Science Information Applications

Ulrich Schäfer

DFKI Language Technology Lab

Paper/bibliographic search

http://academic.research.microsoft.com/

- for many research areas; graphical browsers (Windows only...)
- "explore 37,472,555 publications and 19,327,188 authors" (as of yesterday):
- people, organization, citation network, CfP calendar, research trends

http://scholar.google.com

- textual paper content search, author search

DBLP (http://www.informatik.uni-trier.de/~ley/db/): 1.8 million entries, mainly computer science and related field; only bibl. metadata with links to open or closed access papers

Bielefeld Academic Search (http://www.base-search.net/): 32,663,572 papers from 2085 sources: metadata with links to open or closed access papers

CiteceerX (http://citeseerx.ist.psu.edu/index): digital library, search engine and citation statistics for computer and information science papers, also a software infrastructure

Open Access Portals:

Scientific Commons (http://en.scientificcommons.org): 38,245,864 documents from 1269 sources

ArXiv (http://lanl.arxiv.org): Open access to 728,365 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics

Publisher's Portals

Springer

Elsevier

Thomson-Reuters Web of Science

Universities, e.g. SciDok (SULB)

Thousands of other indexes and portals...

Citation Analysis

Pioneer: Eugene Garfield (1955), see references

founder of ISI (Information Sciences Institute, USC, Marina del Rey, CA)

Related Research fields:

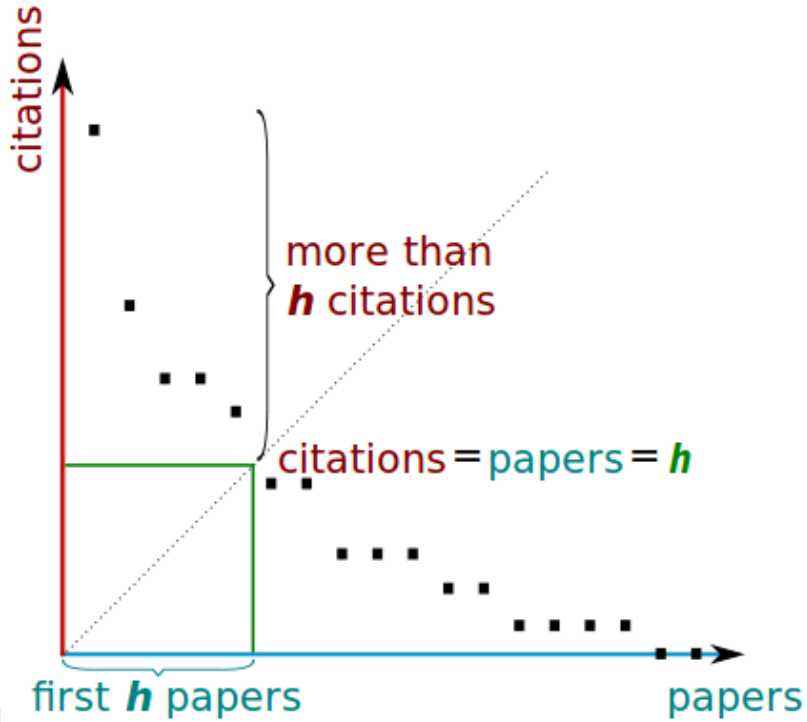
- Scientometrics
- Bibliometrics
- Library Science
- Information Science

Citation Analysis

Citation Index

h-index (or Hirsch index, after Jorge E. Hirsch)

A scientist has index h if h of his/her N papers have at least h citations each, and the other (N – h) papers have no more than h citations each.



The key to (almost) everything in citation analysis and search: String distance metrics...

1. Levenshtein distance: number of edits from s_1 to s_2
2. Jaro distance: $d_j = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right)$

(i.e., normalized metric: 0=no, 1=full match; m=# of matches, t=1/2 # of transpositions)

3. Jaro-Winkler: Jaro with weight for prefix changes

There are many more...

→ Exercise python + external Levenshtein module (src from http://pypi.python.org/pypi/python-Levenshtein/)

Exercise: python-levenshtein library

```
Ubuntu/Debian:
sudo apt-get install python-levenshtein
python
from Levenshtein import distance, hamming, jaro, jaro_winkler
>>> distance("scientometrics", "bibliometrics")
5
>>> hamming("bibliometrics", "scientometric")
13
>>> jaro("scientometrics", "bibliometrics")
0.6672771672771672
>>> jaro_winkler("scientometrics", "bibliometrics")
0.6672771672771672
>>> jaro("scientometrics", "scientomanics")
0.8772893772893773
>>> jaro_winkler("scientometrics", "scientomanics")
0.9754578754578754
```

Java variant (different library): Simmetrics

http://sourceforge.net/projects/simmetrics/

http://web.archive.org/web/20081224234350/
http://www.dcs.shef.ac.uk/~sam/stringmetrics.html

The case of Medical Science

Elaborated Ontologies:

- MeSH (Medical Subject Headlines, http://www.nlm.nih.gov/mesh/)
- UMLS (Unified Medical Language System, http://www.nlm.nih.gov/research/umls/)

Huge text databases: PubMed/Medline (publication metadata and abstracts only...): http://www.ncbi.nlm.nih.gov/pubmed/

There are many more...

