NLP for Science Information Systems

Ulrich Schäfer

DFKI Language Technology Lab

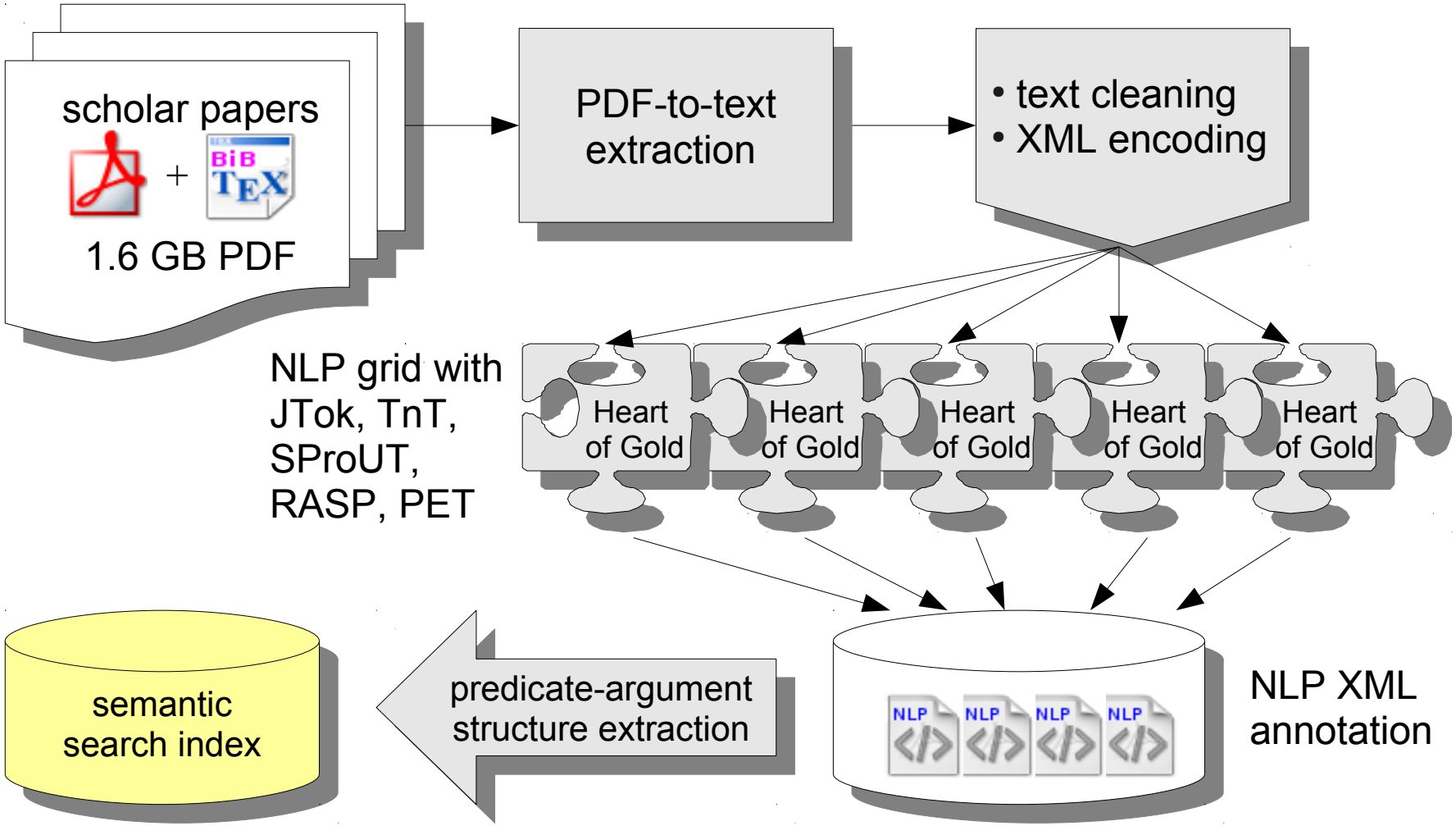
NLP for Science Information Systems

- ☆ 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

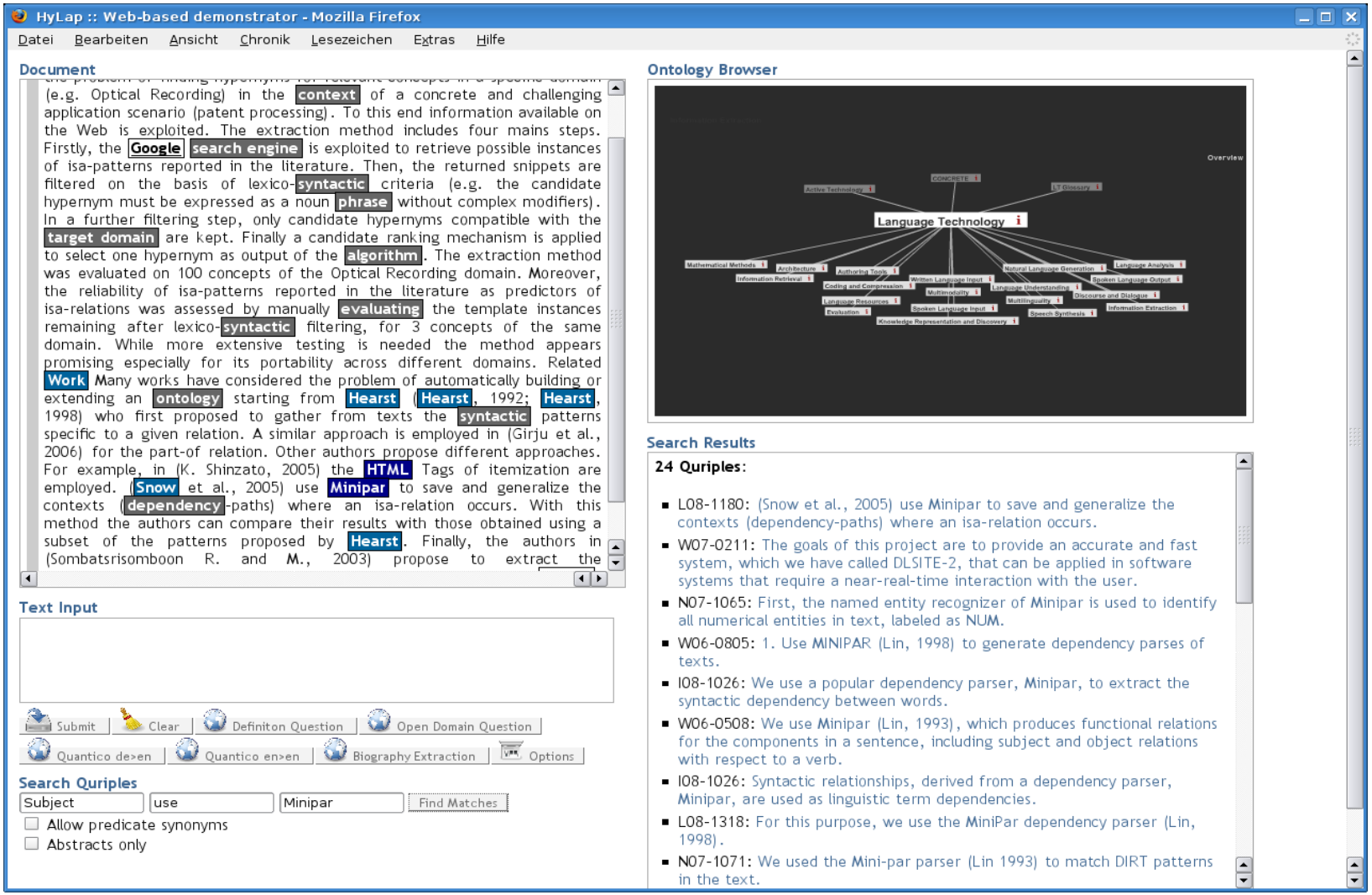
- ☆ corpus: ACL Anthology (initially 6300 CL/LT conference and workshop papers, now extended to 8500, including CL Journal 2002-2009)

