A Graphical Citation Browser for the ACL Anthology

Benjamin Weitz, Ulrich Schäfer

German Research Center for Artificial Intelligence (DFKI), Language Technology Lab
Campus D 3 1, Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
email: {benjamin.weitz,ulrich.schaefer}@dfki.de

Abstract

Navigation in large scholarly paper collections is tedious and not well supported in most scientific digital libraries. We describe a novel browser-based graphical tool implemented using HTML5 Canvas. It displays citation information extracted from the paper text to support useful navigation. The tool is implemented using a client/server architecture. A citation graph of the digital library is built in the memory of the server. On the client side, egdes of the displayed citation (sub)graph surrounding a document are labeled with keywords signifying the kind of citation made from one document to another. These keywords were extracted using NLP tools such as tokenizer, sentence boundary detection and part-of-speech tagging applied to the text extracted from the original PDF papers (currently 22,500). By clicking on an egde, the user can inspect the corresponding citation sentence in context, in most cases even also highlighted in the original PDF layout. The system is publicly accessible as part of the ACL Anthology Searchbench.

Keywords: citation analysis, graphical citation browser, scientific digital libraries

1. Introduction

The aim of the ACL Anthology Searchbench[1] (Schäfer et al., 2011) is to provide targeted, efficient search in the digital library of Computational Linguistics and Language Technology, the ACL Anthology[2] (Bird et al., 2008). This is achieved by combining semantic, full-text and bibliographic search. Another way of searching in digital libraries is by citations. Citations are important means to structure the broad publication space. They are of invaluable importance to beginners in a scientific field as they ultimately point to seminal, original work and knowledge not explicitly available or repeated in every publication. Citations are also the primary discourse links in scientific discussion which typically span over years or even decades. Furthermore, citations are helpful to understand and reproduce findings. Thus, they form a predominant feature for every reader.

2. Related and Previous Work

Wan et al. (2009) present a study on user needs for browsing scientific publications and show that citations play an important role. (Harwood, 2009) investigate the role of citations in academic writing. There is a wealth of work in the academic fields of bibliometrics and scientometrics, going back until the seminal work of Eugene Garfield (Garfield, 1955; Garfield, 1965).

The AAN (ACL Anthology Network)[3] (Radev et al., 2009) allows to browse textually and with hyperlinks through a citation graph that has been built for the ACL Anthology. Its advantage is that manual correction has been applied to ensure high quality of the data, but usability of the user interface is limited.

Schäfer and Kasterka (2010) have suggested a novel, graphical user interface for navigating in typed citation graphs. The citations in it are typed which means they were classified according to a schema dividing classifications into categories such as refutal, use, neutral etc. using a simple rule-based approach.

Rule-based classification has drawbacks as reported in that previous work. Approaches with classification based on machine learning have been shown to deliver better results (Teufel et al., 2006; Dong and Schäfer, 2011), yet errors remain, and manual correction is unrealistic here.

Wrong automatic classification may lead to dissatisfied paper authors and system users, so we decided not to show a classification in our public system, but the keywords or phrases that may have lead to classification – with a fallback to finite verb if no specific keyword matched. As these keywords always are verbatim in the citation sentence, they cannot be wrong (however, sometimes potentially misleading).

Other graphical citation navigation tools for large citation networks are presented in Elmqvist and Tsigas (2004) and Bergström and Jr. (2006). These are less suited for inspecting individual publications in detail, but rather focus on larger-scale overviews and statistics.

3. System and Implementation

The overall preprocessing workflow of our new system is shown in Fig. 1. From the HTML output of a commercial PDF-to-text extraction tool, text is input to both the CRF-based citation reference matcher ParsCit (Councill et al., 2008) and a tokenizer and sentence boundary detector. Each sentence containing a citation plus up to three previous and subsequent sentences are then stored together with ParsCit's XML annotation. These can be inspected in the citation context view (Figure 3). The citations extracted by ParsCit are merged with the manually annotated ACL Anthology Network (Radev et al., 2009) in order to cover citations which were not found by ParsCit. The complete graph for the currently 22,500 papers contains approximately 125,000 nodes and about 305,000 edges.