Related research field: Literature analysis/text mining as subfield of Bioinformatics

Computational Linguistics

LT World (http://www.lt-world.org)

- Underlying ontology and data: people, organisations, projects, conferences, news, links, resources, tools, etc.
- Largely hand-crafted content, limited terminology resources, no publication metadata nor publication content

ACL Anthology (http://www.aclweb.org/anthology)

- Open access digital library of more than 23,000 CL papers from 1967 until today, including the complete CL Journal.
- Content search via Google custom search and DFKI's Searchbench
- Incomplete publication metadata (will be improved)
- Citation Network: http://clair.si.umich.edu/clair/anthology/

Using more NLP for Science Information Application

Motivation: go beyond citation graphs and indexes, text retrieval/fulltext and metadata search

Users want to see original, full content of papers, not just bibliographic metadata, abstracts and references

Interesting areas for NLP:

- improve search → semantic search ("find what I mean")
    - search for complex propositions, synonyms, in context
    - preprocess textual content: parsing, coreferences, etc.
- automatic terminology, taxonomy & ontology extraction from text
- qualitative citation analysis
- automatic summarization
- question answering, learning by reading, expert systems, …

Parsing Science with NLP (more or less...)

MEDIE is a semantic search engine to retrieve biomedical correlations from MEDLINE articles (Sætre et al., 2008)

SciBorg: UK-based research project on parsing and named entity recognition of chemistry papers from a publisher

Wolfram Alpha: Question answering, specialized tools and database: http://www.wolframalpha.com/

NLP Pipeline and before

Preprocessing 1: Text extraction from digital and scanned documents

commercial (O)CR:

- Omnipage, Abbyy

Open source (O)CR:

- Tesseract (http://code.google.com/p/tesseract-ocr/)

Open source layout recognition on top of Tesseract:

- Ocropus (http://code.google.com/p/ocropus/)

Alternatives for native (not scanned) PDF:

- Apache PDFbox: http://pdfbox.apache.org/
- Poppler/Xpdf: http://poppler.freedesktop.org/

Text and metadata extraction from office file formats etc.:

- Apache POI (http://projects.apache.org/projects/poi.html),
- Aperture (http://aperture.sourceforge.net/)

NLP Pipeline

Preprocessing 2:

- text filtering (remove non-text character sequences)
- de-hyphenation
- XML Markup (optional, e.g. TEI P5, Docbook,...), containing information on section headings, footnotes, tables, character styles such as *Italics,* page numbers, figures and tables, captions, ...
  Potentially useful for detecting argumentative zones, citation classification, emphasized tokens marked for parsing, etc.

- Example: XML file: paper.xml

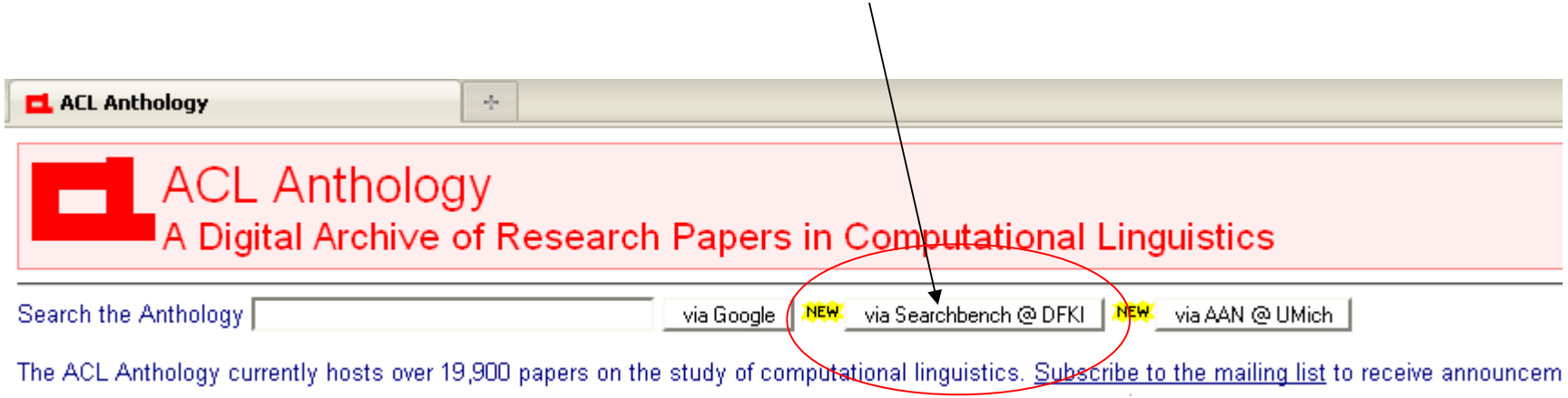
NLP Pipeline

Preprocessing 3:

