, once published, their impact to subsequent or derived research. While it is still hard if not impossible to capture and formalize the semantic content of a scientific publication automatically, at least citation properties and derived scientific impact can be and usually are measured automatically on the basis of simple citation graphs. In other words, these graphs can be used to describe I/O behavior of publications in a very simple way.

However, just counting citations is a very coarse approach and does not tell much about the reasons for citing one's work in a specific situation. Moreover, once such measure is formalized and standardized e.g. for science evaluation, it can be exploited to tune up statistics. Since the first proposal of the Science Citation Index (Garfield, 1955), it has also provoked criticism.

In the bibliometrics and computational linguistics literature, many proposals are available on how citations could be further classified by careful analysis of citation sentences and context (Garfield, 1965; Garzone, 1996; Mercer and Di Marco, 2004; Teufel et al., 2006; Bornmann and Daniel, 2008).

The number of different classes proposed varies from 3 to 35. Different authors try to identify dimensions and mutually exclusive classes, but the more classes a schema contains, the more difficult becomes the automatic classification.

The focus of our paper is to combine automatic classification approaches with a tool that supports scientists in graphically navigating through *typed citation graphs (TCG)*. Such TCGs can be generated

Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing, pages 7–14,
Los Angeles, California, June 2010.

by augmenting a simple citation graph with information synonymously called citation *function* (Teufel et al., 2006), citation *relation* (Mercer and Di Marco, 2004) or citation *sentiment*, forming the labels of the graph's edges. In the following, we use the more neutral and general term *citation type*.

The idea is to help scientists, especially those not so familiar with an area, understanding the relations between publications and quickly get an overview of the field. Moreover, the goal is to embed this tool in a larger framework for scientists that also supports semantic search assisted by domain ontologies and further tools for authoring support (Schäfer et al., 2008).

Our paper is structured as follows. In Section 2, we describe how we automatically compute the typed citation graph from the raw text content of a scientific paper corpus to generate realistic data for testing the visualization and navigation tool. Section 3 contains an evaluation of the quality of the extracted unlabeled graphs and of the citation classification step. We then describe in Section 4 the ideas of efficient and at the same time well-arranged visualization and navigation in the typed citation graph. We compare with related work in Section 5. Finally, we conclude and give an outlook to future work in Section 6.

2 Data Preparation and Automatic Citation Type Classification

Our corpus is based on 6300 electronically-available papers, a subset (published 2002-2008) of the ACL Anthology (Bird et al., 2008), a comprehensive collection of scientific conference and workshop papers in the area of computational linguistics and language technology.

The overall workflow of the employed tools and data is shown in Fig. 1.

We ran the open source tool ParsCit (Councill et al., 2008) to extract references lists and corresponding citation sentences from raw paper texts. To build the citation graph, we used the Levenshtein distance (Levenshtein, 1966) to find and match titles and authors of identical papers yet tolerating spelling and PDF extraction errors.

To increase robustness, publication years were not considered as they would hinder matches for

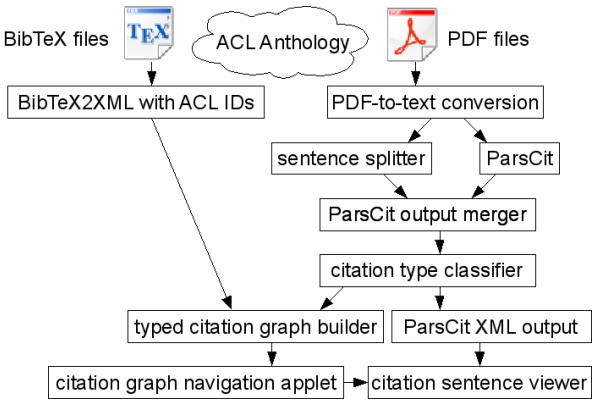

Figure 1: Workflow from ACL Anthology data (top) to citation graph navigation applet and citation sentence viewer (bottom)

delayed journal publications. Generation of the graph edges, i.e. matching of papers and reference strings, is performed by means of the ACL ID, a unique identifier for each paper, available for the PDF (source nodes of references) and BibTeX files (targets of references).

