

# NLP for Science Information Systems

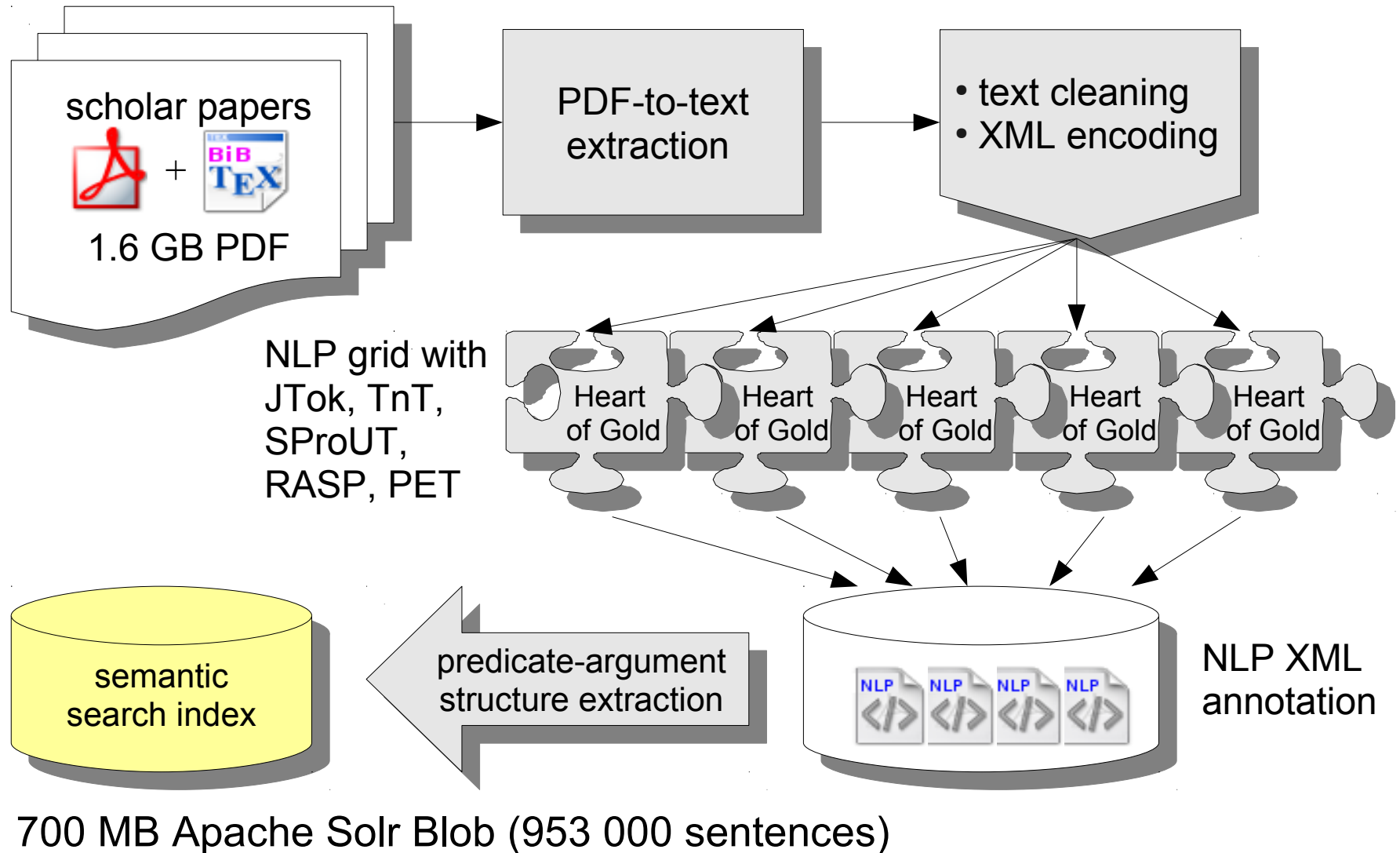
Ulrich Schäfer

DFKI Language Technology Lab





- ☆ extend and improve Scientist's Workbench by using more NLP
  - improve **semantic search** in paper contents
  - new application: **typed citation analysis (+ demo)**
  
- ☆ research focus
  - multi-word **term extraction** as preparation step
  - automatic **ontology learning** and extension from text
  - **coreference resolution** as auxiliary technology to increase redundancy and robustness
  
- ☆ corpus: ACL Anthology (initially 6300 CL/LT conference and workshop papers, now extended to 8500, including CL Journal 2002-2009)





## Semantic Search in 6300 ACL Anthology papers

HyLap :: Web-based demonstrator - Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

**Document**

the problem of finding hypernyms for relevant concepts in a specific domain (e.g. Optical Recording) in the **context** of a concrete and challenging application scenario (patent processing). To this end information available on the Web is **exploited**. The extraction method includes four main steps. Firstly, the **Google search engine** is exploited to retrieve possible instances of isa-patterns reported in the literature. Then, the returned snippets are filtered on the basis of lexico-**syntactic** criteria (e.g. the candidate hypernym must be expressed as a noun **phrase** without complex modifiers). In a further filtering step, only candidate hypernyms compatible with the **target domain** are kept. Finally a candidate ranking mechanism is applied to select one hypernym as output of the **algorithm**. The extraction method was evaluated on 100 concepts of the Optical Recording domain. Moreover, the reliability of isa-patterns reported in the literature as predictors of isa-relations was assessed by manually **evaluating** the template instances remaining after lexico-**syntactic** filtering, for 3 concepts of the same domain. While more extensive testing is needed the method appears promising especially for its portability across different domains. Related **Work** Many works have considered the problem of automatically building or extending an **ontology** starting from **Hearst** **Hearst**, 1992; **Hearst**, 1998) who first proposed to gather from texts the **syntactic** patterns specific to a given relation. A similar approach is employed in (Girju et al., 2006) for the part-of relation. Other authors propose different approaches. For example, in (K. Shinzato, 2005) the **HTML** Tags of itemization are employed. (**Snow** et al., 2005) use **Minipar** to save and generalize the contexts (**dependency**-paths) where an isa-relation occurs. With this method the authors can compare their results with those obtained using a subset of the patterns proposed by **Hearst**. Finally, the authors in (Sombatsrisomboon R. and M., 2003) propose to extract the

**Text Input**

**Search Quiriles**

Subject

☐ Allow predicate synonyms

☐ Abstracts only

**Ontology Browser**

**Search Results**

**24 Quiriles:**

- L08-1180: (Snow et al., 2005) use Minipar to save and generalize the contexts (dependency-paths) where an isa-relation occurs.
- W07-0211: The goals of this project are to provide an accurate and fast system, which we have called DLSITE-2, that can be applied in software systems that require a near-real-time interaction with the user.
- N07-1065: First, the named entity recognizer of Minipar is used to identify all numerical entities in text, labeled as NUM.
- W06-0805: 1. Use MINIPAR (Lin, 1998) to generate dependency parses of texts.
- I08-1026: We use a popular dependency parser, Minipar, to extract the syntactic dependency between words.
- W06-0508: We use Minipar (Lin, 1993), which produces functional relations for the components in a sentence, including subject and object relations with respect to a verb.
- I08-1026: Syntactic relationships, derived from a dependency parser, Minipar, are used as linguistic term dependencies.
- L08-1318: For this purpose, we use the MiniPar dependency parser (Lin, 1998).
- N07-1071: We used the Mini-par parser (Lin 1993) to match DIRT patterns in the text.



## Semantic Search in 6300 ACL Anthology papers

HyLap :: Web-based demonstrator - Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

Document

Ontology Browser

103%

4 / 4

Text Input

Submit

Quantico de...

Search Quirples

Subject

☐ Allow predicate

☐ Abstracts only

6 Conclusion

- Considering multiple stemming possibilities for each word seems important.
- Alternating between increasing likelihood and decreasing error rate is a useful training approach which can be used for many problems.
- Our model and training method improve upon a strong baseline for producing 1-to-many alignments.
- Our model and training method can be used with the "intersection" heuristic to produce higher quality 1-to-1 alignments
- Models which can directly model many-to-many alignments and do not require heuristic symmetrization are needed to produce higher quality many-to-many alignments. Our training method can be used to train them.