[1] http://aclasb.dfki.de

[2] http://aclweb.org/anthology

[3] http://clair.si.umich.edu/clair/anthology

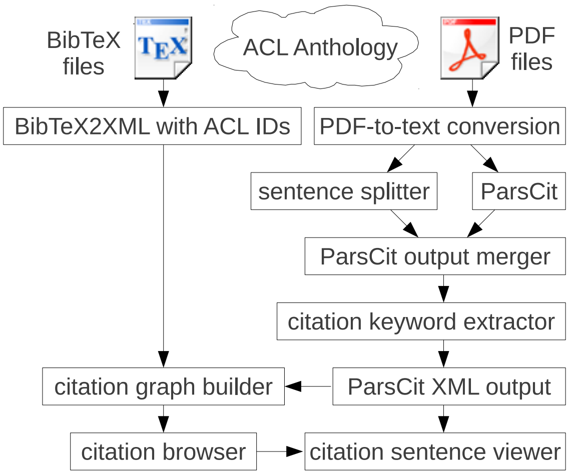


Figure 1: Workflow from ACL Anthology data (top) to citation browser and citation sentence viewer

In addition, matching of bibliographic metadata from the references with BibTeX from the ACL Anthology is performed using Levenshtein distance (Levenshtein, 1966) in order to add the unique ACL document IDs. Through these IDs, interlinking with ACL Anthology, ACL Anthology Searchbench and ACL Anthology Network is achieved by generating appropriate URLs.

The preprocessed information altogether is used to build the citation graph in memory on the server. Loading it for 22,500 papers including all edges and metadata takes less than 10 seconds. On request of the client, the server provides the respective subgraph for the paper of which the citation relations shall be visualized.

The graphical browser on client side is implemented using the Infovis system JIT[4] which is based on HTML5-Canvas[5]. This is advantageous as every modern Web browser, also in smartphones and tablet computers, comes with HTML5/Canvas support built-in. It is more convenient and loads much faster than the previous implementation as a Java applet.

Communication is based on the lightweight Javascript Object Notation, JSON (Crockford, 2002) with transport via HTTP. In the citation graph, each node represents a paper and the edges represent citation relations between the papers.

The layout algorithm is a variant of the fan-out algorithm described by Schäfer and Kasterka (2010). It always has one paper in the center, and cited papers left and citing papers on the right hand side, with arrow heads indicating the citation direction.

The advantage of the fan-out layout is that it avoids overlapping vertical edges in case a citing paper cites another paper that also cites the paper in the center (analogously for cited papers). In such a case, the graph is expanded horizontally to provide space for the intermediate node, instead of arranging it vertically.

In addition, the citation depth for citing and cited papers can be modified (default is 1) in the user interface by adjusting the number with a slider.

Instead of programming the whole graph drawing algorithm from scratch, the fan-out layout is implemented on top of JIT's RadialGraph implementation. Before drawing the graph, the positions of edges and nodes are rearranged according to the fan-out constraints. In addition, another modification replaces straight edges by Bézier curves (Bézier, 1968) in order to avoid overlapping (mostly horizontal) edges. This makes it easier to select labelled edges for inspection of the citation context (cf. next section).

A screenshot is displayed in Figure 2. The nodes show meta-information (first author, year of publication, ACL ID) about the papers they represent, and when hovering them with the mouse, further information such as full author list, title and conference are displayed in a pop-up box. Clicking on a node brings the respective paper into the center with its local citation graph.

4. Displaying Citation Context and Constraining Search

We used the rule-based classifier from Schäfer and Kasterka (2010) and extract the keywords or phrases that would have lead to a classification into a citation class (but do not compute the class itself as mentioned above). In case no pattern matched, the main (finite) verb of the citation sentence is determined using the statistical part-of-speech tagger TnT (Brants, 2000). The resulting keyword is displayed as edge label of the citation link.

Multiple citations to the same reference are not shown on the egde label, but will be enumerated when the user moves the mouse over an egde.

By clicking on an egde, the user can inspect the corresponding citation sentence in context (Figure 3) – in most cases even highlighted in the original PDF layout. The latter requires the Adobe Acrobat Reader plug-in for the Web browser.