We evaluated the generated graph against the one that was corrected manually by the ACL Anthology Network (AAN) group (Radev et al., 2009) and found that 10821 citation links were shared between both and can be considered correct[1].

3883 additional ones were in the AAN but not recognized by us, the other way round, 1021 discovered by us were not in the AAN. In addition, the publication bases were not identical. The anthology network data ends in February 2007 but covers years before 2002, while our data covers 2002-2008 inclusively. Given the fact that our graph is computed fully automatically, the result can be considered very good.

In the next step, we augmented the citation graph by types for each edge. In contrast to other approaches, we currently only consider the citation sentence itself to determine the citation type, neither a wider context, its position nor the abstract, title or content of the cited paper. A reference (from the references section at the end of a paper) may be associated with several citation sentences mentioning the paper referenced at the end.

[1]We only consider intra-network links here, not those pointing to books or other publications outside the corpus.

In only considering the citation sentence itself, we may lose some citation type information, as it may be (also) contained in follow-up sentences referring to the citation using a pronoun ("they", "their approach" etc.). Considering follow-up or even preceding sentences is planned to be addressed in future work.

After consulting the rich literature on citation classification (Bornmann and Daniel, 2008; Garzone, 1996; Teufel et al., 2006), we derived a simplified classification schema consisting of the following five classes.

- Agree: The citing paper agrees with the cited paper
- PRecycle: The citing paper uses an algorithm, data, method or tool from the cited paper
- Negative: The paper is cited negatively/contrastively
- Neutral: The paper is cited neutrally
- Undef: impossible determine the sentiment of the citation (fallback)

Then, we used a blend of methods to collect verbal and non-verbal patterns (cue words) and associated each with a class from the aforementioned schema.

- A list from (Garzone, 1996) devised for biomedical texts; it is largely applicable to the computational linguistics domain as well.
- Simple negation of positive cue words to obtain negative patterns.
- A list of automatically extracted synonyms and antonyms (the latter for increasing number of patterns for negative citations) from WordNet (Miller et al., 1993).
- Automatically computed most frequent cooccurrences from all extracted citation sentences of the corpus using an open source cooccurrence tool (Banerjee and Pedersen, 2003).
- Inspection: browse and filter cue words manually, remove redundancies.

3 Results: Distribution and Evaluation

These pattern where then used for the classification algorithm and applied to the extracted citation sentences. In case of multiple citations with different classes, a voting mechanism was applied were the 'stronger' classes (Agree, Negative, PRecycle) won in standoff cases. For the total of 91419 citations we obtained the results shown in Table 1.

Classes	Citations	Percent
Agree	3513	3.8%
Agree, Neutral	2020	2.2%
Negative	1147	1.2%
PRecycle	10609	11.6%
PRecycle, Agree	1419	1.6%
PRecycle, Agree, Neutral	922	1.0%
PRecycle, Neutral	3882	4.2%
Neutral	13430	14.7%
Undef	54837	60.0%

Table 1: Citation classification result

The numbers reflect a careful classification approach where uncertain citations are classified as Undef. In case of multiple matches, the first (leftmost) was taken to achieve a unique result.

The results also confirm obervations made in other works: (1) citation classification is a hard task, (2) there are only a few strongly negative citations which coincides with observations made by (Teufel et al., 2006), (Pendlebury, 2009) and others, (3) the majority of citations is neutral or of unknown type.

An evaluation on a test set of 100 citations spread across all the types of papers with a manual check of the accuracy of the computed labels showed an overall accuracy of 30% mainly caused by the fact that 90% of undefined hits were in fact neutral (i.e., labeling all undefs neutral would increase accuracy). Negative citations are sparse and unreliable (33%), neutral ones are about 60% accurate, PRecycle: 33%, Agree: 25%.

To sum up, our automatic classification approach based on only local citation information could surely be improved by applying methods described in the literature, but it helped us to quickly (without annotation effort) generate a plausible data set for the main task, visualization and navigation in the typed citation graphs.