Semantic Search Index Generation

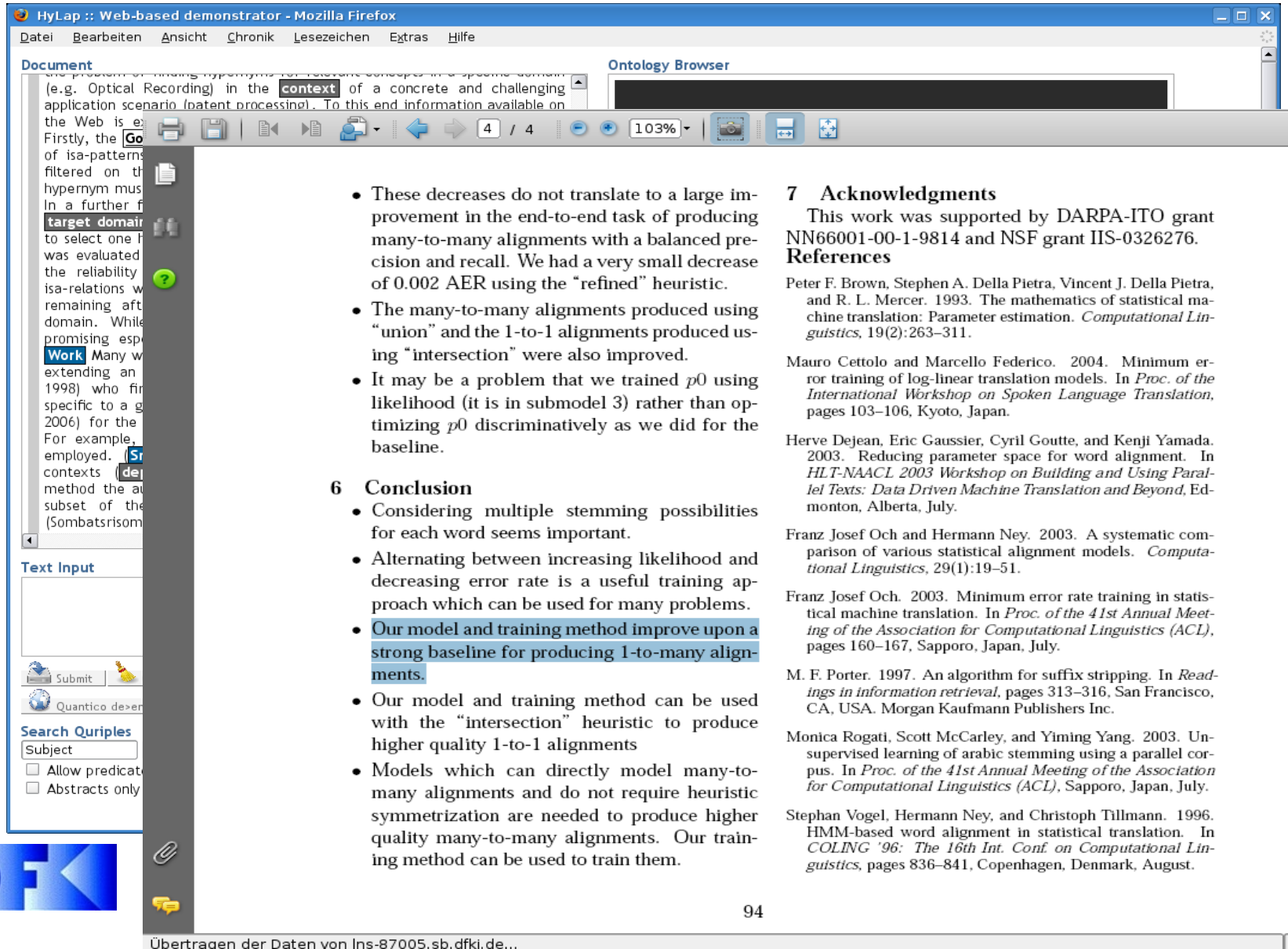


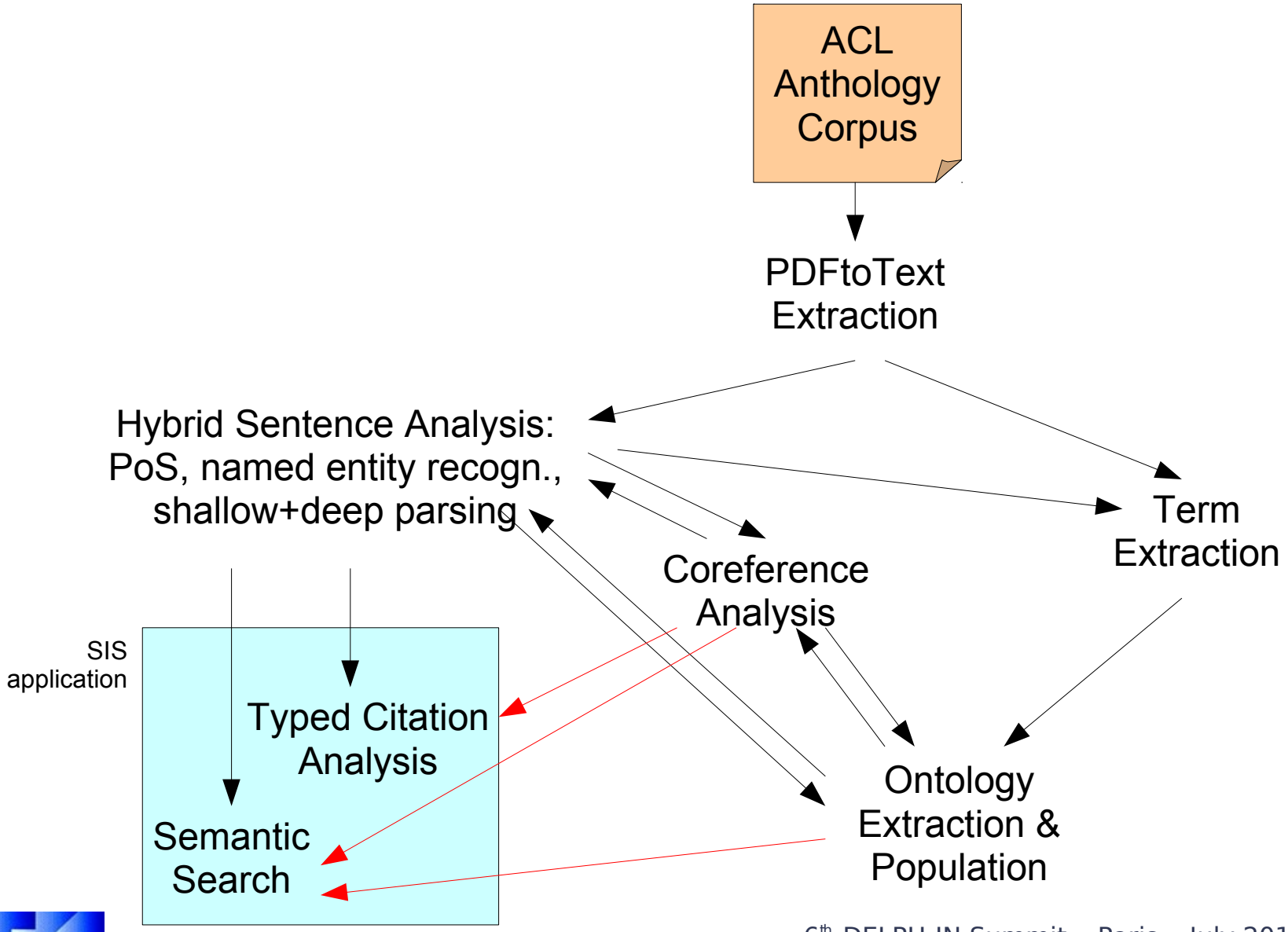
700 MB Apache Solr Blob (953 000 sentences)