One can use the time range slider on the bottom of the user interface to reduce the size of the graph, limiting it to publications in the specified time range. For larger graphs, only a filtered graph is shown by default, because larger graphs can be unclear and confusing and can take long to load on slow systems.

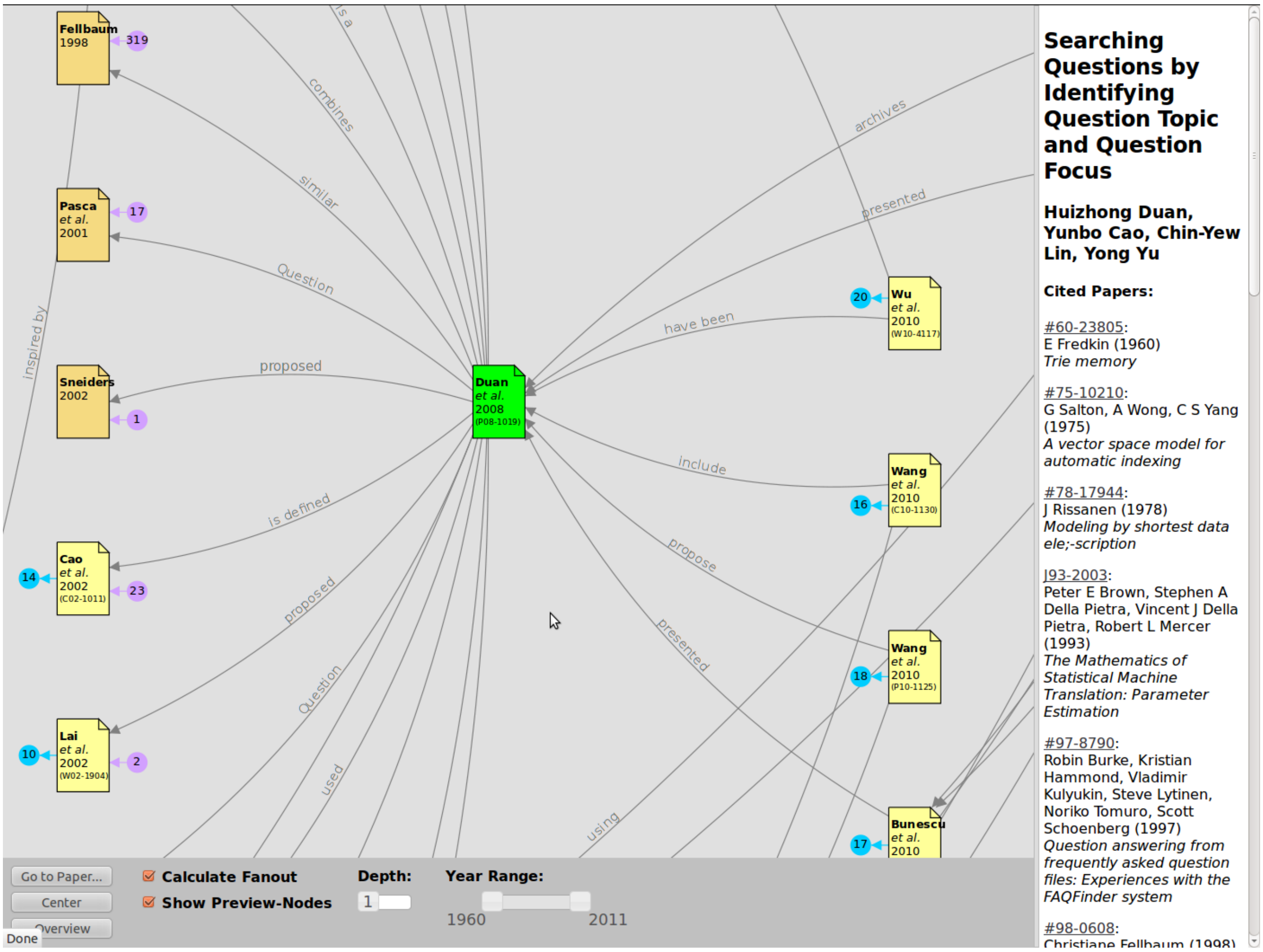
Filtering is done so that the highest possible number of papers below a configurable threshold is displayed using papers from the years around the year of the centered paper. The user can then choose to display a larger graph by using the time range slider.