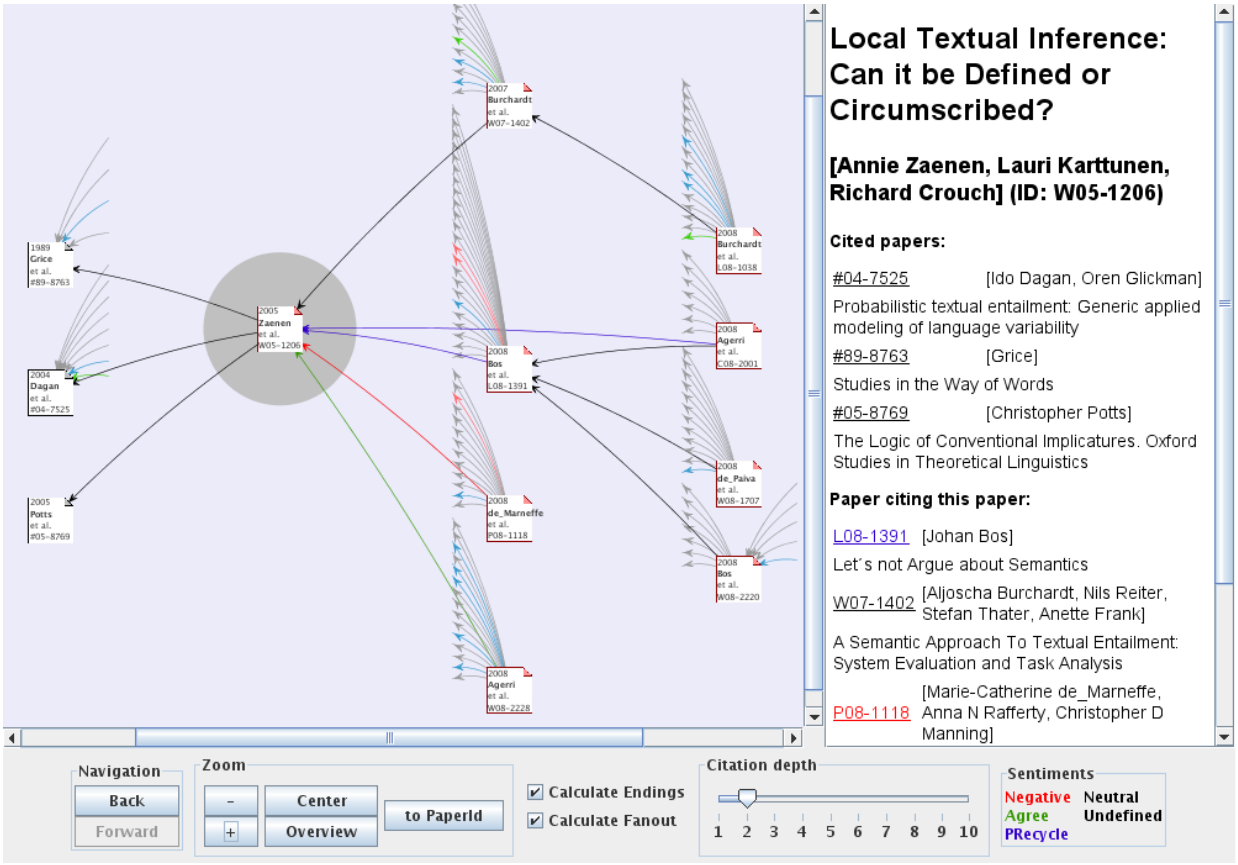

Figure 2: Typed citation graph navigator applet

4 Visualization Algorithm and Navigation User Interface

The overall idea of the citation visualization and navigation tool is simple and intuitive. Each paper is represented by a node, all citations between papers are drawn as edges between nodes where the color of the edge indicates the computed (overall) citation type, e.g. green for agree, red for negative, blue for recycle and black for neutral or undefined.

To cope with flexible layouts and scalability of the graph, we decided to use the open source tool Java Universal Network/Graph Framework (JUNG, http://jung.sourceforge.net). Its main advantages over similar tools are that it supports user interaction (clicking on nodes and edges, tool tips) and user-implemented graph layout algorithms. A screenshot of the final user interface is presented in Figure 2.