Semantic Search in 6300 ACL Anthology papers



Semantic Search in 6300 ACL Anthology papers




ACL Anthology Corpus
PDFtoText Extraction
Hybrid Sentence Analysis: PoS, named entity recogn., shallow+deep parsing
Term Extraction
Coreference Analysis
SIS application
Typed Citation Analysis
Semantic Search
Ontology Extraction & Population

Motivation for Term Extraction: Automatic Ontology Extraction/Learning/Population

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

Term extraction method

- ☆ 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

TAKE Term Extraction as Preparation for Automatic Ontology Extension/Learning/Population

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

international conference	5505.3
word sense	5226.8
named entity	5067.4
information retrieval	4810.6
word sense disambiguation	4715.3
training set	4228.3
question answering	4159.2
annual meeting	4102.0
speech recognition	3826.6
word alignment	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

Coreference Resolution - Motivation

- ☆ 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]

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

Pronoun Resolution: A first result with BART

#	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)

Citation Classification Method

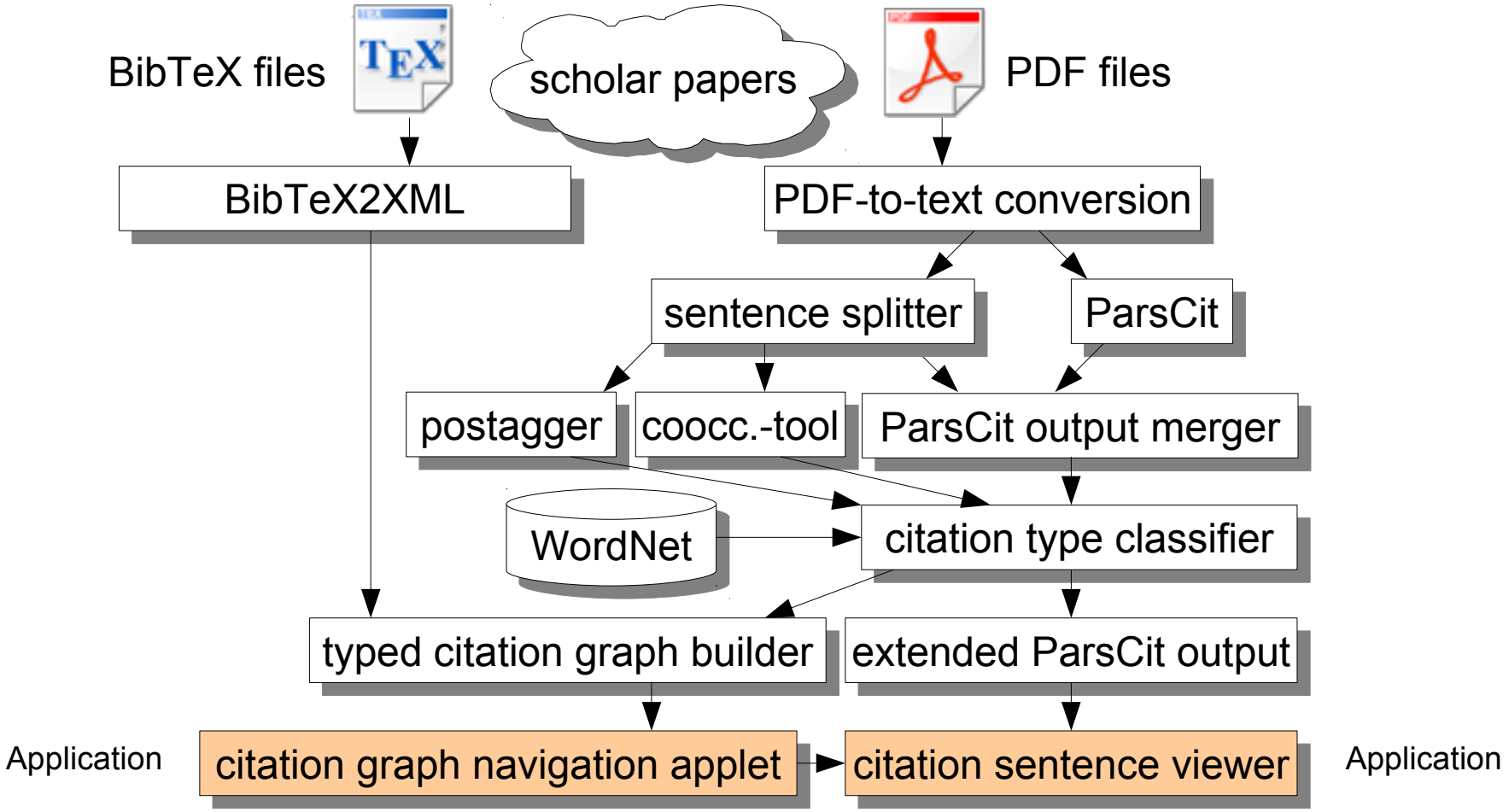
- ☆ Blend of methods to collect verbal and non-verbal patterns (cue words):
  - WordNet synonyms and antonyms (the latter for increasing number of patterns for negative citations)
  - DiMarco/Garzone list devised for biomedical texts; largely applicable to computational linguistics
  - Negating positive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

Evaluation

- 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 Classification Workflow & Application



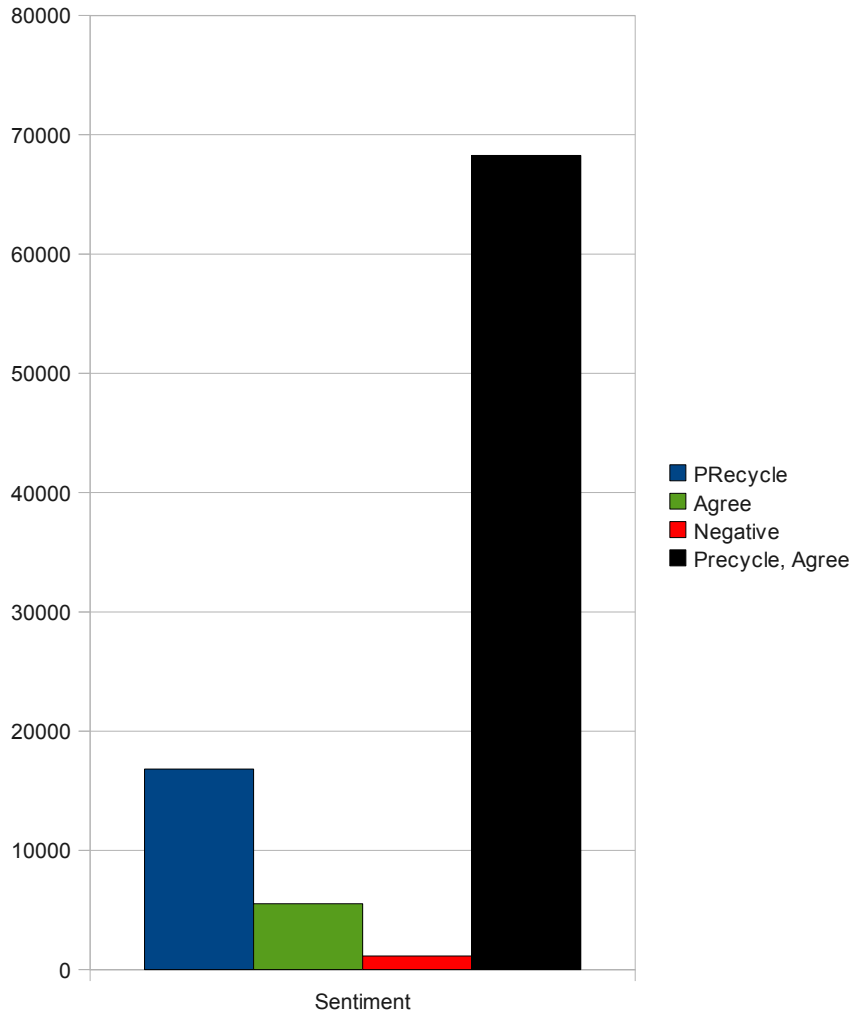
ParsCit Output Extended with ACL Document ID and Citation Types

```
<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 ofSemantic 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 &amp;quot;embedded contradiction.&amp;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>
```