7 Acknowledgments

This work was supported by DARPA-ITO grant NN66001-00-1-9814 and NSF grant IIS-0326276.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Mauro Cettolo and Marcello Federico. 2004. Minimum error training of log-linear translation models. In *Proc. of the International Workshop on Spoken Language Translation*, pages 103–106, Kyoto, Japan.

Herve Dejean, Eric Gaussier, Cyril Goutte, and Kenji Yamada. 2003. Reducing parameter space for word alignment. In *HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta, July.

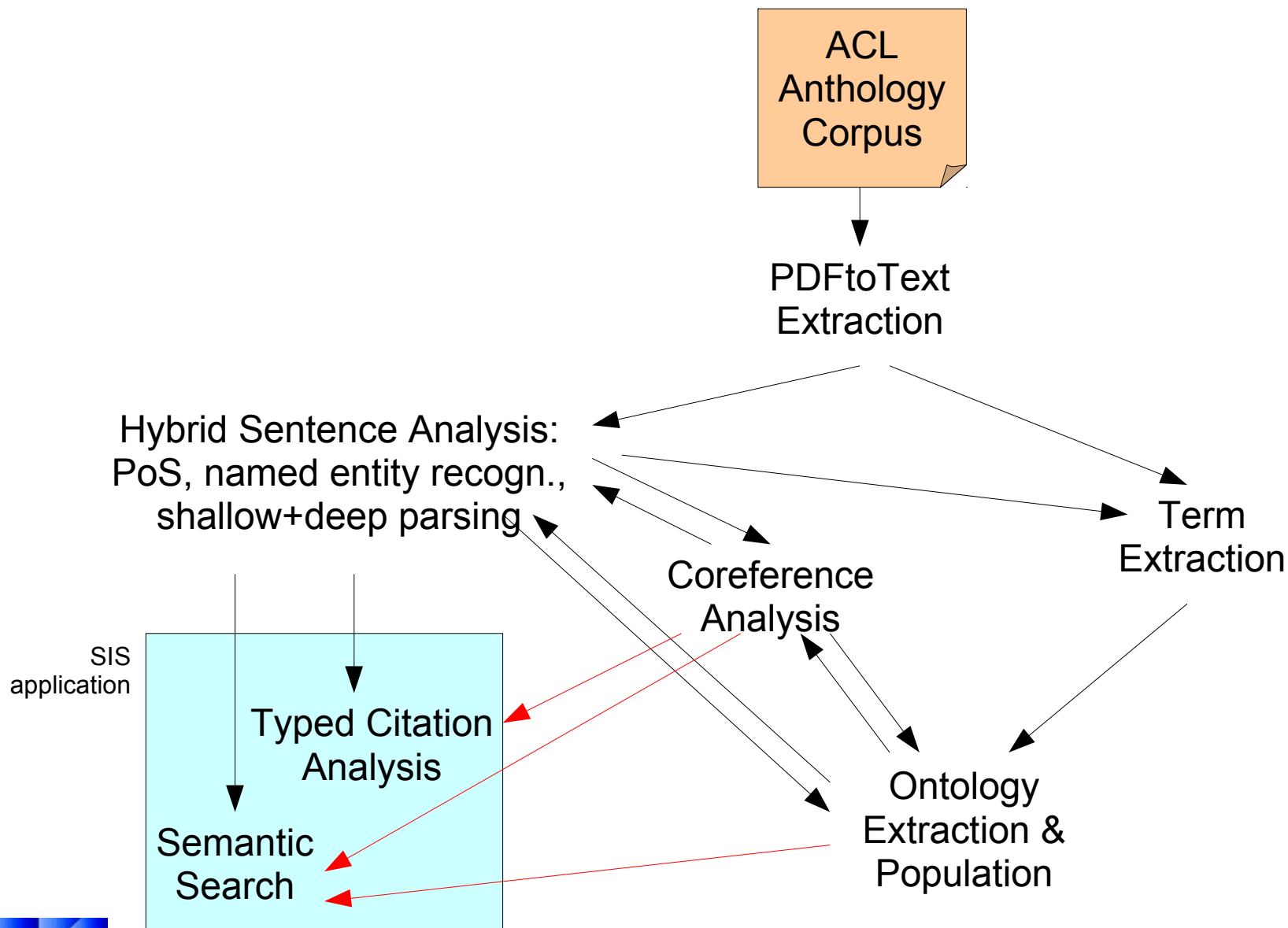
Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

M. F. Porter. 1997. An algorithm for suffix stripping. In *Readings in information retrieval*, pages 313–316, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Monica Rogati, Scott McCarley, and Yiming Yang. 2003. Unsupervised learning of arabic stemming using a parallel corpus. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, July.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, Denmark, August.





Example: the following two contiguous paragraphs from N04-1022 contain four definitions

"Word Error Rate (WER) is the ratio of the string-edit distance between the reference and the hypothesis word strings to the number of words in the reference. String-edit distance is measured as the minimum number of edit operations needed to transform a word string to the other word string.

Position-independent Word Error Rate (PER) measures the minimum number of edit operations needed to transform a word string to any permutation of the other word string. The PER score (Och, 2002) is then computed as a ratio of this distance to the number of words in the reference word string."

→ convert to RDF using hybrid NLP



- ☆ We adapted the method from Frantzi, Ananiadou & Mima (2000) to our corpus.  
It uses frequencies of 2-,3-,4-grams with special emphasis on contained n-grams
- ☆ Very good results when based on the recently extracted texts
- ☆ The quality of the raw text extracted from PDF matters, e.g. bad results from Omnipage (hyphenation problems etc.)
- ☆ Future extensions: shallow parsing, named entities
- ☆ Evaluation against external source: Jurafsky & Martin 2007: Speech and Language Processing Introduction, electronic draft
  - multiwords from border texts
  - same term extraction method on full book text

## SYNTACTIC DISAMBIGUATION

tences is an extremely irksome problem that affects all parsers. Ultimately, most natural language processing systems need to be able to choose the correct parse from the multitude of possible parses via process known as **syntactic disambiguation**. Unfortunately, effective disambiguation algorithms generally require statistical, semantic, and pragmatic knowledge not readily available during syntactic processing (techniques for making use of such knowledge will be introduced later, in Ch. 14 and Ch. 18).

Lacking such knowledge we are left with the choice of simply returning all the possible parse trees for a given input. Unfortunately, generating all the possible parses from robust, highly ambiguous, wide-coverage grammars such as the Penn Treebank grammar described in Ch. 12 is problematic. The reason for this lies in the potentially exponential number of parses that are possible for certain inputs. Consider the following ATIS example:

(13.4) Show me the meal on Flight UA 386 from San Francisco to Denver.

The recursive  $VP \rightarrow VP PP$  and  $Nominal \rightarrow Nominal PP$  rules conspire with the three prepositional phrases at the end of this sentence to yield a total of 14 parse trees for this sentence. For example *from San Francisco* could be part of the *VP* headed by *show* (which would have the bizarre interpretation that the showing was happening from San Francisco). Church and Patil (1982) showed that the number of parses for sentences of this type grows exponentially at the same rate as the number of parenthesizations of arithmetic expressions.