The decision for and development of the visualization and navigation tool was mainly driven by the fact that citation graphs quickly grow and become unmanagable by humans when extended to the transitive closures of citing or cited papers of a given publication. The sheer number of crossing edges would make the display unreadable.

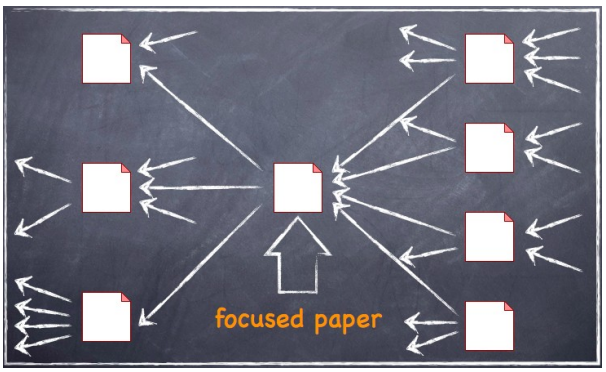

Figure 3: Focused paper in the center

The main design goal therefore was reducing the number of nodes and edges where possible and (by default) have only one paper in focus (Fig. 3), with

all cited papers on the left side (Fig. 4), and all citing papers on the right (Fig. 5).

This also reflects a time line order where the origin (oldest papers) is on the left. In the graphical user interface, the citation depth (default 1) is adjustable by a slider to higher numbers. The graph display is updated upon change of the configured depth.

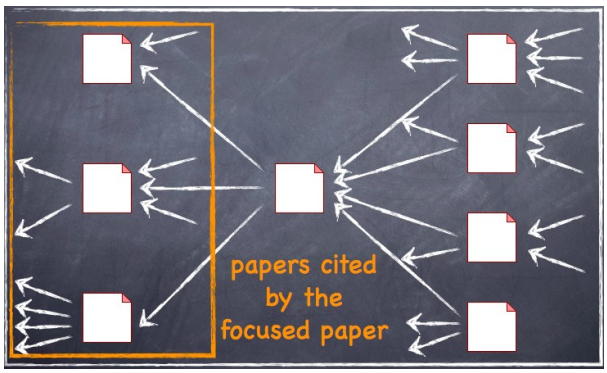

Figure 4: Papers cited by the focused paper

At level 1, papers citing the citing papers (analogously for cited papers), are not fully drawn as nodes, but only adumbrated by short ingoing or outgoing edges (arrows). However, the color of these short edges still signifies the citation type and may attract interest which can easily be satisfied by clicking on the edge's remaining node (cf. screenshot in Figure 2). When the mouse is moved over a node, a tooltip text display pops up displaying full author list and paper title.

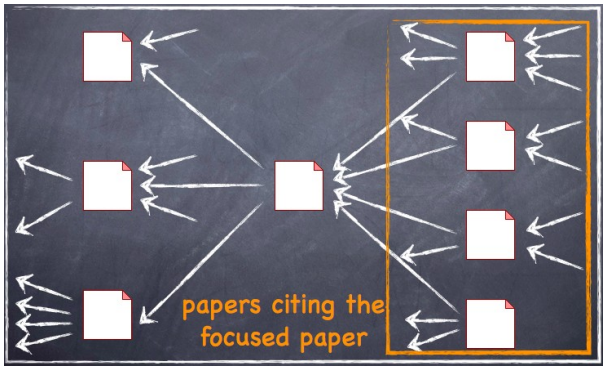

Figure 5: Papers citing the focused paper

To avoid crossing edges caused by citations at the same level (citing or cited papers also cite each other), we devised a fan-out layout generation (Fig. 6). It increases the width of the displayed graph, but leads to better readability. Fan-out layout can also be switched off in the user interface.