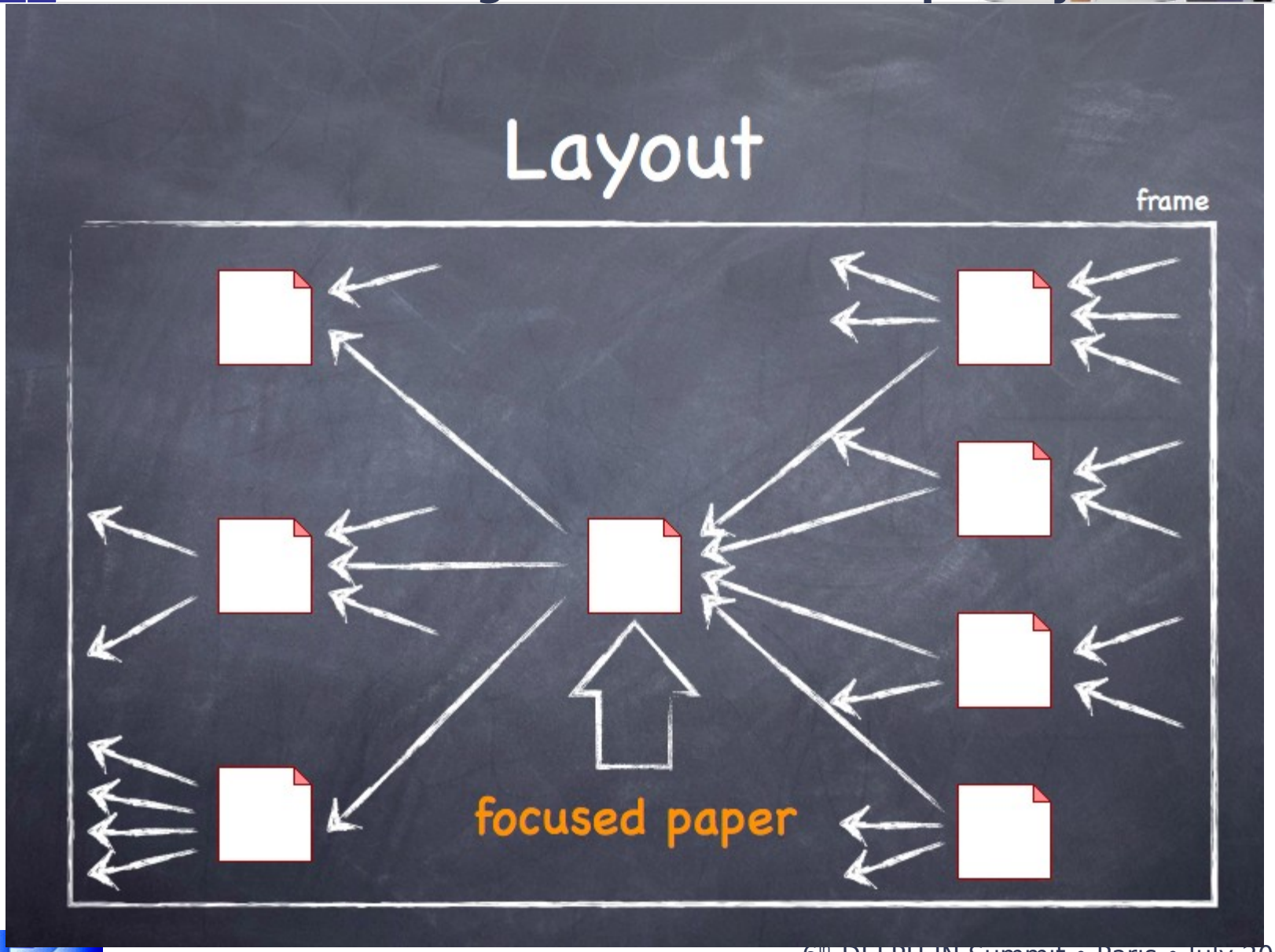
Automatic Citation Classification

Current numbers out of a total of 91419 citations:

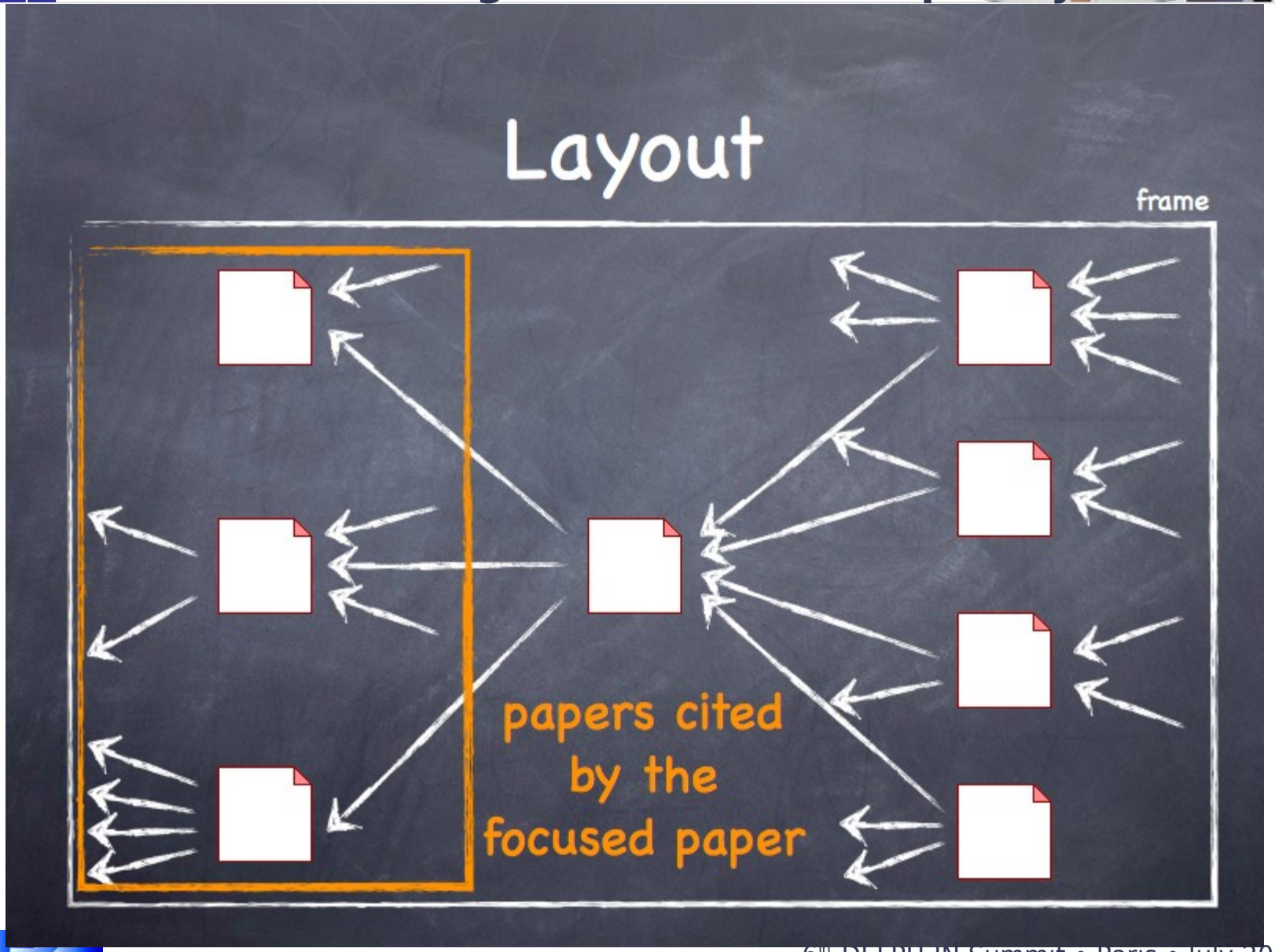
- ☆ Agree: 3513 (3.8%)
- ☆ Agree, Neutral: 2020 (2.2%)
- ☆ Negative: 1147 (1.2%)
- ☆ PRecycle: 10609 (11.6%)
- ☆ PRecycle, Agree: 1419 (1.6%)
- ☆ PRecycle, Agree, Neutral: 922 (1.0%)
- ☆ PRecycle, Neutral: 3882 (4.2%)
- ☆ Neutral: 13430 (14.7%)
- ☆ undef: 54837 (60.0%)