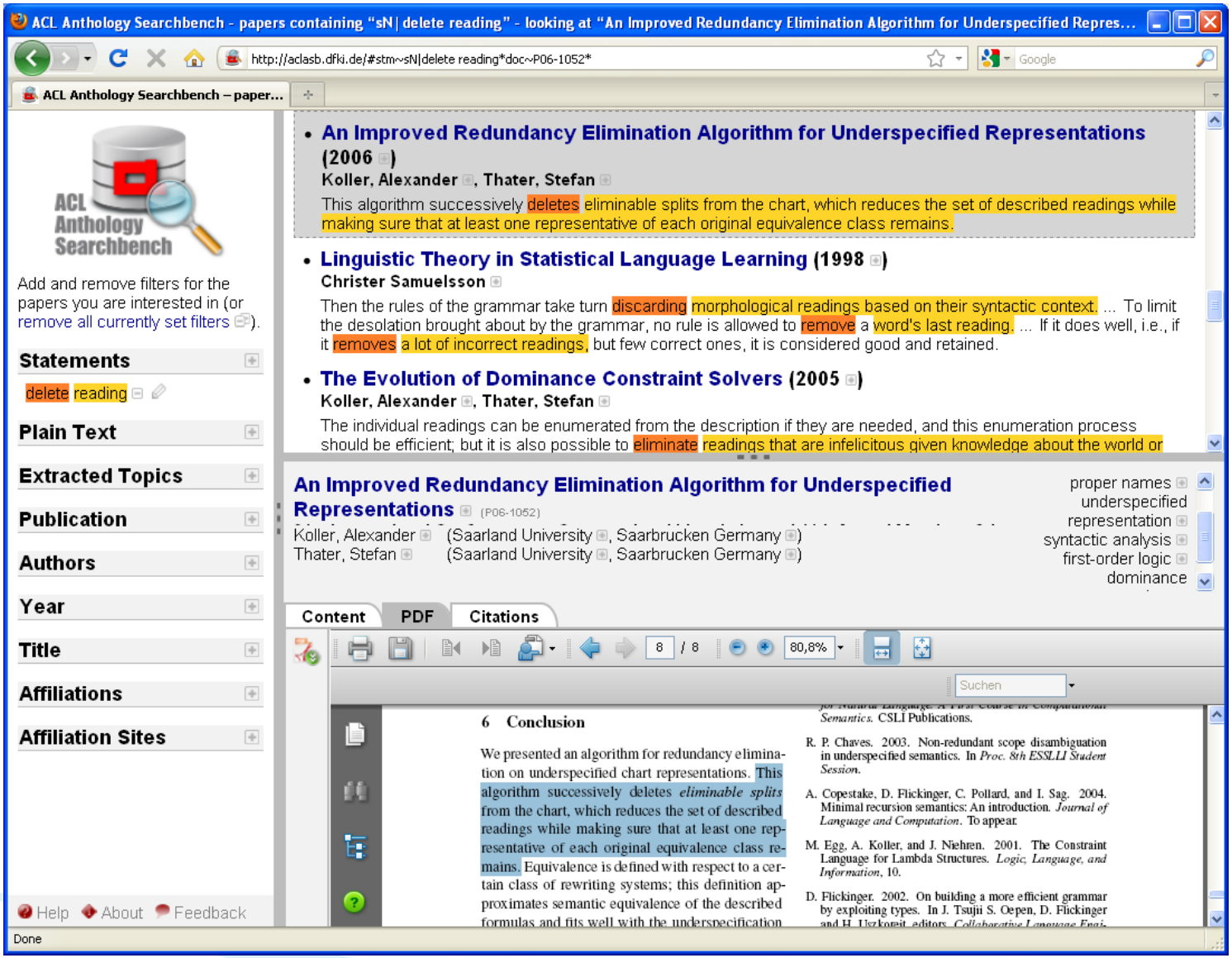
- Sentence boundary recognition
- Tokenization
- PoS tagging (for unknown word guessing, term extraction, ...)
- Named entity recognition
- Parsing
- Semantics extraction
- Index preparation
- (Structured) indexing with Apache Lucene/Solr

ACL Anthology Searchbench

- http://aclasb.dfki.de
- Released at ACL-2011
- Combines semantic, full-text and bibliographic search in 19,000 papers of the ACL Anthology from the past 46 years, incl. CL journal
- ACL Anthology start page links to it!