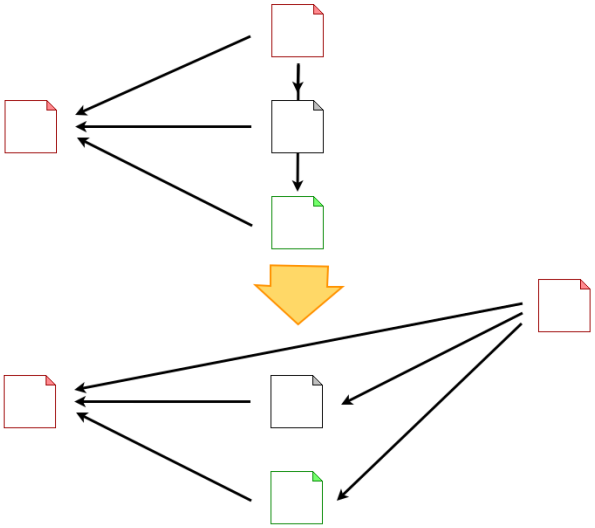

Figure 6: Fan-out layout: avoid crossing edges caused by citations on the same level

In addition, the graph layout algorithm orders papers chronologically in the vertical direction. Here, we have implemented another technique that helps to avoid crossing edges. As shown in Fig. 7, we sort papers vertically by also taking into account the position of its predecessor, the cited paper. It often leads to less crossing edges.

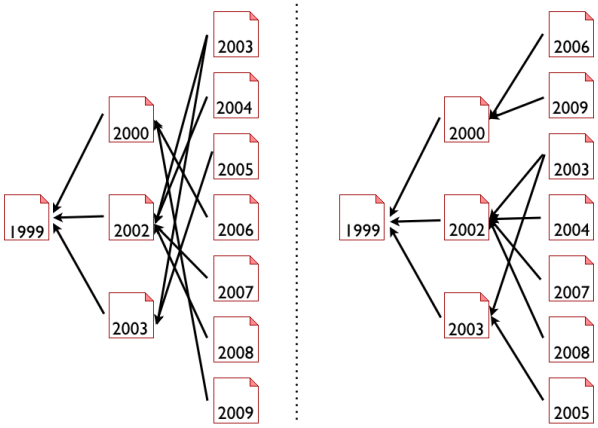

Figure 7: Order: avoid crossing edges by ordering chronologically (strict, simple variant on the left for comparison), taking into account the position of the cited paper on the previous level (right)

By double-clicking on any node representing a paper (cited or citing), this node can be made the new center and the graph is re-arranged accordingly.

Zooming in and out is possible via mouse wheel or shortcut buttons ('overview', 'center').

Using the right mouse button context menu on a node, it is possible to open a details page for the selected paper with bibliographic metadata and all citations and types. All references in the document with their citation sentences identified are displayed in a structured list.

The citation context around a citation sentence is shown as well, while the citation sentence itself is colored according to the citation type color and clickable. If clicked, the original PDF document opens with the citation sentence highlighted (Fig. 8; currently only possible in Acrobat Reader).

By clicking on an edge instead of a node, only the citations between the two papers at both ends are displayed, in the same way as described above for all citations of a document.

5 Related Work

Our paper touches and combines results of three disciplines, (1) bibliometrics, (2) computational linguistics, and (3) information visualization. We briefly discuss related and mostly recent literature, being aware of the fact that this list is necessarily incomplete.

(Garfield, 1965) is probably the first to discuss an automatic computation of citation types. He is also the founder of citation indexing and the Institute of Scientific Information (ISI). His first publication on science citation indexing appeared in 1955 (Garfield, 1955) and he remained the probably most influential scientist in this field for decades. (Bornmann and Daniel, 2008) is a comprehensive recent metastudy on citing behavior.

Investigating citation classification has a long tradition in bibliometrics and information science and in the last 20 years also attracted computational linguistics researchers trying to automate the task based on rhetorics of science, statistical methods and sentence parsing.

There is much more work than we can cite here on citation function computation worth combination with our approach (Bornmann and Daniel, 2008; Garzone, 1996; Teufel et al., 2006) – using our tool one can easily browse to further publications!

There is little work on innovative layout techniques for displaying and navigating citation graphs. We found three independent approaches to citation graph visualization: CiteViz (Elmqvist and Tsigas, 2004), CircleView (Bergström and Jr., 2006), and (Nguyen et al., 2007). They share a disadvantageous property in that they try to visualize too much information at the same time. In our opinion, this contradicts the need to navigate and keep control over displayable parts of large paper collections.

Moreover, these approaches do not provide information on citation types derived from text as our system does. Further ideas on visualizing science-related information such as author co-citation networks are also discussed and summarized in (Chen, 2006).