## LOCAL AMBIGUITY

Even if a sentence isn't ambiguous (ie. it doesn't have more than one parse in the end), it can be inefficient to parse due to **local ambiguity**. Local ambiguity occurs when some part of a sentence is ambiguous, that is, has more than one parse, even



|                                 |         |
|---------------------------------|---------|
| computational linguistics       | 17174.4 |
| natural language                | 14603.3 |
| <b>machine translation</b>      | 11205.8 |
| natural language processing     | 10384.7 |
| training data                   | 10363.8 |
| language processing             | 8774.0  |
| <b>language model</b>           | 7700.8  |
| <b>test set</b>                 | 7253.1  |
| machine learning                | 6574.7  |
| statistical machine translation | 6513.7  |

|                                  |        |
|----------------------------------|--------|
| international conference         | 5505.3 |
| <b>word sense</b>                | 5226.8 |
| <b>named entity</b>              | 5067.4 |
| information retrieval            | 4810.6 |
| <b>word sense disambiguation</b> | 4715.3 |
| training set                     | 4228.3 |
| <b>question answering</b>        | 4159.2 |
| annual meeting                   | 4102.0 |
| speech recognition               | 3826.6 |
| <b>word alignment</b>            | 3768.4 |

Ongoing experiments: clustering document sets for sub-domains to extract important terms for main research fields (e.g. "statistical machine translation"), finding suitable cluster sizes

- by ACL Anthology Network citation graphs: depth 2,3,4
- by workshop/conference session title



- ☆ Semantic Search and Citation Sentiment Analysis would benefit from resolved pronouns, synonyms etc.
  - Semantic Search: replace pronouns etc. in follow-up sentences by antecedent → more robustness, redundancy
  - Citation Type Analysis: also include follow-up sentences with e.g. pronouns to compute sentiment more reliably
  
- ☆ Not only pronouns – coreference resolution itself also has to rely on ontology information (multi-word terms!):

"For all performance metrics, we show the 70% **confidence interval** with respect to the MAP baseline computed using bootstrap resampling (Press et al., 2002; Och, 2003). We note that **this significance level** does meet the customary criteria for **minimum significance intervals** of 68.3% (Press et al., 2002)." [N02-1022]

# TAKE Coreference Annotation: Preliminary Figures

| types    | instances |    | word counts |       | sentence spans |        |
|----------|-----------|----|-------------|-------|----------------|--------|
|          | #         | %  | mean        | sd    | mean           | sd     |
| def-np   | 10891     | 36 | 3.77        | 3.38  | 99.43          | 89.65  |
| ne       | 7427      | 25 | 2.87        | 2.25  | 140.75         | 76.63  |
| pper     | 6194      | 21 | 1.02        | 0.45  | 163.25         | 87.52  |
| ppos     | 2393      | 8  | 1.01        | 0.24  | 145.71         | 100.92 |
| indef-np | 2012      | 7  | 6.08        | 6.47  | 75.13          | 95.17  |
| conj-np  | 545       | 2  | 11.47       | 13.30 | 44.18          | 79.97  |
| other    | 316       | 1  | 2.21        | 2.64  | 106.66         | 94.58  |

| # | types               | tp  | fp  | fn  | Prec  | Recall | F1    |
|---|---------------------|-----|-----|-----|-------|--------|-------|
| 1 | Overall             | 188 | 343 | 228 | 0.354 | 0.452  | 0.397 |
| 2 | Overall(with "we")  | 812 | 562 | 438 | 0.591 | 0.650  | 0.619 |
| 3 | "it" in Overall     | 74  | 153 | 127 | 0.326 | 0.368  | 0.346 |
| 4 | citation in Overall | 58  | 77  | 73  | 0.430 | 0.443  | 0.436 |
| 5 | Upperbound          | 188 | 135 | 228 | 0.582 | 0.452  | 0.509 |

(tp: true positives; fp: false positives; fn: false negatives).

"citation": result on pronoun resolution in a 3-sentences window for citation sentences

Corpus: 63 papers of ACL Anthology for training + 12 for testing



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

Ulrich Schäfer

DFKI Language Technology Lab, Saarbrücken

Uwe Kasterka

Computer Science Dept., Saarland University

# TAKE Typed Citation Analysis: 5 Categories



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

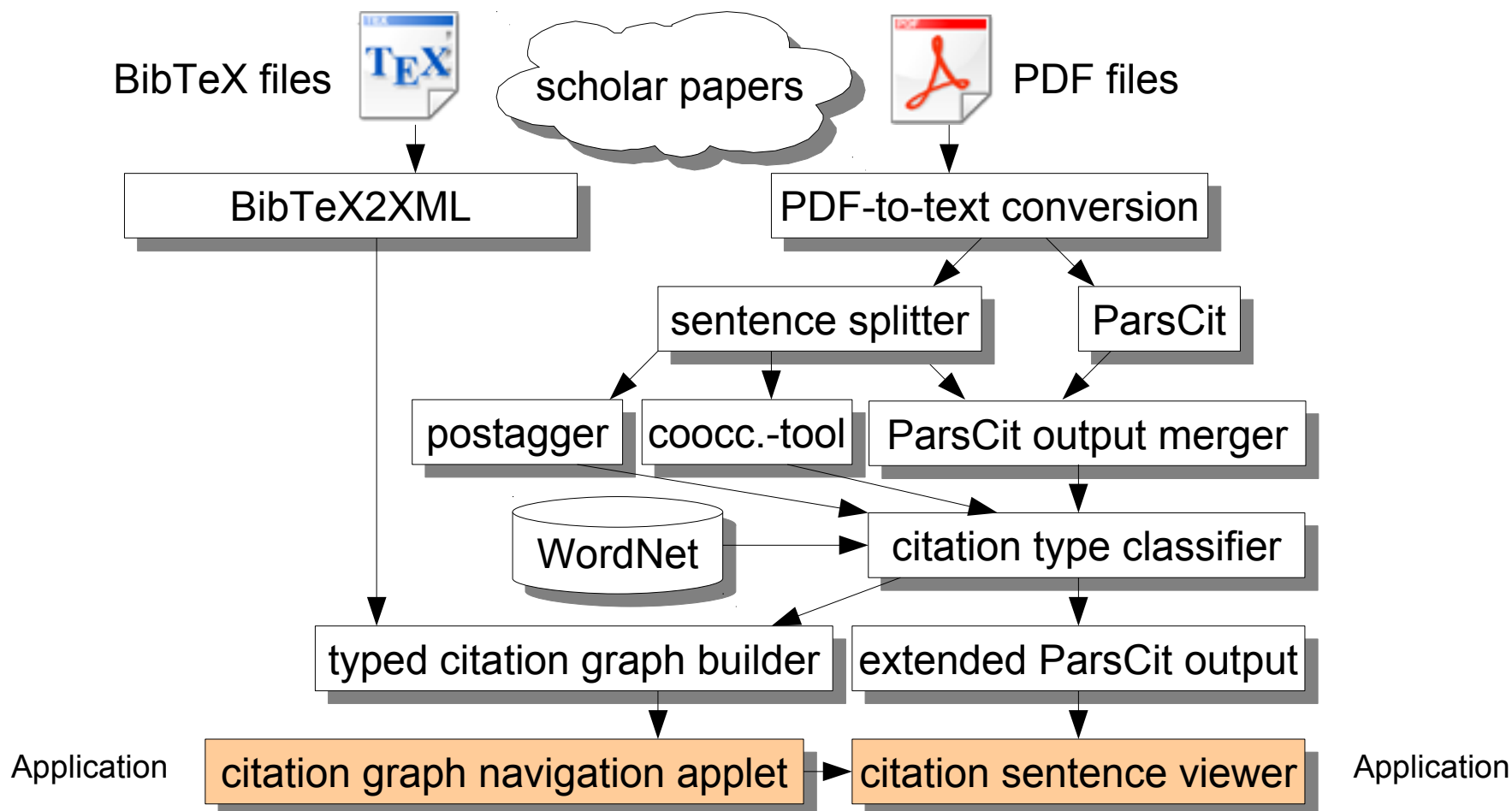


☆ Blend of methods to collect verbal and non-verbal patterns (cue words):

- WordNet synonyms and antonyms (the latter for increasing number of patterns for negative citations)
- DiMarco/Garzone list devised for biomedical texts; largely applicable to computational linguistics
- Negating positive cue words
- Using automatically extracted cooccurrences (Ted Pedersen's cooccurrence tool) on citation sentences
- Inspection: browse and filter cue words manually



- Graph building:
  - 10821 links shared with ACL Anthology Network (AAN; Radev et al 2009)
  - 3883 in AAN not recognized by us, 1021 by us not in AAN
  - Different subsets, no gold data for publications outside ACLA
  
- Citation classification
  - Evaluation on 100 citations
  - 30% correct (90% undef are neutral, negative unreliable @33%, neutral: 60% correct, Precycle 33%, Agree 25%)