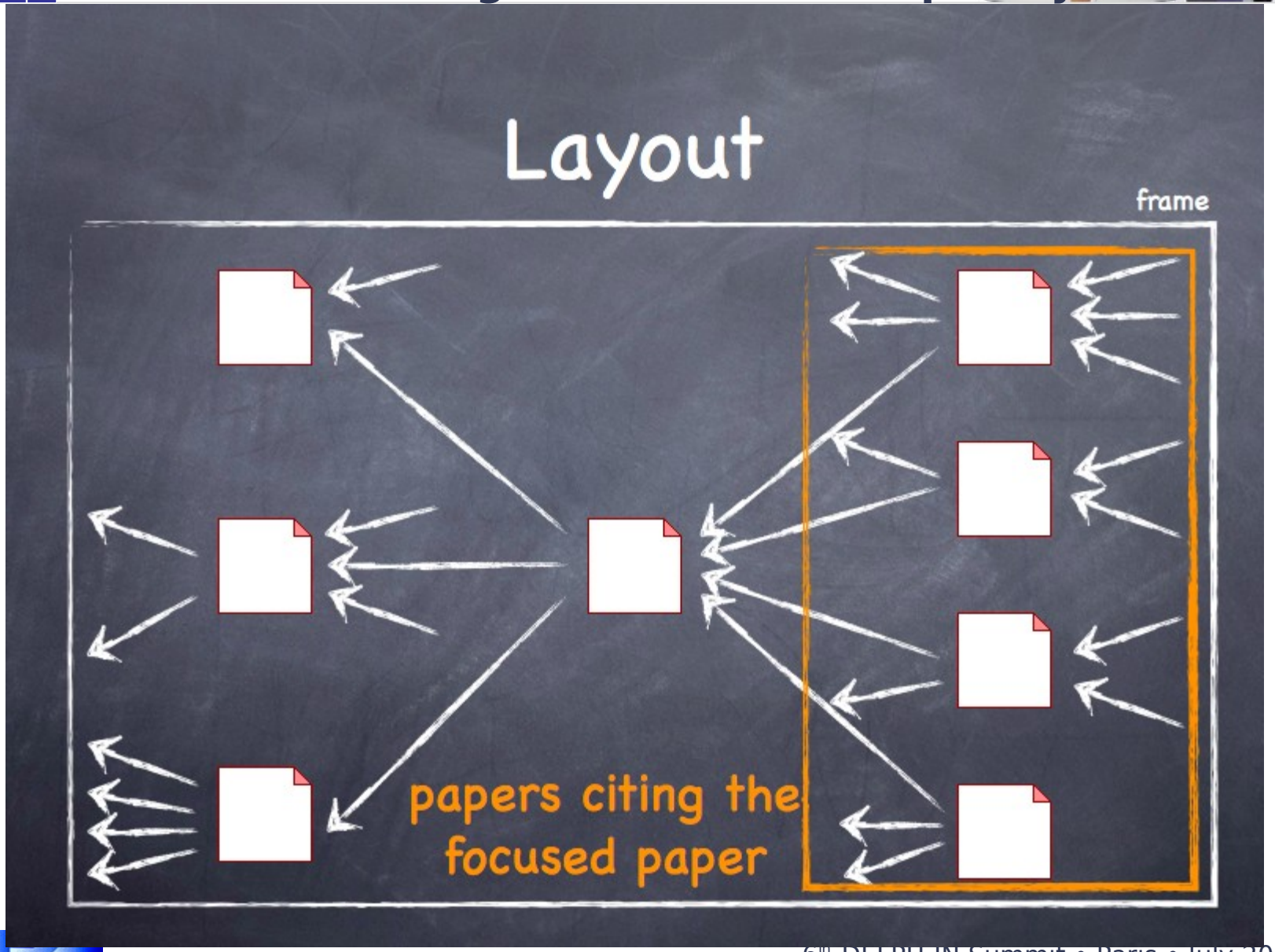
TeeCeeGeeNavigator: Citation Graph Layout



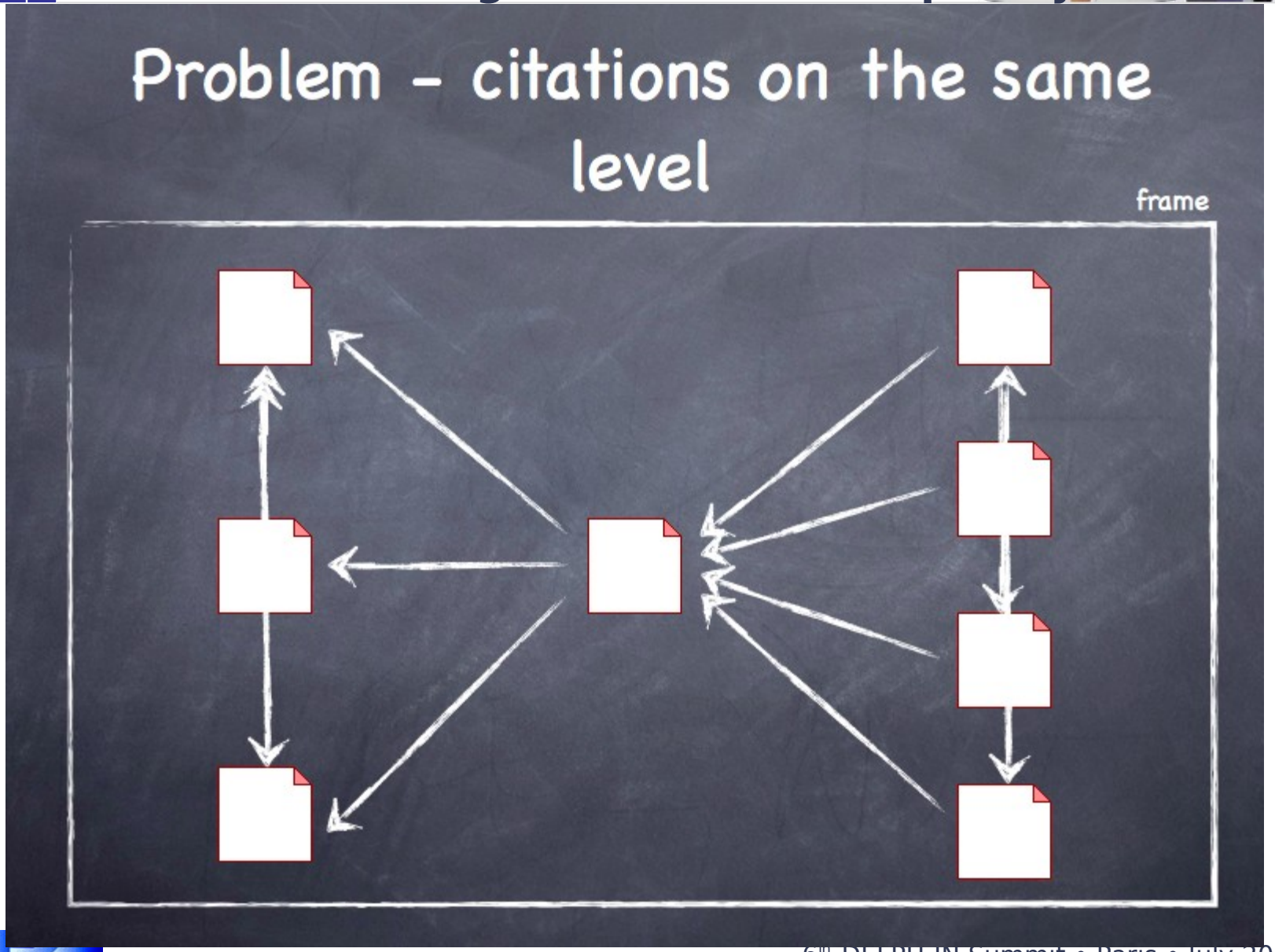
TeeCeeGeeNavigator: Citation Graph Layout