The citation graph is updated automatically when new papers are added to the Searchbench by being part of the general Searchbench update workflow. An update requires BibTeX, fulltext, Jtok (tokenizer/sentence splitter), TnT (PoS tagger) and ParsCit input.

It generates an extended ParsCit XML with citation sentences and other additional information to be displayed in the graph/citation context viewer, via an XSLT transformation to HTML. An XML example fragment for a single citation context is shown in the Appendix (Figure 4).

[4] http://thejit.org

[5] http://www.w3.org/TR/html-markup



☑ Calculate Fanout
☑ Show Preview-Nodes
Depth: 1

Searching Questions by Identifying Question Topic and Question Focus

Huizhong Duan, Yunbo Cao, Chin-Yew Lin, Yong Yu

Cited Papers:

#60-23805:
E Fredkin (1960)
Trie memory

#75-10210:
G Salton, A Wong, C S Yang (1975)
A vector space model for automatic indexing

#78-17944:
J Rissanen (1978)
Modeling by shortest data ele;-scription

J93-2003:
Peter E Brown, Stephen A Della Pietra, Vincent J Della Pietra, Robert L Mercer (1993)
The Mathematics of Statistical Machine Translation: Parameter Estimation

#97-8790:
Robin Burke, Kristian Hammond, Vladimir Kulyukin, Steve Lytinen, Noriko Tomuro, Scott Schoenberg (1997)
Question answering from frequently asked question files: Experiences with the FAQFinder system

#98-0608:
Christiane Fellbaum (1998)

Figure 2: Citation graph browser

5. Summary

We have implemented a novel user interface for browsing the citation graph of the ACL Anthology. It can be used for research and education in Computational Linguistics and Language Technology. The system is publicly accessible as part of the Searchbench at `http://aclasb.dfki.de` (Citations tab in the document view) and is smoothly interlinked with it, as well as with the ACL Anthology itself and the ACL Anthology network. As the employed technology is mostly domain-independent, the tool could also be applied to digital libraries of other scientific fields.

6. Acknowledgments

The authors would like to thank Jörg Diesinger, Andreas Martin, Christian Spurk and Jörg Steffen for useful comments on the implementation. Christian also helped with the integration into the ACL Anthology Searchbench and its update workflow. We are also grateful to the LREC reviewers and Searchbench users for their feedback. The work described in this paper has been funded by the German Federal Ministry of Education and Research, projects TAKE (FKZ 01IW08003) and Deependance (FKZ 01IW11003).

7. References

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.