# TAKE ParsCit Output Extended with ACL Document ID and Citation Types

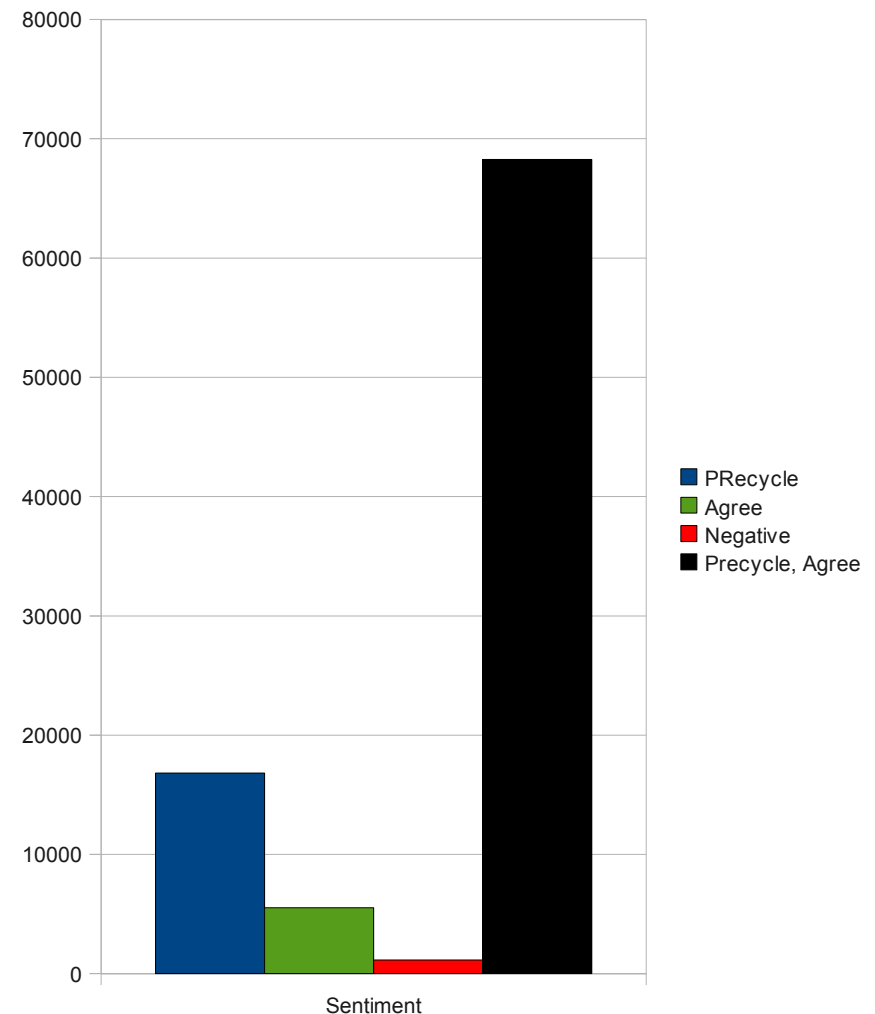


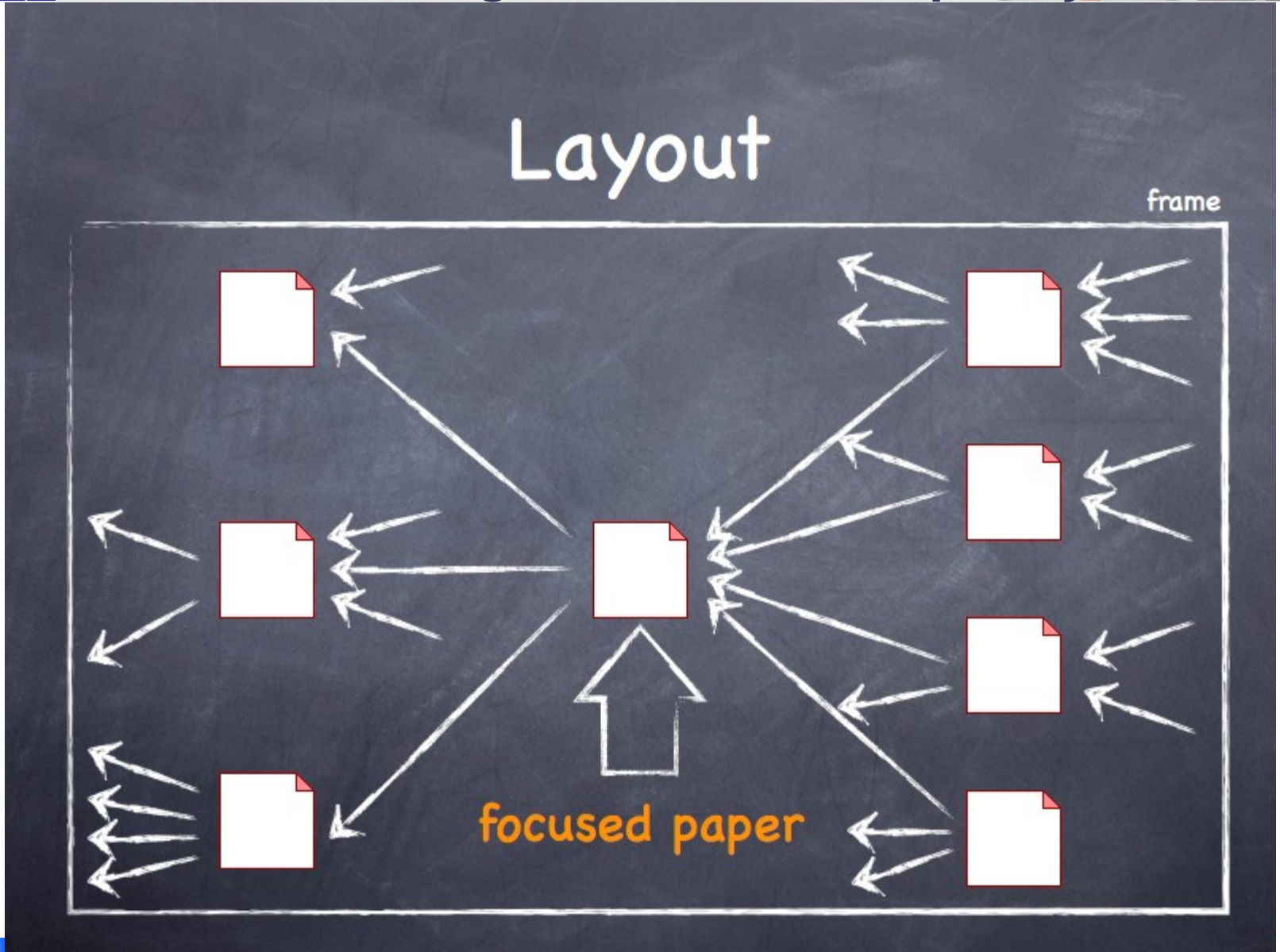
```
<citation sentiment="negative" valid="true">
  <title>Local textual inference: can it be defined or circumscribed</title>
  <rawString>Annie Zaenen, Lauri Karttunen, and Richard S. Crouch. 2005. Local
    textual inference: can it be defined or circumscribed? In ACL 2005 Workshop
    on Empirical Modeling of Semantic Equivalence and Entailment.</rawString>
  <marker>Zaenen, Karttunen, Crouch, 2005</marker>
  <authors>
    <author>Annie Zaenen</author>
    <author>Lauri Karttunen</author>
    <author>Richard S Crouch</author>
  </authors>
  <aclId>W05-1206</aclId>
  <date>2005</date>
  <booktitle>In ACL 2005 Workshop on Empirical Modeling of Semantic Equivalence
    and Entailment</booktitle>
  <scores Agree="0.0" PRecycle="0.0" negative="0.021" neutral="0.0"/>
  <contexts>
    <context position="7150" sentiment="negative">
      <scores Agree="0.0" PRecycle="0.0" negative="0.021" neutral="0.0"/>
      <text>urnstile at Stockwell subway station. (2) The documents leaked to
        ITV News suggest that Menezes walked casually into the subway station. This
        example contains an &quot;embedded contradiction.&quot; Contrary to
        Zaenen et al. (2005), we argue that recognizing embedded contradictions is
        important for the application of a contradiction detection system: if John
        thinks that he is incompetent, and his boss believes that John is not </text>
```