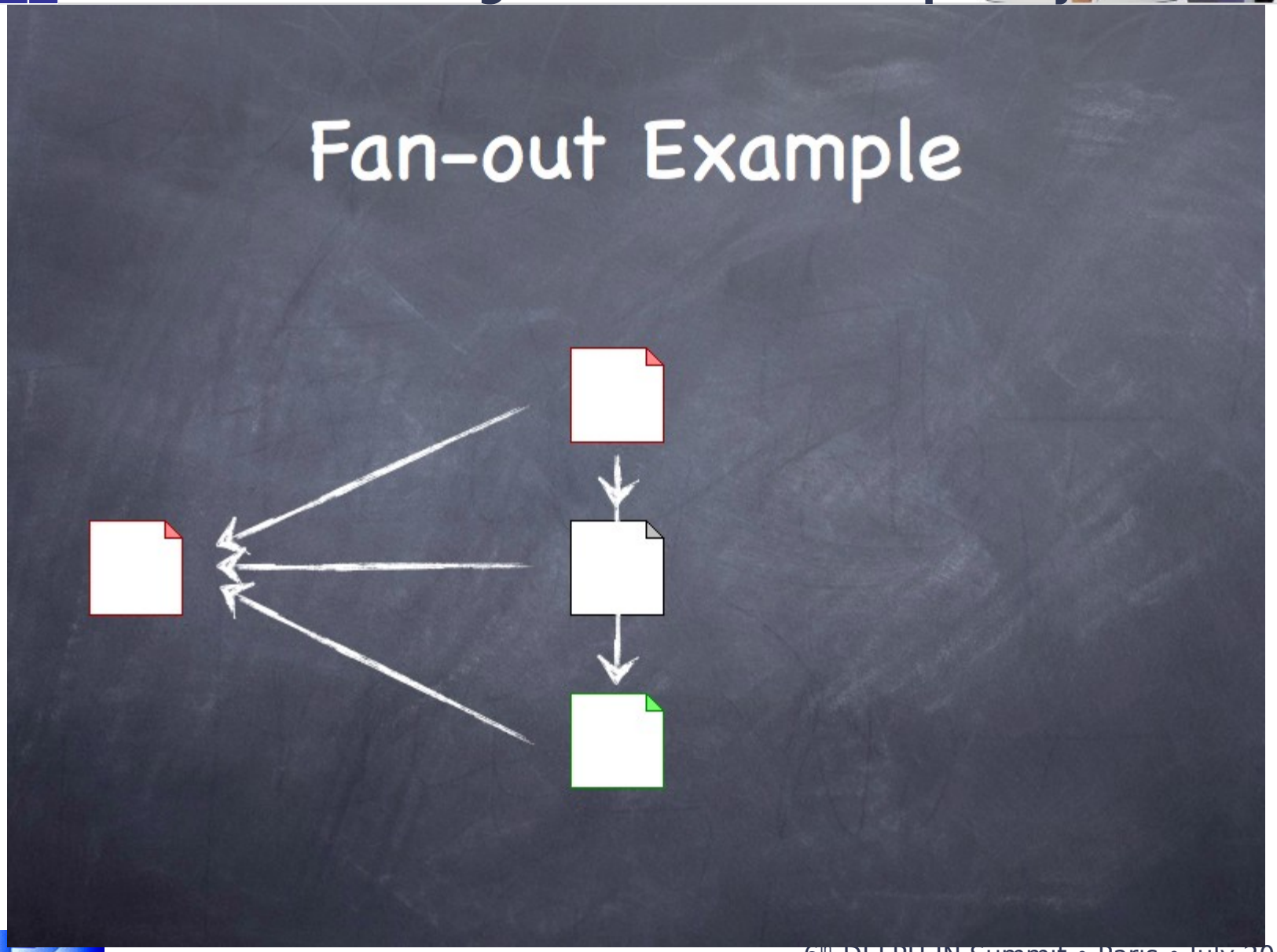
TeeCeeGeeNavigator: Citation Graph Layout



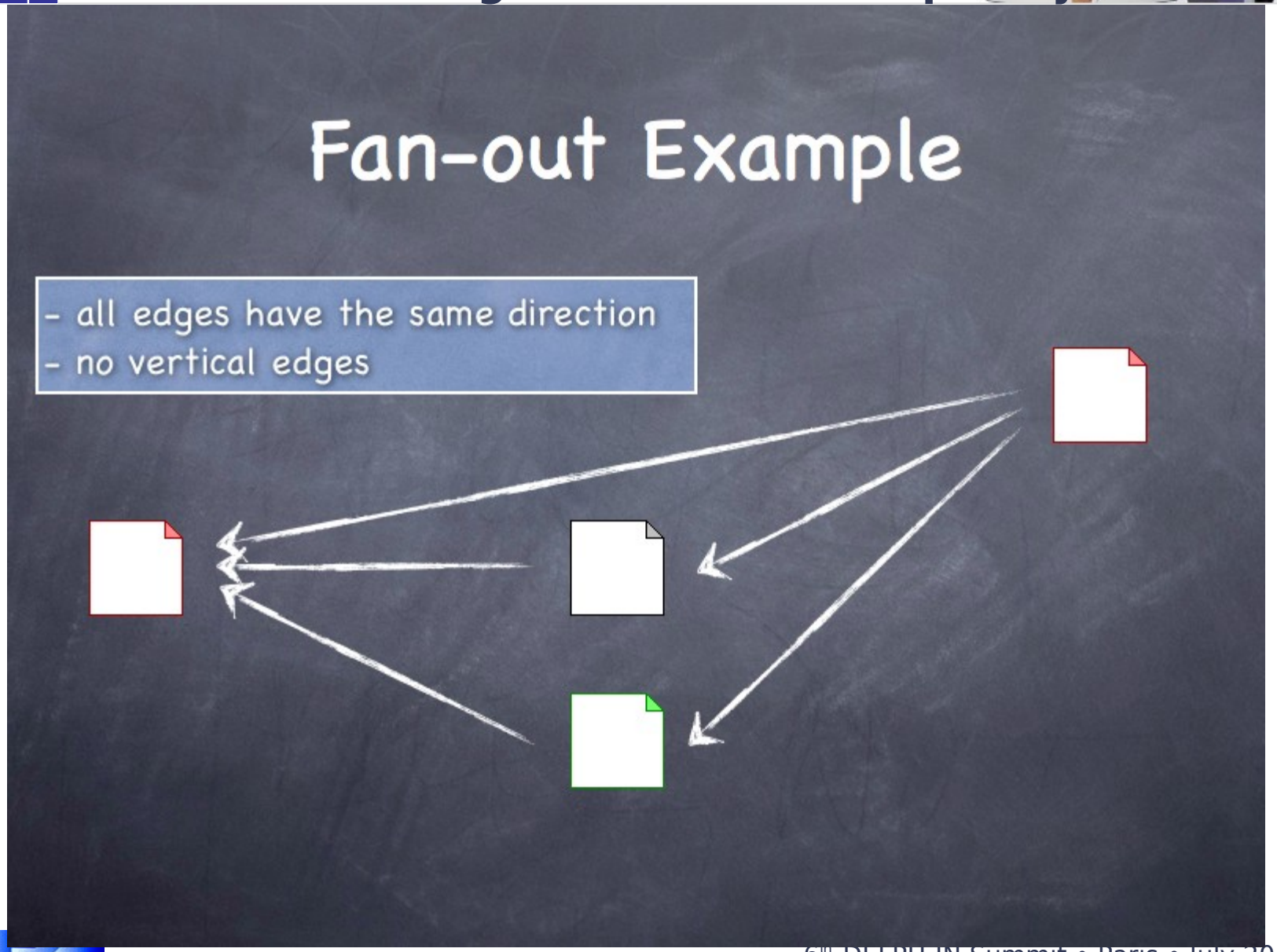
TeeCeeGeeNavigator: Citation Graph Layout