ACL Anthology Searchbench



Research Fields in TAKE

Coreference resolution

Deep parsing and semantic tuple extraction

Combined semantic search

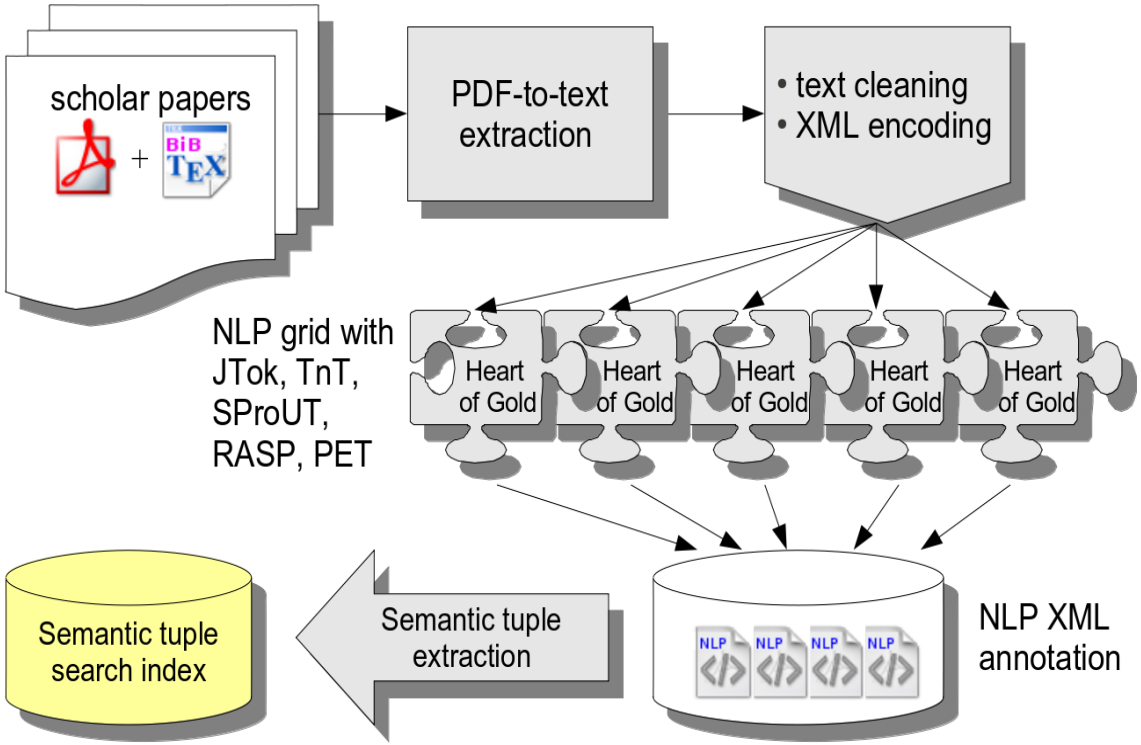
Citation Analysis

Unsupervised multi-word domain term extraction

Taxonomy extraction

Glossary extraction

Paper Parsing Architectue



Common NL Pre-Processing

Boost in Deep Parsing Coverage and Efficiency

ACL Anthology Parsing: breakthrough by combining

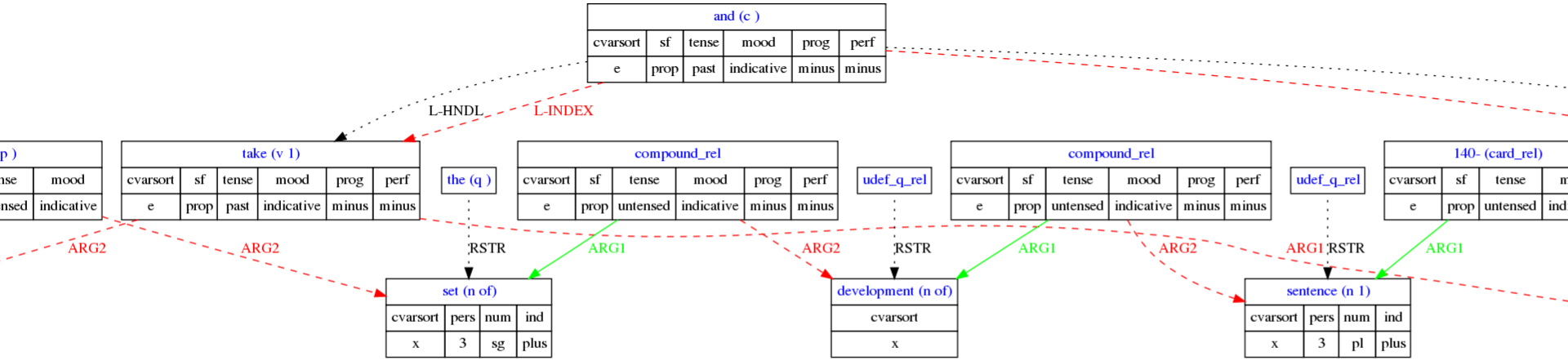
- **chart pruning**: directed search during parsing to increase performance, and also coverage for longer sentences (Cramer & Zhang, 2010)
- **chart mapping**, a novel method for integrating preprocessing information (Adolphs et al, 2008)
- new grammar (ERG) with better handling of open word classes
- fine-grained named entity recognition, including citation patterns (SProUT)
- new parse ranking model (WeScience; Oepen '09)

→ Improvement of overall coverage from 63% to now **>83%** full parses (now 4.3 million sentences)

DMRS to Semantic Tuple Conversion

From W07-1209, section 3

"We took the raw strings from the 140-sentence development set and parsed them with each of the state-of-the-art probabilistic parsers."



Semantic Subject	Semantic Predicate	Semantic First Object	Semantic Second Object	Adjuncts
We	took	the raw strings from the 140-sentence development set		
We	parsed	them		with each of the state-of-the-art probabilistic parsers.

Asking Solr Index (simplified)

Query:
"method improve baseline"

is translated into Apache Solr query:
subj:method +pred:(improve OR ameliorate OR better OR meliorate) +(dobj:baseline OR iobj:baseline OR rest:baseline)

result (1 of 72) →

could also be used for question answering...

```
<doc>  <!-- each doc is a single quriple sentence here -->
  <float name="score">1.2502118</float>
  <date name="timestamp">2009-01-27T10:46:38.452Z
  </date>
  <str name="aclaid">W05-0814</str>
  <int name="offset">198</int>
  <int name="sentno">87</int>
  <int name="page">4</int>
  <str name="prefix">W05-0814-s87-p4</str>
  <str name="qgen">PET</str>
  <str name="sentence">Our model and training method
      improve upon a strong baseline for producing 1-to-
      many alignments.
  </str>
  <str name="subj">Our model training method</str>
  <int name="subj_start">0</int>
  <int name="subj_end">28</int>
  <str name="pred">improve</str>
  <int name="pred_start">30</int>
  <int name="pred_end">36</int>
  <str name="rest">upon a strong baseline for producing
      1-to-many alignments
  </str>
  <int name="rest_start">38</int>
  <int name="rest_end">94</int>
</doc>
```