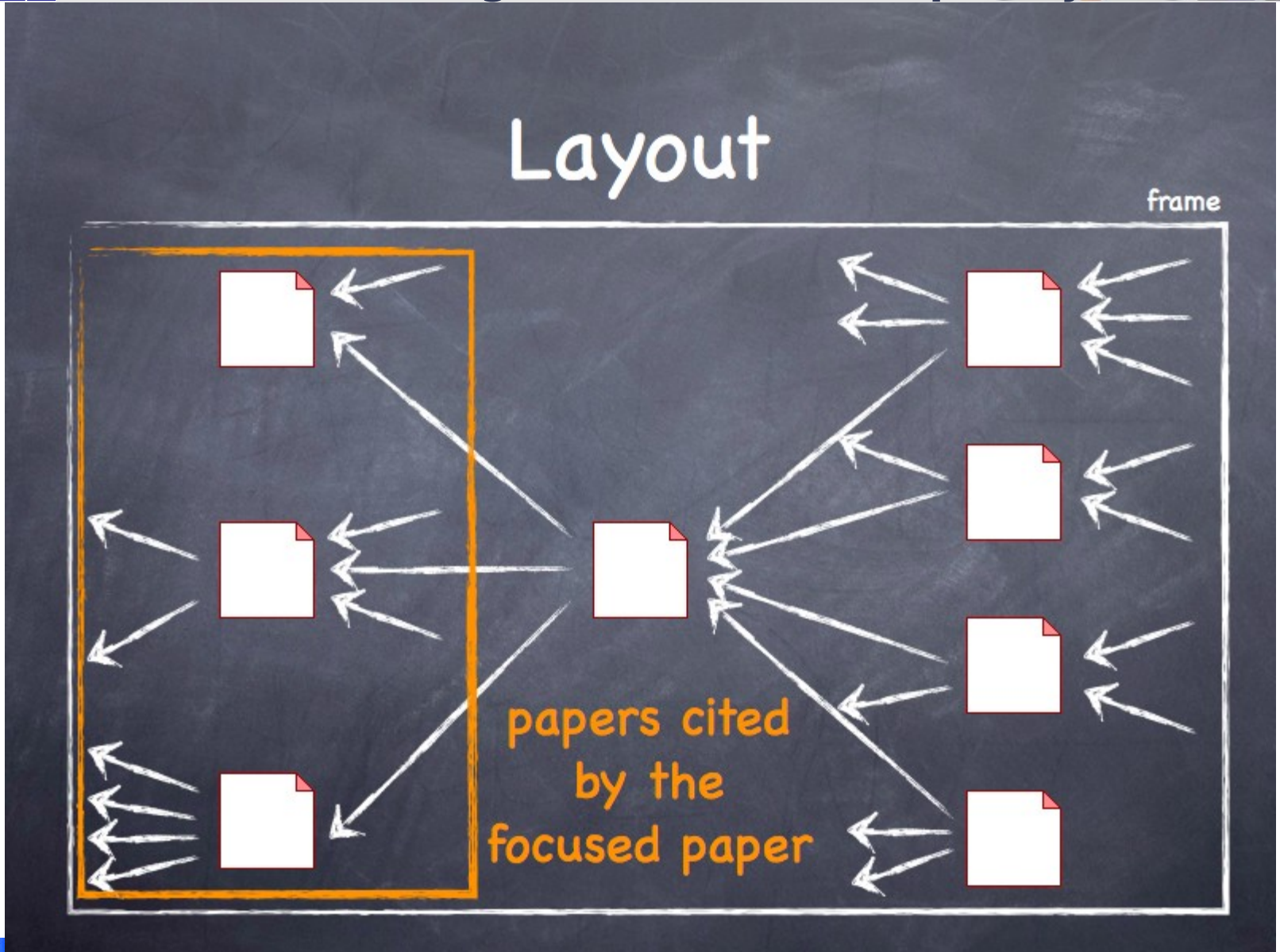


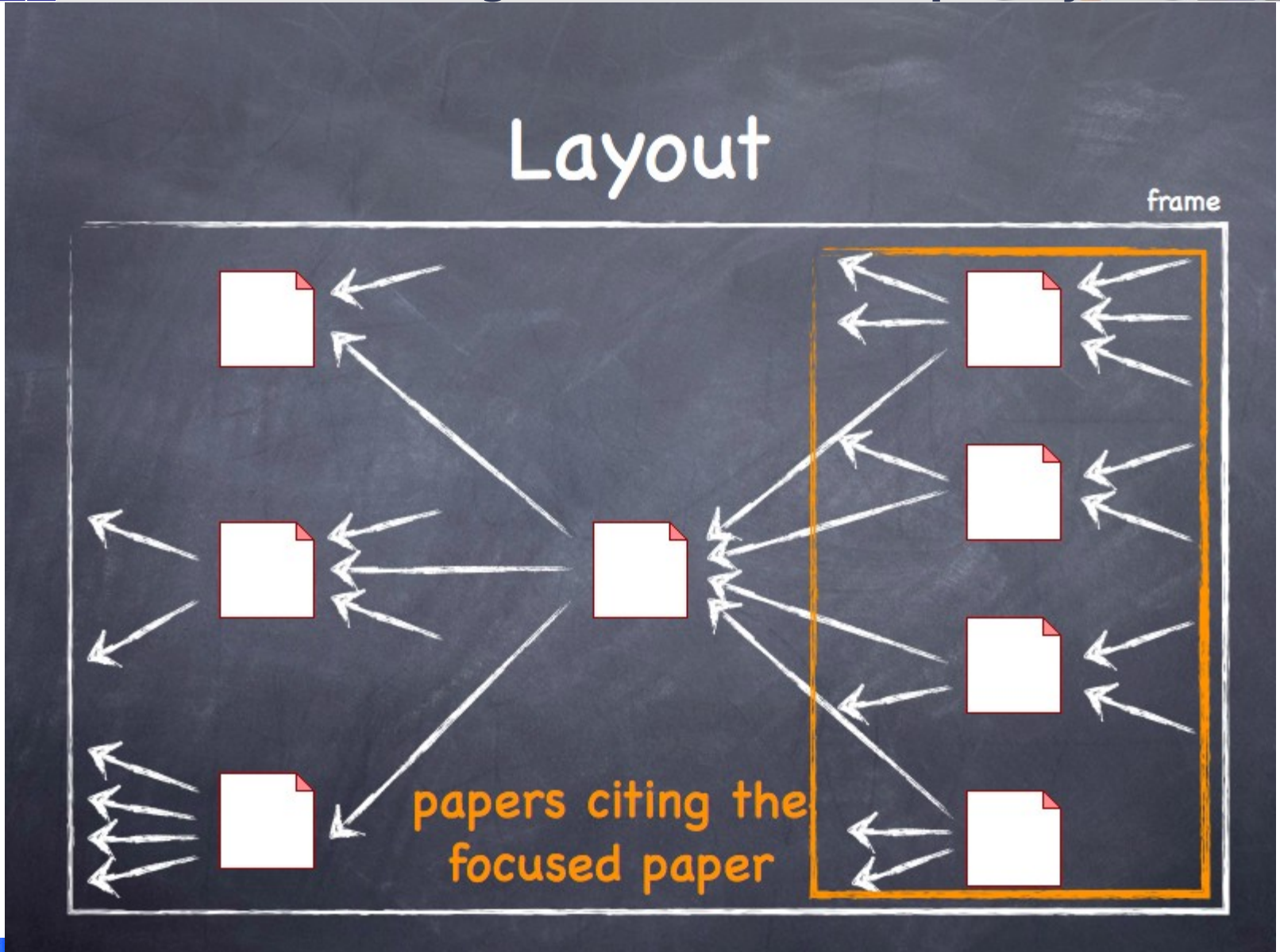
Current numbers out of a total  
of 91419 citations:

- ☆ 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%)