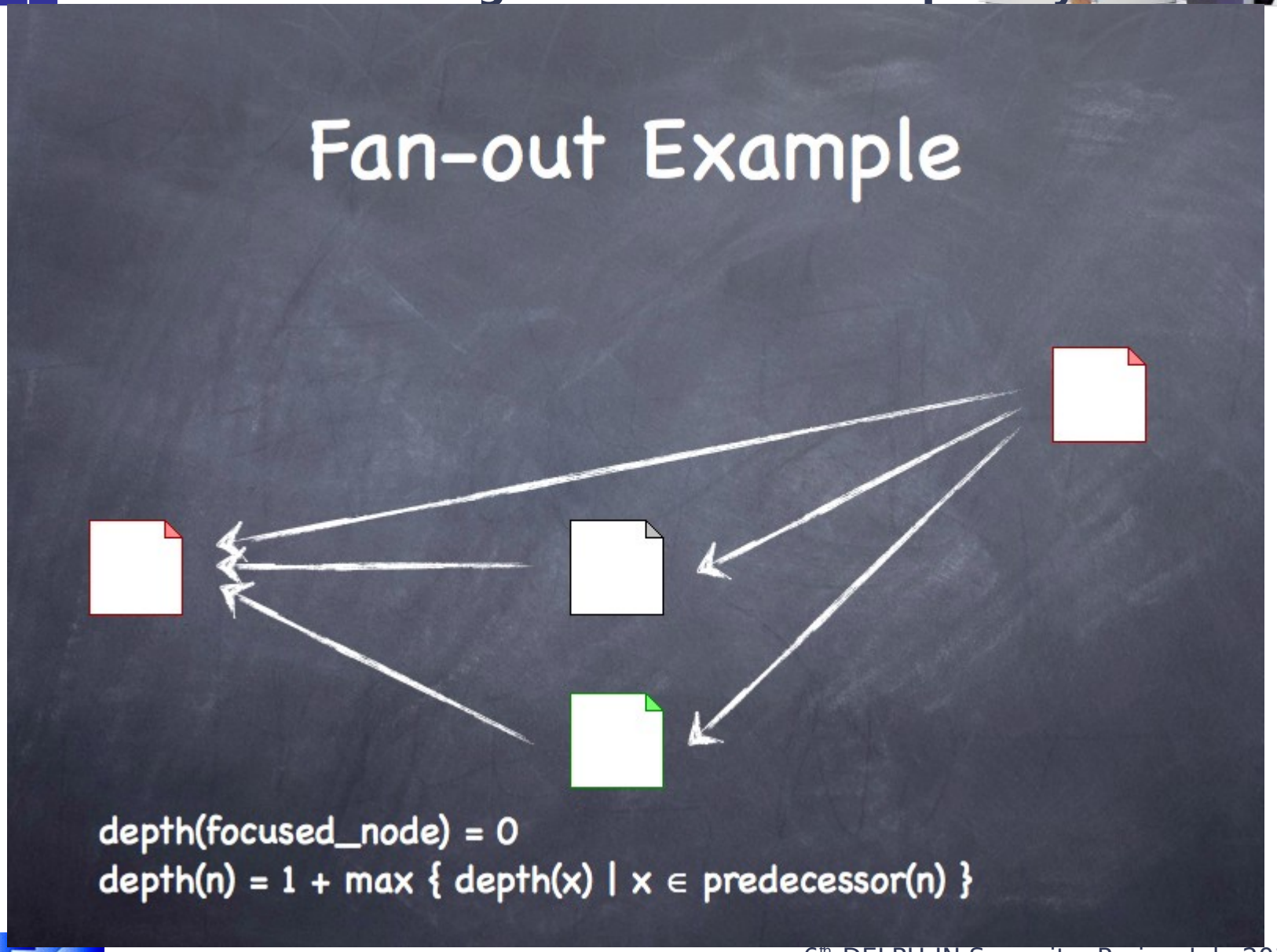
TeeCeeGeeNavigator: Citation Graph Layout



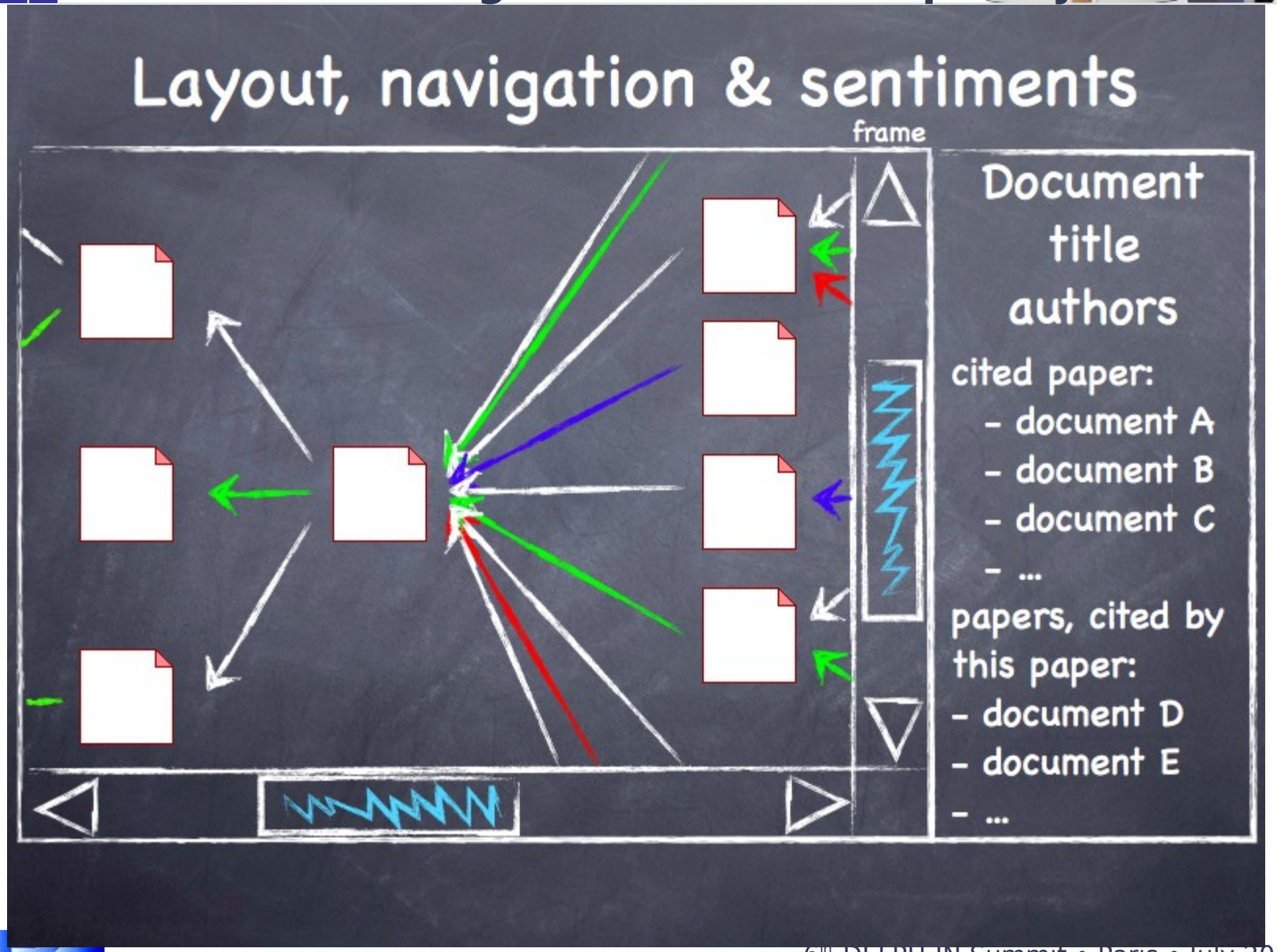
TeeCeeGeeNavigator: Citation Graph Layout

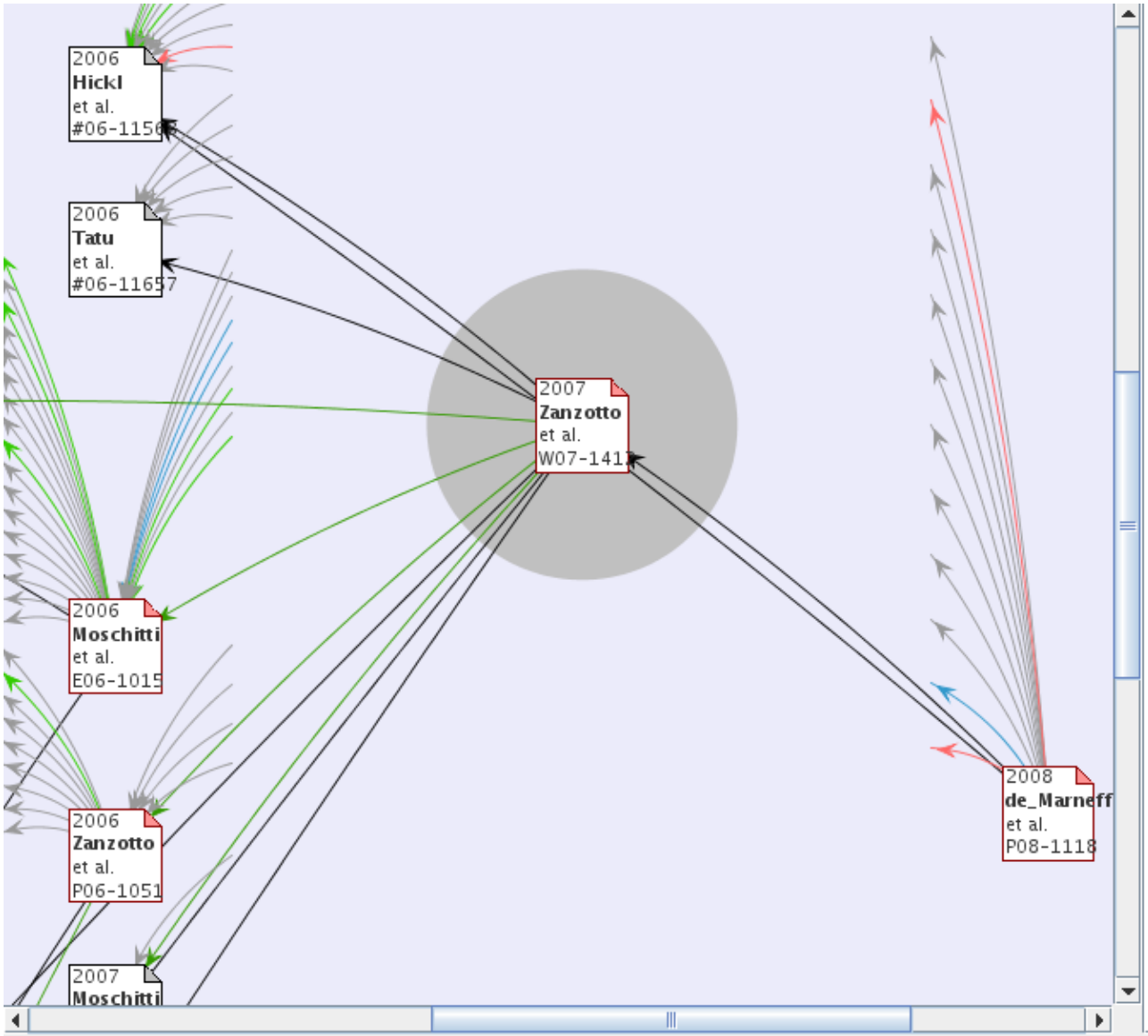


TeeCeeGeeNavigator: Citation Graph Layout



TeeCeeGeeNavigator: Citation Graph Layout





Shallow Semantic in Fast Textual Entailment Rule Learners

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

Cited papers:

#06-3178 [Roy Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, Idan Szpektor]

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

#06-11568 [Andrew Hickl, Jeremy Bensley, John Williams, Kirk Roberts, Bryan Rink, Ying Shi]

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

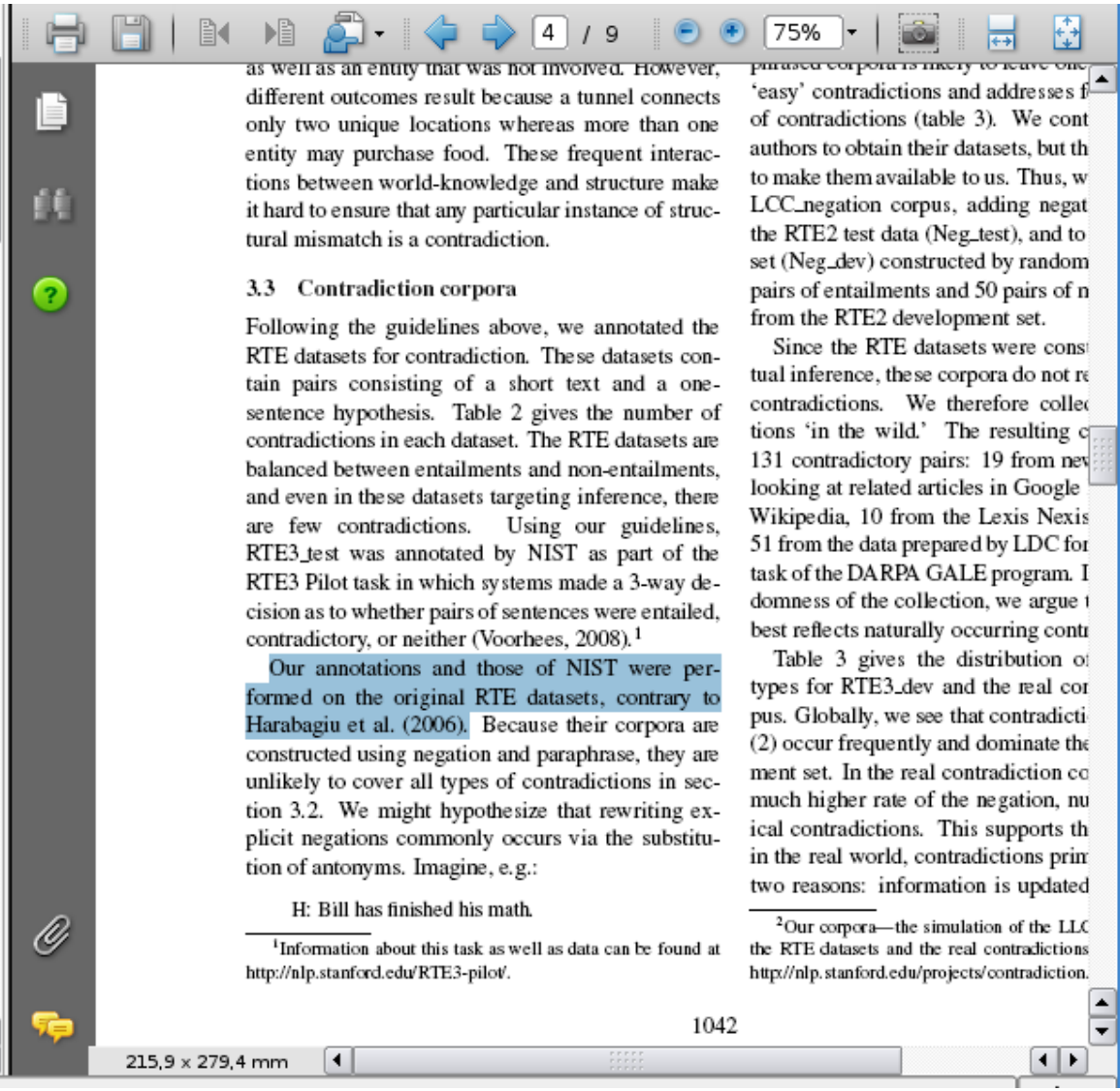
TAKE Citation Sentence Context and PDF Highlighting

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 ofthe 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 (AAAAI-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 ofsentences 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. '



Future Work

- ☆ 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

...