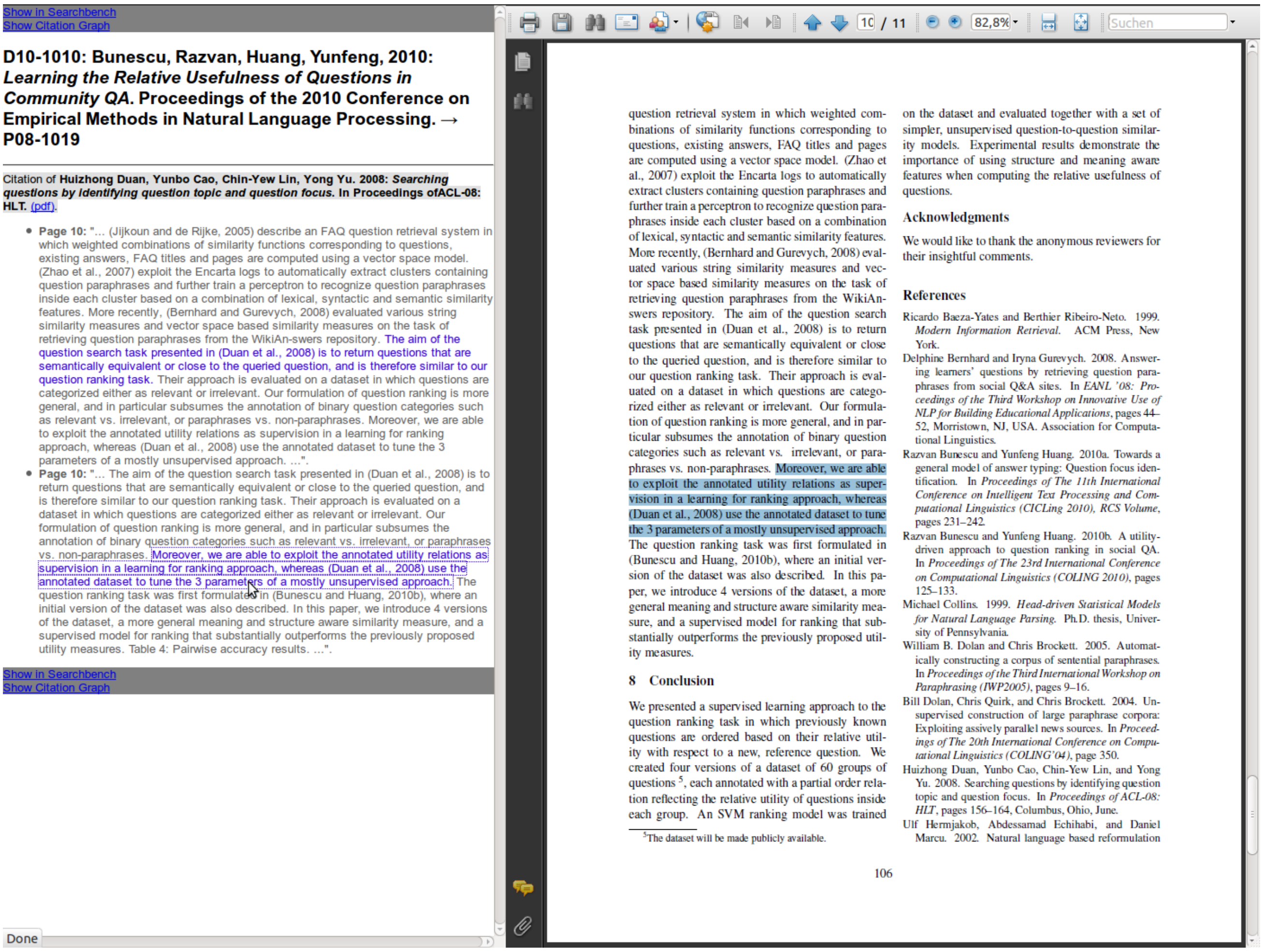
Pierre Bézier. 1968. How Renault uses numerical control for car body design and tooling. In *Society of automotive engineers congress*.

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 LREC-2008*, pages 1755–1759, Marrakesh, Morocco.

Thorsten Brants. 2000. TnT – A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP)*, pages 224–231, Seattle, Washington.

Isaac G. Councill, C. Lee Giles, and Min-Yen Kan. 2008. ParsCit: An open-source CRF reference string parsing package. In *Proceedings of LREC-2008*, pages 661–667, Marrakesh, Morocco.

Douglas Crockford. 2002. JSON (Javascript Object Notation). http://www.json.org.



Show in Searchbench
Show Citation Graph

D10-1010: Bunescu, Razvan, Huang, Yunfeng, 2010: *Learning the Relative Usefulness of Questions in Community QA*. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. → P08-1019

Citation of **Huizhong Duan, Yunbo Cao, Chin-Yew Lin, Yong Yu. 2008: *Searching questions by identifying question topic and question focus.* In Proceedings ofACL-08: HLT.** (pdf).