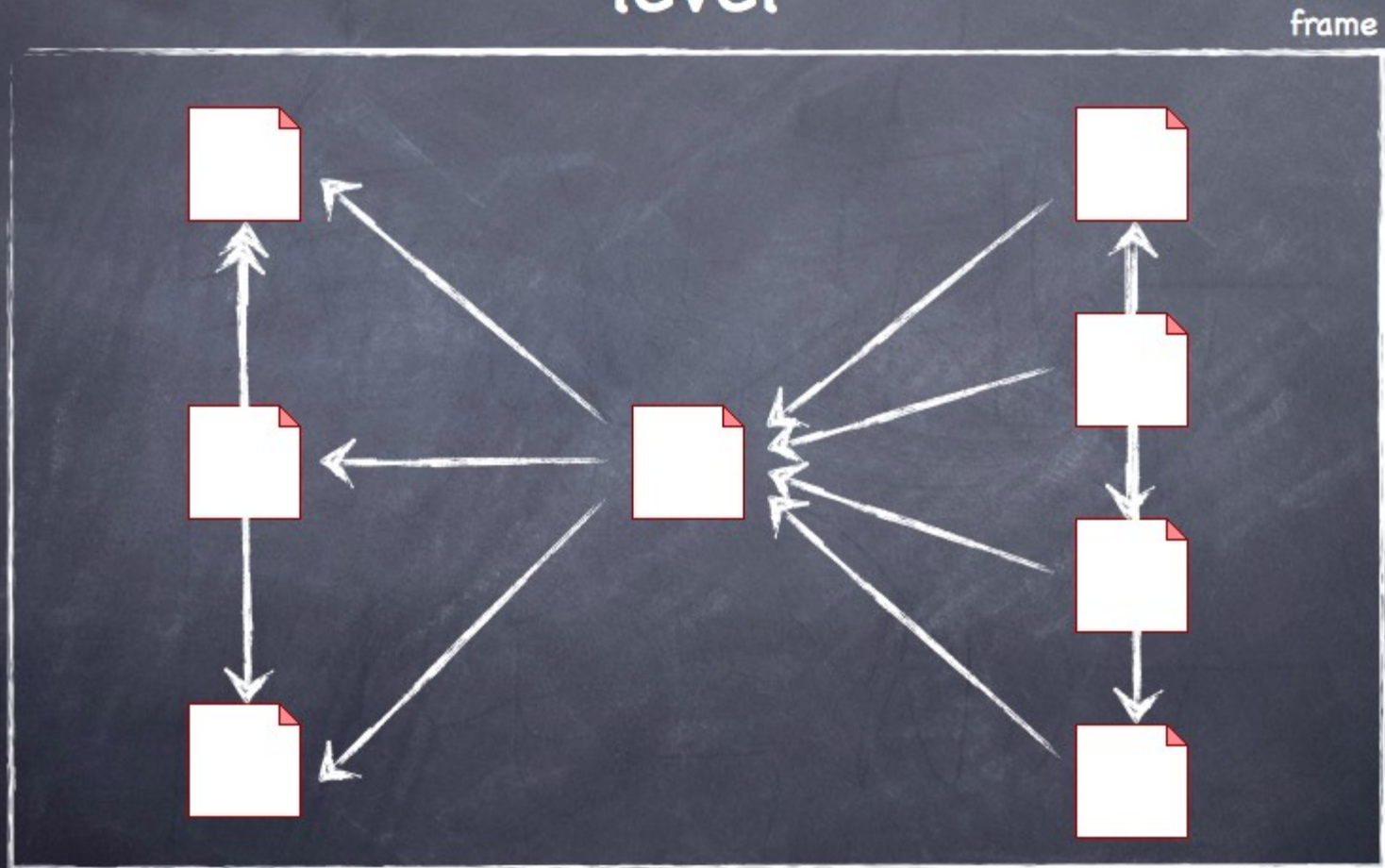




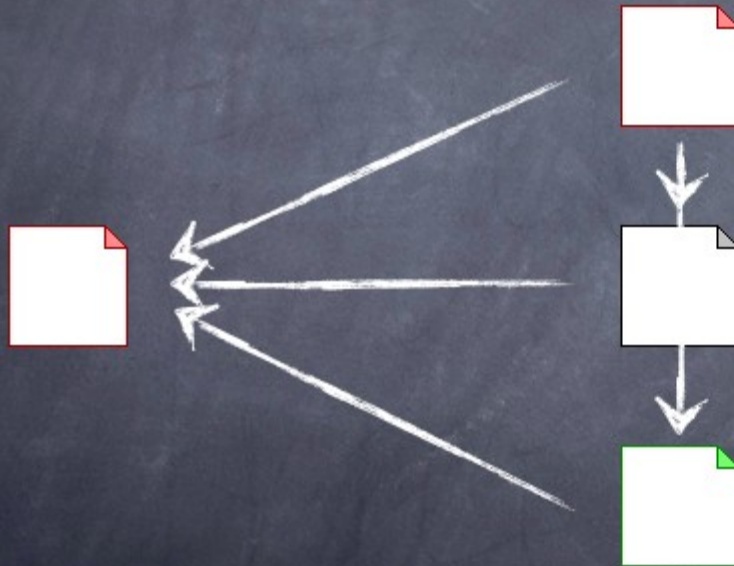




# Problem – citations on the same level

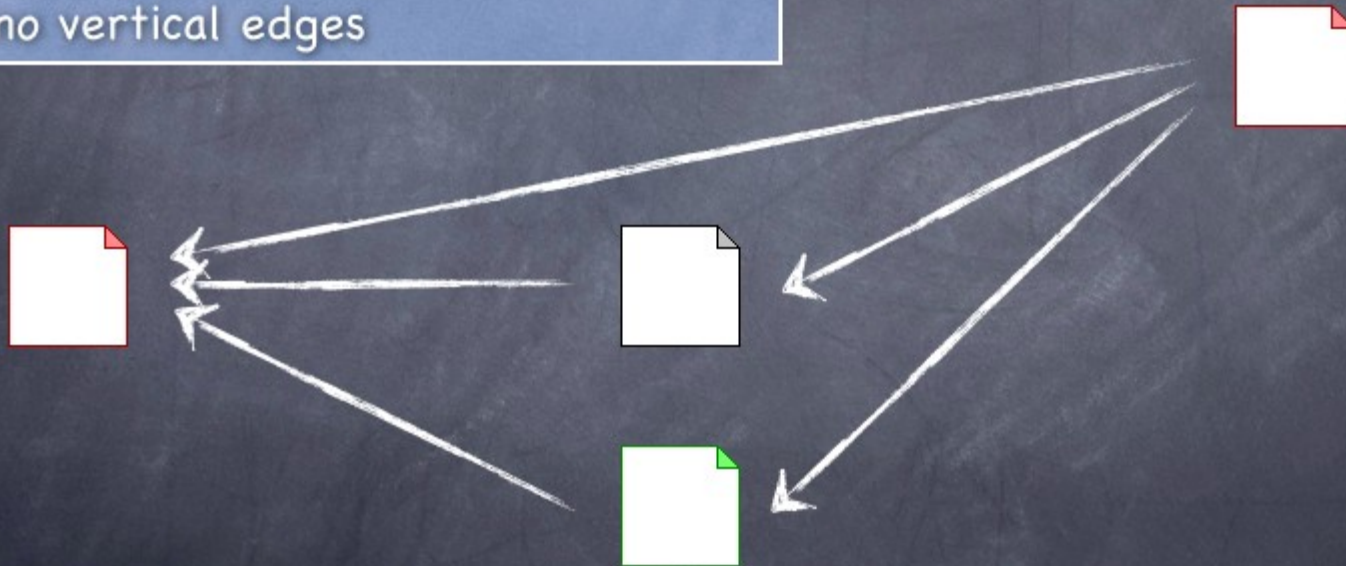


# Fan-out Example

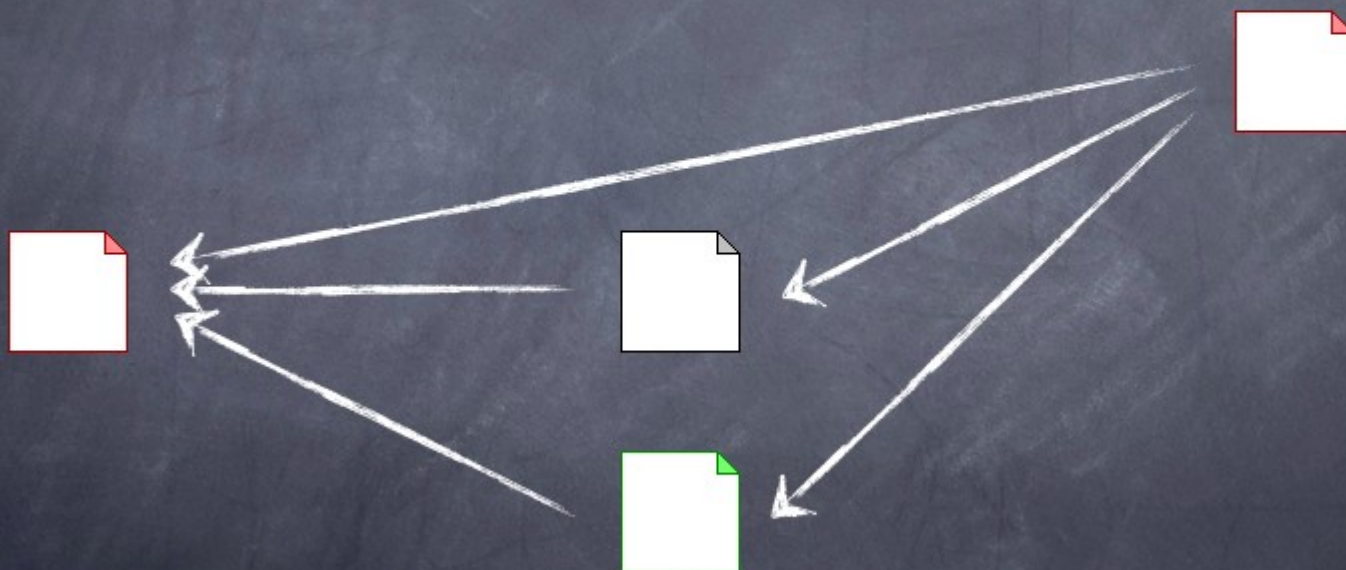