Searchbench: Statement Search Options

strict

only find strictly affirmative statements with a predicate matching only the entered one.

default

find generally affirmative or neutral statements with a predicate matching either the entered one or a synonym of it.

lax

as before, but additionally find statements with negated or neutral predicates matching antonyms of the entered predicate.

maximal

find statements with the entered predicate or a synonym/antonym thereof, irrespective of whether the predicate is negated or not

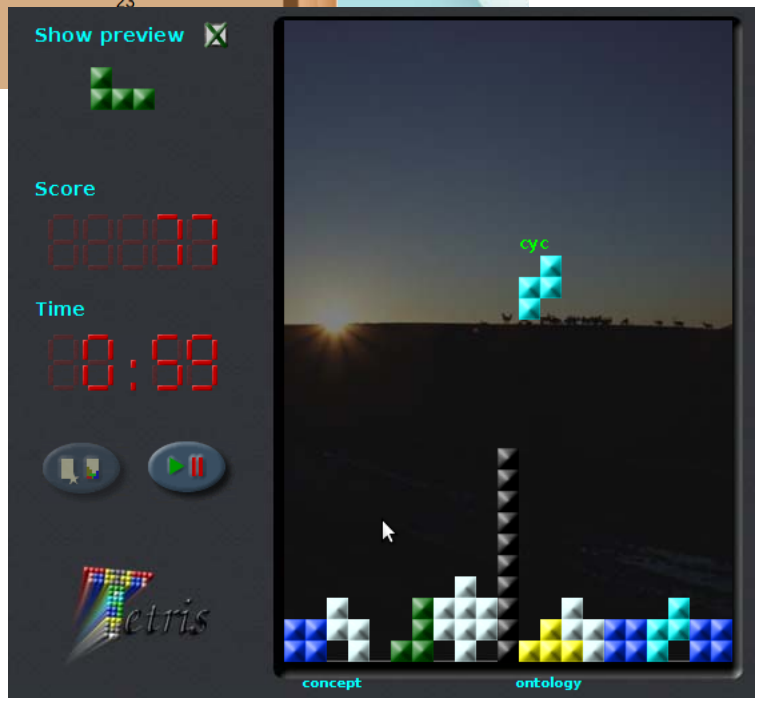
Multiword Domain Term Extraction

Based on an extended implementation of the Frantzi & Ananiadou 2000 approach (C-Value/NC-Value)

Example in Searchbench: „data structure + speech recognition + partial results + ...

Also basis for taxonomy and glossary extraction

Automatic Taxonomy Extraction – Evaluation with OntoGWAP



Examples of extracted hypernym-hyponym pairs (including invalid pairs)