- **Page 10:** "... (Jijkoun and de Rijke, 2005) describe an FAQ question retrieval system in which weighted combinations of similarity functions corresponding to questions, existing answers, FAQ titles and pages are computed using a vector space model. (Zhao et al., 2007) exploit the Encarta logs to automatically extract clusters containing question paraphrases and further train a perceptron to recognize question paraphrases inside each cluster based on a combination of lexical, syntactic and semantic similarity features. More recently, (Bernhard and Gurevych, 2008) evaluated various string similarity measures and vector space based similarity measures on the task of retrieving question paraphrases from the WikiAn-swers repository. The aim of the question search task presented in (Duan et al., 2008) is to return questions that are semantically equivalent or close to the queried question, and is therefore similar to our question ranking task. Their approach is evaluated on a dataset in which questions are categorized either as relevant or irrelevant. Our formulation of question ranking is more general, and in particular subsumes the annotation of binary question categories such as relevant vs. irrelevant, or paraphrases vs. non-paraphrases. Moreover, we are able to exploit the annotated utility relations as supervision in a learning for ranking approach, whereas (Duan et al., 2008) use the annotated dataset to tune the 3 parameters of a mostly unsupervised approach. ...".
- **Page 10:** "... The aim of the question search task presented in (Duan et al., 2008) is to return questions that are semantically equivalent or close to the queried question, and is therefore similar to our question ranking task. Their approach is evaluated on a dataset in which questions are categorized either as relevant or irrelevant. Our formulation of question ranking is more general, and in particular subsumes the annotation of binary question categories such as relevant vs. irrelevant, or paraphrases vs. non-paraphrases. Moreover, we are able to exploit the annotated utility relations as supervision in a learning for ranking approach, whereas (Duan et al., 2008) use the annotated dataset to tune the 3 parameters of a mostly unsupervised approach. The question ranking task was first formulated in (Bunescu and Huang, 2010b), where an initial version of the dataset was also described. In this paper, we introduce 4 versions of the dataset, a more general meaning and structure aware similarity measure, and a supervised model for ranking that substantially outperforms the previously proposed utility measures. Table 4: Pairwise accuracy results. ...".

Show in Searchbench
Show Citation Graph

question retrieval system in which weighted combinations of similarity functions corresponding to questions, existing answers, FAQ titles and pages are computed using a vector space model. (Zhao et al., 2007) exploit the Encarta logs to automatically extract clusters containing question paraphrases and further train a perceptron to recognize question paraphrases inside each cluster based on a combination of lexical, syntactic and semantic similarity features. More recently, (Bernhard and Gurevych, 2008) evaluated various string similarity measures and vector space based similarity measures on the task of retrieving question paraphrases from the WikiAnswers repository. The aim of the question search task presented in (Duan et al., 2008) is to return questions that are semantically equivalent or close to the queried question, and is therefore similar to our question ranking task. Their approach is evaluated on a dataset in which questions are categorized either as relevant or irrelevant. Our formulation of question ranking is more general, and in particular subsumes the annotation of binary question categories such as relevant vs. irrelevant, or paraphrases vs. non-paraphrases. Moreover, we are able to exploit the annotated utility relations as supervision in a learning for ranking approach, whereas (Duan et al., 2008) use the annotated dataset to tune the 3 parameters of a mostly unsupervised approach. The question ranking task was first formulated in (Bunescu and Huang, 2010b), where an initial version of the dataset was also described. In this paper, we introduce 4 versions of the dataset, a more general meaning and structure aware similarity measure, and a supervised model for ranking that substantially outperforms the previously proposed utility measures.

8 Conclusion

We presented a supervised learning approach to the question ranking task in which previously known questions are ordered based on their relative utility with respect to a new, reference question. We created four versions of a dataset of 60 groups of questions [5], each annotated with a partial order relation reflecting the relative utility of questions inside each group. An SVM ranking model was trained on the dataset and evaluated together with a set of simpler, unsupervised question-to-question similarity models. Experimental results demonstrate the importance of using structure and meaning aware features when computing the relative usefulness of questions.

[5] The dataset will be made publicly available.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments.

References

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval.* ACM Press, New York.

Delphine Bernhard and Iryna Gurevych. 2008. Answering learners' questions by retrieving question paraphrases from social Q&A sites. In *EANL '08: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 44–52, Morristown, NJ, USA. Association for Computational Linguistics.

Razvan Bunescu and Yunfeng Huang. 2010a. Towards a general model of answer typing: Question focus identification. In *Proceedings of The 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2010), RCS Volume*, pages 231–242.

Razvan Bunescu and Yunfeng Huang. 2010b. A utility-driven approach to question ranking in social QA. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 125–133.

Michael Collins. 1999. *Head-driven Statistical Models for Natural Language Parsing.* Ph.D. thesis, University of Pennsylvania.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, pages 9–16.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting assively parallel news sources. In *Proceedings of The 20th International Conference on Computational Linguistics (COLING'04)*, page 350.