# Fan-out Example

- all edges have the same direction
- no vertical edges

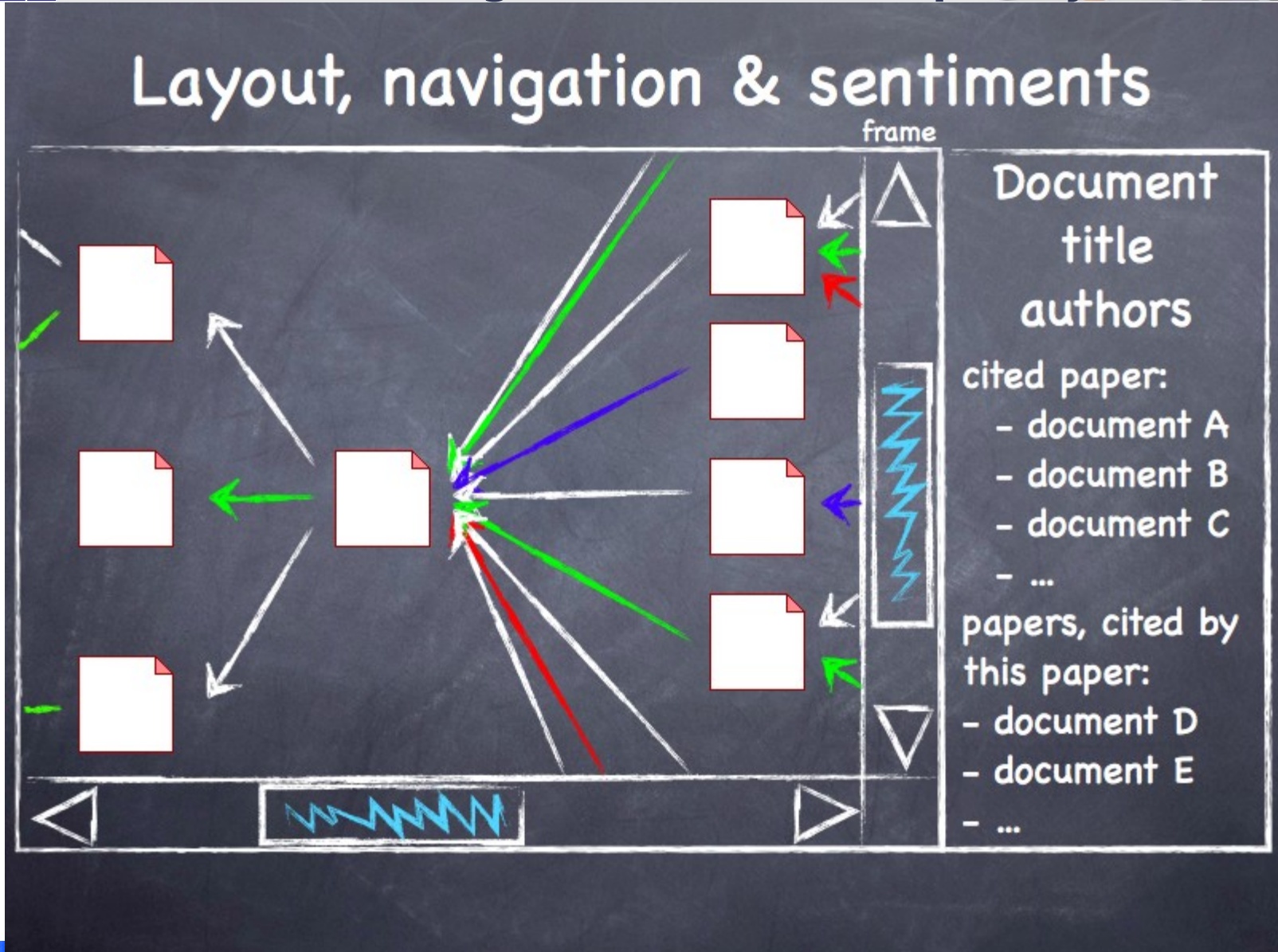


# Fan-out Example



$\text{depth}(\text{focused\_node}) = 0$

$\text{depth}(n) = 1 + \max \{ \text{depth}(x) \mid x \in \text{predecessor}(n) \}$