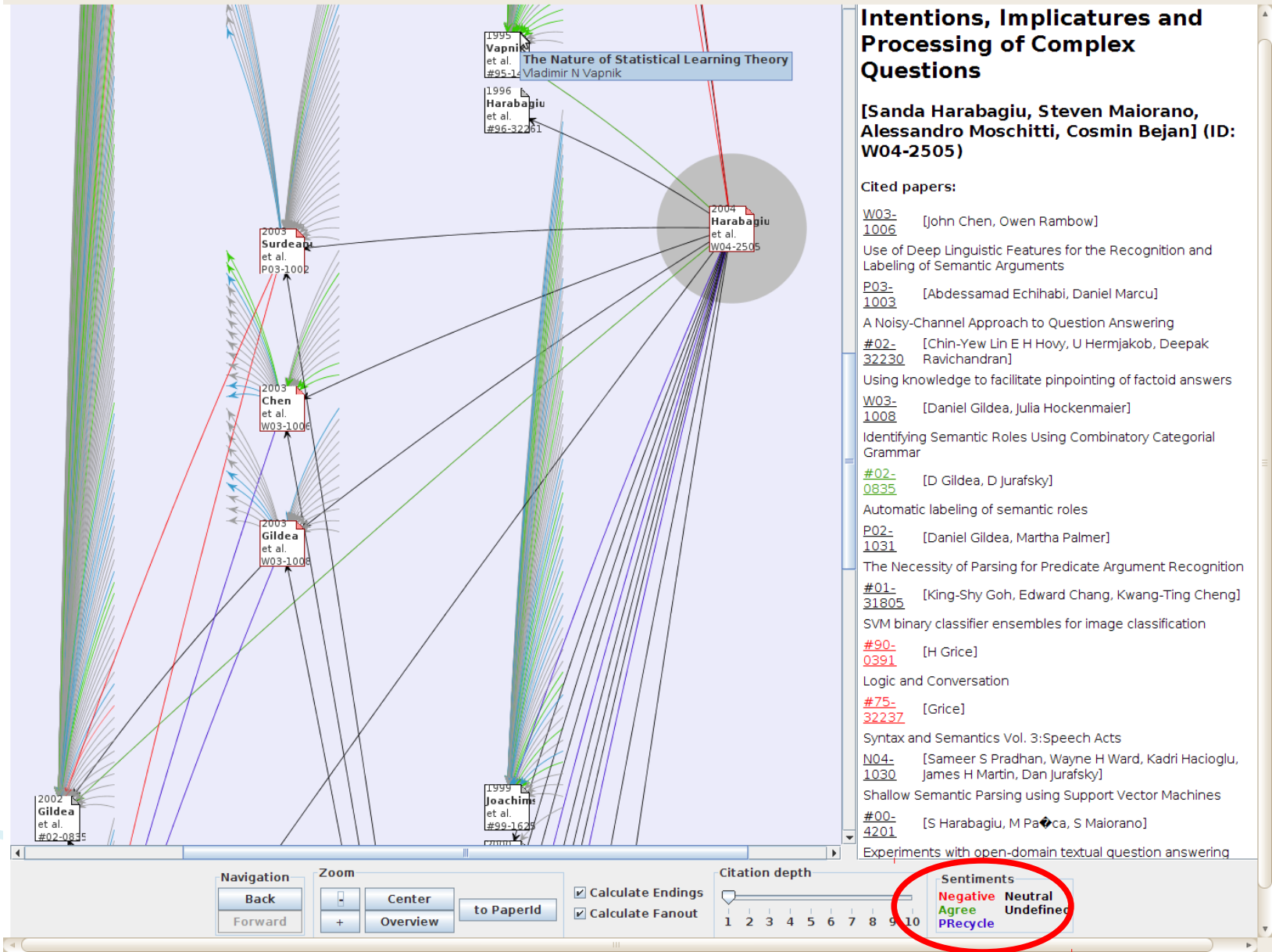
Hypernym	Hyponyms
natural language processing application	information extraction, question answering, machine translation, information retrieval, document summarization, speech recognition, pos tagging, named entity recognition, question answering system, open-domain question-answering, text mining, named entity extraction, question-answering, automatic lexical acquisition, text summarization, document clustering, language model building, word sense disambiguation, annotation projection, cross language information retrieval, . . .
agglutinative language	korean, basque, chinese, hungarian, japanese, thai
web search engine	google, yahoo, altavista
classifier	svm, decision tree, support vector machine, naive bayes, conditional random field, maximum entropy classifier, dependency path, probabilistic classifier, pruned decision tree, timbl, k-nn, acoustic confidence score
vector distance measure	euclidean distance, cosine
dependency relation	subj, subject, object, arg, obj, head-modifier
open-class word	adjective, adverb, verb, common noun, proper name
morphological feature	number, gender, person, case, aspect, pos, tense, count, voice
sequence labeling task	named entity recognition, pos tagging, chunking, syntactic chunking
evaluation metric	nist, bleu

Hyper-/Hyponym Extraction: Evaluation

The competition for prizes lasted 10 days. 61 players participated. 32 Tetris players, 10 Invaders players, 26 quiz participants. Only one played all games.

	Category	Value
Data statistics	No. presented pairs	2940
	% of entire set	31%
	No. annotations	6782
Precision results	No. 3-way agreements	639
	of which, no. valid *is-a* pairs	490
	no. invalid pairs	149
	3-way precision	**77%**
	No. 5-way agreements	298
	of which, no. valid *is-a* pairs	239
	no. invalid pairs	59
	5-way precision	**80%**

Citation Classification & Navigation



Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Typed (Qualified) Citation Classification

Classify citation sentences into categories such as use, refutation, neutral, confirmative, ...

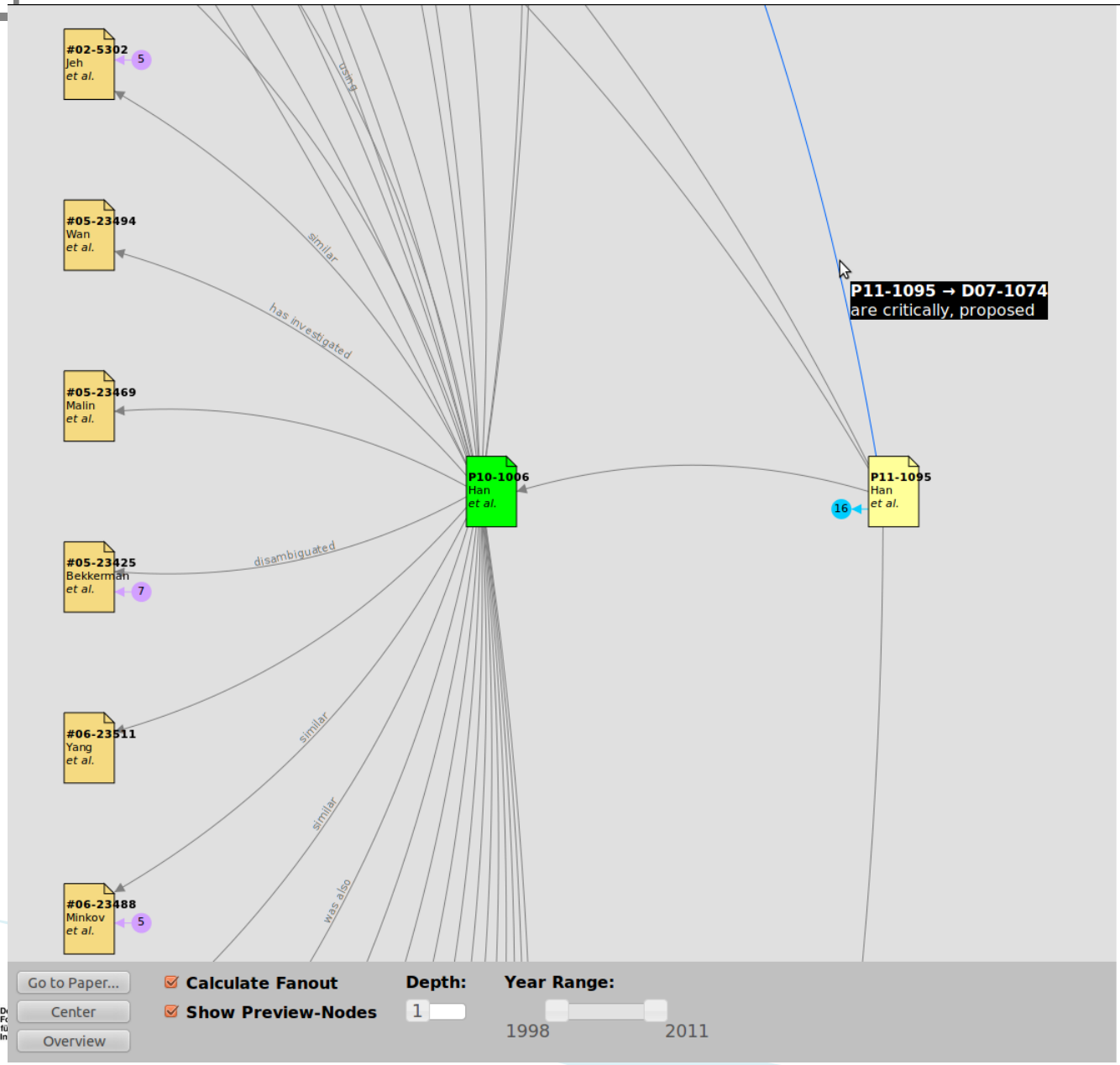
Possibly several categorized citations contribute to an overall classification of the reference from one paper to another (colored edge in the graphical user interface)

Rule-based approaches with PoS-, lexical, syntactical patterns:
not robust, low overall recall and precision

→ Novel approach with semi-supervised learning on citation classification addresses two problems:

- expensive manual annotation
- unbalanced class distribution

New Citation Browser for ACL Searchbench



Structural Semantic Relatedness: A Knowledge-Based Method to Named Entity Disambiguation

Xianpei Han, Jun Zhao

Cited Papers:

#98-0614:
Christiane Fellbaum (1998)
WordNet: An Electronic Lexical Database

#98-0143:
Dekang Lin (1998)
An information-theoretic definition of similarity

P98-1012:
Amit Bagga, Breck Baldwin (1998)
Entity-Based Cross-Document Core f erencing Using the Vector Space Model

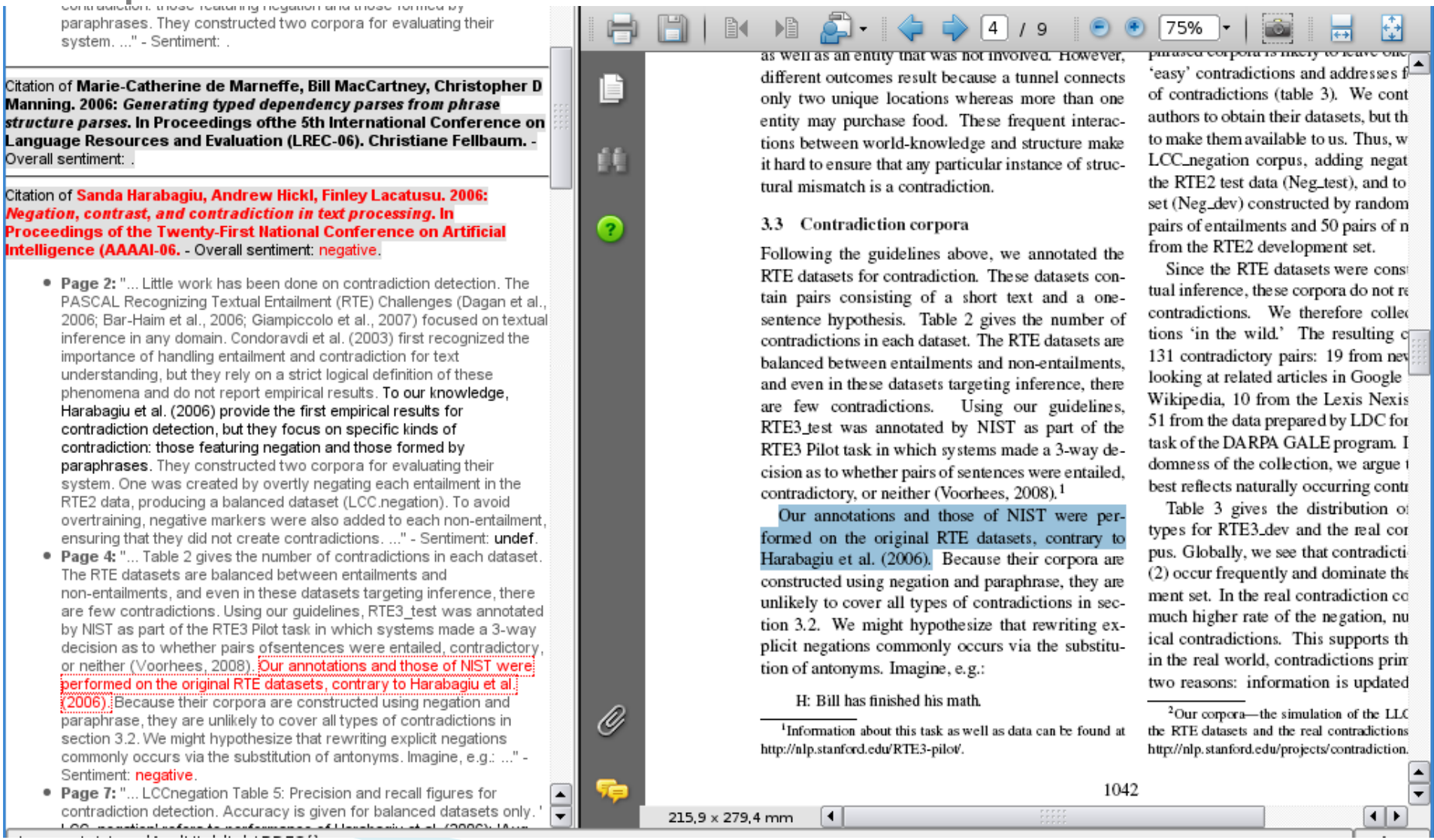
#99-1464:
Ricardo A Baeza-Yates, Berthier Ribeiro-Neto (1999)
Modern Information Retrieval

#02-5302:
Glen Jeh, Jennifer Widom (2002)
Simrank: A measure of structural-context similarity

W03-0405:
Gideon Mann, David Yarowsky (2003)
Unsupervised Personal Name Disambiguation

P04-1076:
Cheng Niu, Wei Li, Rohini K Srihari (2004)
Weakly Supervised Learning for Cross-document Person Name Disambiguation Supported by Information Extraction

View Citations Sentences in Context



Exercise 2: Searchbench

Formulate Searchbench queries from questions

Literature

Lutz Bornmann and Hans-Dieter Daniel. 2008. What do citation counts measure? A review of studies on 13 citing behavior. Journal of Documentation, 64(1):45–80. DOI 10.1108/00220410810844150.

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.

K. Frantzi, S. Ananiadou, and H. Mima. 2000. Automatic recognition of multi-word terms: the Cvalue/NC-value method. International Journal on Digital Libraries, 3:115–130.

M. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In Proc. of the 14th Coling Conference, pages 539–545.

Z. Kozareva, E. Riloff, and E. Hovy. 2008. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. In Proc. of ACL, pages 1048–1056.

Ann Copestake and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In Proceedings of the 2nd Conference on Language Resources and Evaluation (LREC-2000), pages 591–598, Athens, Greece.

Ann Copestake, Dan Flickinger, Ivan A. Sag, and Carl Pollard. 2005. Minimal recursion semantics: an introduction. Journal of Research on Language and Computation, 3(2–3):281–332.

CJ Rupp, Ann Copestake, Peter Corbett, and Ben Waldron. 2007. Integrating general-purpose and domain-specific components in the analysis of scientific text. In Proceedings of the UK e-Science Programme All Hands Meeting 2007 (AHM2007), Nottingham, UK.

Rune Sætre, Sagae Kenji, and Jun'ichi Tsujii. 2008. Syntactic features for protein-protein interaction extraction. In Christopher J.O. Baker and Su Jian, editors, Short Paper Proceedings of the 2nd International Symposium on Languages in Biology and Medicine (LBM 2007), pages 6.1–6.14, Singapore, 1. ISSN 1613-0073319.

Literature

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.

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

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.

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.

Ulrich Schäfer, Bernd Kiefer, Christian Spurk, Jörg Steffen, Rui Wang: The ACL Anthology Searchbench. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011), System Demonstrations, pages 7-13, 2011. ISBN 978-1-932432-90-9. Portland, OR, USA.

Magdalena Wolska, Ulrich Schäfer, The Nghia Pham: Bootstrapping a Domain-specific Terminological Taxonomy from Scientific Text. 9th International Conference on Terminology and Artificial Intelligence (TIA), pages 17-23, Paris, France, 2011.

Literature

Adolphs, P., Oepen, S., Callmeier, U., Crysmann, B., Flickinger, D., Kiefer, B.: Some fine points of hybrid natural language parsing. In: Proc. of LREC. pp. 1380-1387. Marrakesh, Morocco (2008).

Callmeier, U.: PET – A platform for experimentation with ecient HPSG processing techniques. Natural Language Engineering 6(1), 99-108 (2000).

Cramer, B., Zhang, Y.: Constraining robust constructions for broad-coverage parsing with precision grammars. In: Proc. of COLING. pp. 223-231. Beijing, China (2010).

Flickinger, D., Oepen, S., Ytrestøl, G.: WikiWoods: Syntacto-semantic annotation for English Wikipedia. In: Proc. of LREC. pp. 1665-1671. Valletta, Malta (2010).