6 Summary and Outlook

We have presented an innovative tool to support scientific authors in browsing graphically through large collections of publications by means of typed citation graphs. To quickly generate a realistic data set, we devised a classification approach avoiding manual annotation and intervention.

Our classification results cannot compete with approaches such as (Teufel et al., 2006) based on considerable manual annotation for machine learning. However, we think that our application could be combined with this or other approaches described for classifying citations between scientific papers.

We envisage to integrate the navigation tool in a larger framework supporting scientific authoring (Schäfer et al., 2008). When publishing a service of this kind on the Web, one would be faced with ethical issues such as the problem that authors could feel offended by wrongly classified citations.

The reason is that citation type classification is potentially even more subjective than a bare citation index—which itself is already highly controversal, as discussed in the introduction. Moreover, there is not always a single, unique citation type, but often vagueness and room for interpretation.

Therefore, we suggest to augment such a service by a Web 2.0 application that would allow registered users to confirm, alter and annotate precom-

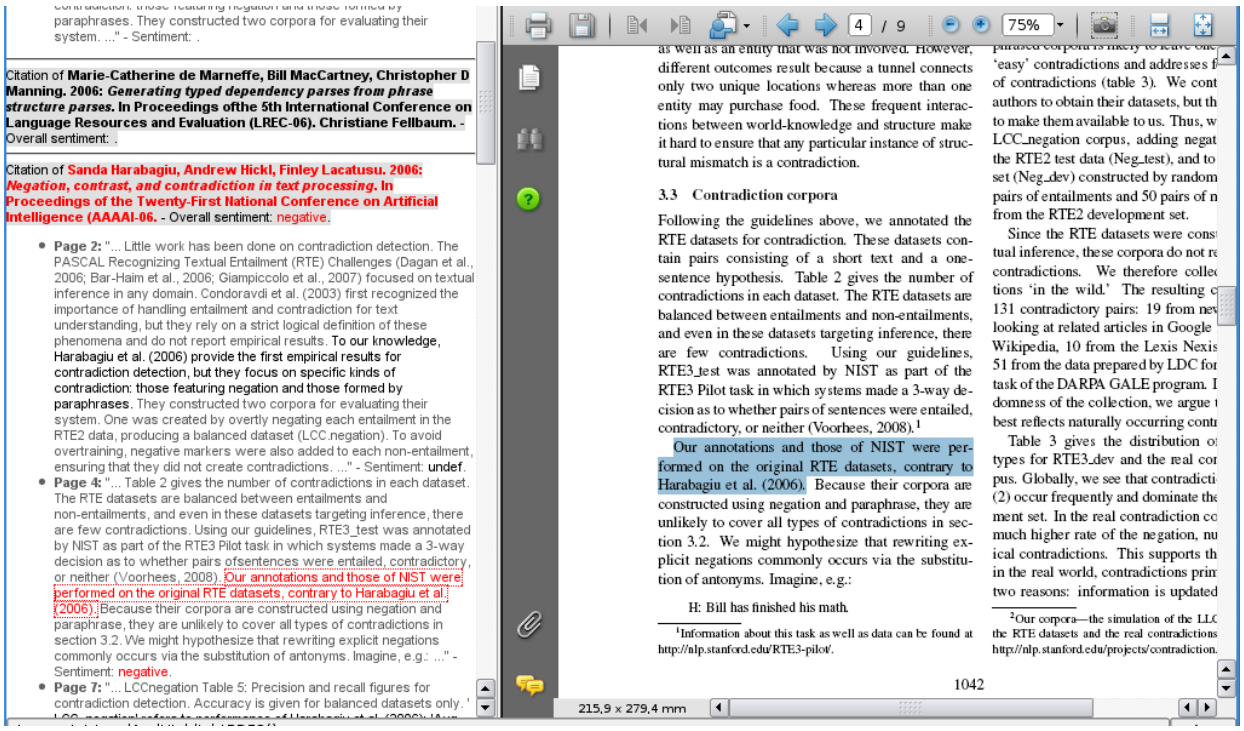

Figure 8: Citation sentence viewer; citation sentence in context on the left, highlighted in PDF on the right when selected on the left

puted citation classifications. In this community application, all citation links in the automatically generated graph could be represented by dashed arrows initially, and users could turn them solid by confirming or correcting the citation type and also adding a comment text.