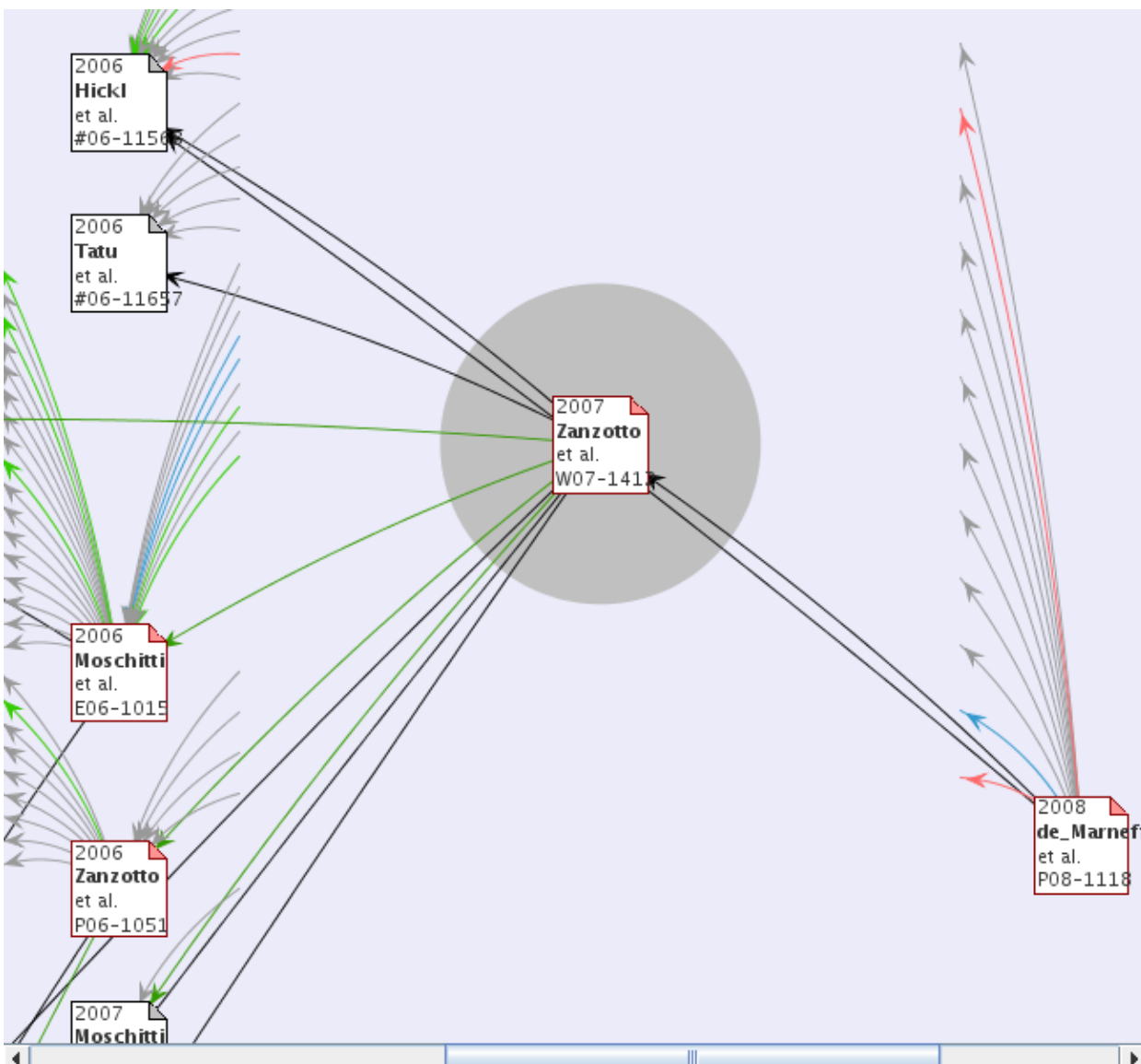


# Shallow Semantic in Fast Textual Entailment Rule Learners

[**Fabio Massimo Zanzotto**, **Marco Pennacchiotti**, **Alessandro Moschitti**] (ID: W07-1412)

## Cited papers:

- [Roy Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, Idan Szpektor]  
[#06-3178](#)
- The II PASCAL RTE challenge  
[P02-1034](#) [Michael Collins, Nigel Duffy]
- New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron  
[W05-1203](#) [Courtney Corley, Rada Mihalcea]
- Measuring the Semantic Similarity of Texts  
[Andrew Hickl, Jeremy Bensley, John Williams, Kirk Roberts, Bryan Rink, Ying Shi]  
[#06-11568](#)
- Recognizing Textual Entailment with LCC's Groundhog System  
[#99-1625](#) [T Joachims]
- Making large-scale SVM learning practical



## Navigation

Back

Forward

## Zoom



Center

Overview

to PaperId

☒ Calculate Endings

☒ Calculate Fanout

## Citation depth



1 2 3 4 5 6 7 8 9 10

## Sentiments

Negative Neutral  
Agree Undefined  
PRecycle

# TAKE Citation Sentence Context and PDF Highlighting

contradiction, those featuring negation and those formed by paraphrases. They constructed two corpora for evaluating their system. ... - Sentiment: .

Citation of **Marie-Catherine de Marneffe, Bill MacCartney, Christopher D Manning. 2006: *Generating typed dependency parses from phrase structure parses*. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06)*. Christiane Fellbaum. - Overall sentiment: .**

Citation of **Sanda Harabagiu, Andrew Hickl, Finley Lacatusu. 2006: *Negation, contrast, and contradiction in text processing*. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*. - Overall sentiment: **negative**.**

- **Page 2:** "... Little work has been done on contradiction detection. The PASCAL Recognizing Textual Entailment (RTE) Challenges (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007) focused on textual inference in any domain. Condoravdi et al. (2003) first recognized the importance of handling entailment and contradiction for text understanding, but they rely on a strict logical definition of these phenomena and do not report empirical results. To our knowledge, Harabagiu et al. (2006) provide the first empirical results for contradiction detection, but they focus on specific kinds of contradiction: those featuring negation and those formed by paraphrases. They constructed two corpora for evaluating their system. One was created by overtly negating each entailment in the RTE2 data, producing a balanced dataset (LCC.negation). To avoid overtraining, negative markers were also added to each non-entailment, ensuring that they did not create contradictions. ... - Sentiment: **undef**.
- **Page 4:** "... Table 2 gives the number of contradictions in each dataset. The RTE datasets are balanced between entailments and non-entailments, and even in these datasets targeting inference, there are few contradictions. Using our guidelines, RTE3\_test was annotated by NIST as part of the RTE3 Pilot task in which systems made a 3-way decision as to whether pairs of sentences were entailed, contradictory, or neither (Voorhees, 2008). **Our annotations and those of NIST were performed on the original RTE datasets, contrary to Harabagiu et al. (2006).** Because their corpora are constructed using negation and paraphrase, they are unlikely to cover all types of contradictions in section 3.2. We might hypothesize that rewriting explicit negations commonly occurs via the substitution of antonyms. Imagine, e.g.: ... - Sentiment: **negative**.
- **Page 7:** "... LCCnegation Table 5: Precision and recall figures for contradiction detection. Accuracy is given for balanced datasets only. 'LCC.negation' reflects performance of Harabagiu et al. (2006). ...

as well as an entity that was not involved. However, different outcomes result because a tunnel connects only two unique locations whereas more than one entity may purchase food. These frequent interactions between world-knowledge and structure make it hard to ensure that any particular instance of structural mismatch is a contradiction.

### 3.3 Contradiction corpora

Following the guidelines above, we annotated the RTE datasets for contradiction. These datasets contain pairs consisting of a short text and a one-sentence hypothesis. Table 2 gives the number of contradictions in each dataset. The RTE datasets are balanced between entailments and non-entailments, and even in these datasets targeting inference, there are few contradictions. Using our guidelines, RTE3\_test was annotated by NIST as part of the RTE3 Pilot task in which systems made a 3-way decision as to whether pairs of sentences were entailed, contradictory, or neither (Voorhees, 2008).<sup>1</sup>

Our annotations and those of NIST were performed on the original RTE datasets, contrary to Harabagiu et al. (2006). Because their corpora are constructed using negation and paraphrase, they are unlikely to cover all types of contradictions in section 3.2. We might hypothesize that rewriting explicit negations commonly occurs via the substitution of antonyms. Imagine, e.g.:

H: Bill has finished his math.

<sup>1</sup>Information about this task as well as data can be found at <http://nlp.stanford.edu/RTE3-pilot/>.

paraphrase corpora is likely to leave one 'easy' contradictions and addresses f of contradictions (table 3). We cont authors to obtain their datasets, but th to make them available to us. Thus, w LCC.negation corpus, adding negat the RTE2 test data (Neg\_test), and to set (Neg\_dev) constructed by random pairs of entailments and 50 pairs of n from the RTE2 development set.

Since the RTE datasets were consi tual inference, these corpora do not re contradictions. We therefore collec tions 'in the wild.' The resulting c 131 contradictory pairs: 19 from new looking at related articles in Google Wikipedia, 10 from the Lexis Nexis 51 from the data prepared by LDC for task of the DARPA GALE program. I domness of the collection, we argue t best reflects naturally occurring contr

Table 3 gives the distribution of types for RTE3\_dev and the real corpus. Globally, we see that contradicti (2) occur frequently and dominate the ment set. In the real contradiction co much higher rate of the negation, nu ical contradictions. This supports th in the real world, contradictions prim two reasons: information is updated

<sup>2</sup>Our corpora—the simulation of the LLC the RTE datasets and the real contradictions <http://nlp.stanford.edu/projects/contradiction>.



- ☆ Incorporate results from deep parsing in coreference resolution, ontology extraction, citation classification
- ☆ Re-parse extended corpus with FSC-PET, improved shallow-deep integration
- ☆ Increase coverage on long sentences by pre-structuring with Stanford Parser
- ...