Huizhong Duan, Yunbo Cao, Chin-Yew Lin, and Yong Yu. 2008. Searching questions by identifying question topic and question focus. In *Proceedings of ACL-08: HLT*, pages 156–164, Columbus, Ohio, June.

Ulf Hermjakob, Abdessamad Echihabi, and Daniel Marcu. 2002. Natural language based reformulation

106

Figure 3: Citation sentences in context view

Cailing Dong and Ulrich Schäfer. 2011. Ensemble-style self-training on citation classification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 623–631, Chiang Mai, Thailand.

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.

Nigel Harwood. 2009. An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics*, 41(3):497–518.

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

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 and Uwe Kasterka. 2010. Scientific authoring support: A tool to navigate in typed citation graphs. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Linguistics and Writing*, pages 7–14, Los Angeles, CA.

Ulrich Schäfer, Bernd Kiefer, Christian Spurk, Jörg Steffen, and Rui Wang. 2011. The ACL Anthology Searchbench. In *Proceedings of ACL-HLT 2011, System Demonstrations*, pages 7–13, Portland, Oregon, June.

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.

Stephen Wan, Cécile Paris, Michael Muthukrishna, and Robert Dale. 2009. Designing a citation-sensitive research tool: an initial study of browsing-specific information needs. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09, pages 45–53, Singapore.

Appendix: Sample Citation Context XML Fragment

```
<citation valid="true">
  <aclId>P03-1054</aclId>
  <authors>
    <author>D Klein</author>
    <author>C D Manning</author>
  </authors>
  <location>Morristown, NJ, USA</location>
  <verbs undef="combines"/>
  <institution>for Computational Linguistics</institution>
  <title>Accurate unlexicalized parsing</title>
  <rawString>D. Klein and C. D. Manning. 2003. Accurate unlexicalized parsing.
             In ACL ’03: Proceedings of the 41st Annual Meeting on Association
             for Computational Linguistics, pages 423430, Morristown, NJ, USA.
             Association for Computational Linguistics.</rawString>
  <marker>Klein, Manning, 2003</marker>
  <date>2003</date>
  <publisher>Association</publisher>
  <pages>423--430</pages>
  <booktitle>In ACL ’03: Proceedings of the 41st Annual Meeting on Association
             for Computational Linguistics</booktitle>
  <contexts>
    <context position="10486">
      <verbs undef="combines"/>
      <text> on a testsuite created using that tool. 3.1. The Afazio RTE system
            Similarly to the Nutcracker system (Curran et al., 2007), the Afazio
            RTE system combines a statistical parser (the Stanford parser (Klein
            and Manning, 2003)) with a symbolic semantic component. First, a
            system of cascaded rewrite modules is used to rewrite the output of
            the parser into a &amp;quot;normalised&amp;quot; semantic
            representation intended to abstract away from</text>
      <sentences center-no="78" center-page="3">
        <sentence no="77" page="3">The Afazio RTE system</sentence>
        <sentence no="78" page="3">Similarly to the Nutcracker system (Curran et
            al., 2007), the Afazio RTE system combines a statistical parser (the
            Stanford parser (Klein and Manning, 2003)) with a symbolic semantic
            component.</sentence>
        <sentence no="80" page="3">Special emphasis is placed on capturing
            syntax based equivalences such as syntactic (e.g., active/passive)
            variations, redistributions and noun/verb variants.</sentence>
        <sentence no="79" page="3">First, a system of cascaded rewrite modules
            is used to rewrite the output of the parser into a "normalised"
            semantic representation intended to abstract away from surface
            differences and assign paraphrases the same representation
            (Bedaride and Gardent, 2009a; Bedaride and Gardent, 2009b).</sentence>
        <sentence no="81" page="3">Next, automated reasoning is used to check
            entailment.</sentence>
        <sentence no="75" page="3">To illustrate this, we evaluate the Afazio
            RTE system on a testsuite created using that tool.</sentence>
        <sentence no="76" page="3">3.1.</sentence>
      </sentences>
    </context>
  </contexts>
</citation>
```

Figure 4: A citation context from paper L10-1259, citing paper P03-1054, as generated by ParsCit, extended by Citation Browser and Searchbench code with ACL ID, context information and context sentences.