Line thickness could be increased (up to an appropriate maximum) each time another user confirms a classified citation type. The results could then also be employed for active learning and help to improve the automatic classification procedure.

Acknowledgments

First of all, we are indebted to the anonymous reviewers for their useful, encouraging and detailed comments. Many thanks also to Donia Scott for her feedback on an earlier version of the tool and helpful comments on terminology. We would like to thank Madeline Maher and Boris Fersing for generating and evaluating the citation type data on a subcorpus of the ACL Anthology.

The work described in this paper has been carried out in the context of the project TAKE (Technologies for Advanced Knowledge Extraction), funded under contract 01IW08003 by the German Federal Ministry of Education and Research.

References

Satanjeev Banerjee and Ted Pedersen. 2003. The design, implementation, and use of the ngram statistics package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381, Mexico City.

Peter Bergström and E. James Whitehead Jr. 2006. CircleView: Scalable visualization and navigation of citation networks. In *Proceedings of the 2006 Symposium on Interactive Visual Information Collections and Activity IVICA*, College Station, Texas.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research. In *Proceedings of the Language Resources and Evaluation Conference (LREC-2008)*, Marrakesh, Morocco.

Lutz Bornmann and Hans-Dieter Daniel. 2008. What do citation counts measure? A review of studies on

citing behavior. *Journal of Documentation*, 64(1):45–80. DOI 10.1108/00220410810844150.

Chaomei Chen. 2006. *Information Visualization: Beyond the Horizon*. Springer. 2nd Edition, Chapter 5.

Isaac G. Councill, C. Lee Giles, and Min-Yen Kan. 2008. ParsCit: An open-source CRF reference string parsing package. In *Proceedings of the Language Resources and Evaluation Conference (LREC-2008)*, Marrakesh, Morocco.

Niklas Elmqvist and Philippas Tsigas. 2004. CiteWiz: A tool for the visualization of scientific citation networks. Technical Report CS:2004-05, Department of Computing Science, Chalmers University of Technology and Göteborg University, Göteborg, Sweden.

Eugene Garfield. 1955. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 123:108–111.

Eugene Garfield. 1965. Can citation indexing be automated? In Mary Elizabeth Stevens, Vincent E. Giuliano, and Laurence B. Heilprin, editors, *Statistical Association Methods for Mechanical Documentation*. National Bureau of Standards, Washington, DC. NBS Misc. Pub. 269.

Mark Garzone. 1996. Automated classification of citations using linguistic semantic grammars. Master's thesis, Dept. of Computer Science, The University of Western Ontario, Canada.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Robert. E Mercer and Chrysanne Di Marco. 2004. A design methodology for a biomedical literature indexing tool using the rhetoric of science. In Lynette Hirschman and James Pustejovsky, editors, *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*, pages 77–84, Boston, Massachusetts, USA.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1993. Five papers on WordNet. Technical report, Cognitive Science Laboratory, Princeton University.

Quang Vinh Nguyen, Mao Lin Huang, and Simeon Simoff. 2007. Visualization of relational structure among scientific articles. *Advances in Visual Information Systems*, pages 415–425. Springer LNCS 4781, DOI 10.1007/978-3-540-76414-4_40.

David A. Pendlebury. 2009. The use and misuse of journal metrics and other citation indicators. *Archivum Immunologiae et Therapiae Experimentalis*, 57(1):1–11. DOI 10.1007/s00005-009-0008-y.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL anthology network corpus. In *Proceedings of the ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, Singapore.

Ulrich Schäfer, Hans Uszkoreit, Christian Federmann, Torsten Marek, and Yajing Zhang. 2008. Extracting and querying relations in scientific papers. In *Proceedings of the 31st Annual German Conference on Artificial Intelligence, KI 2008*, pages 127–134, Kaiserslautern, Germany. Springer LNAI 5243. DOI 10.1007/978-3-540-85845-4_16.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia.