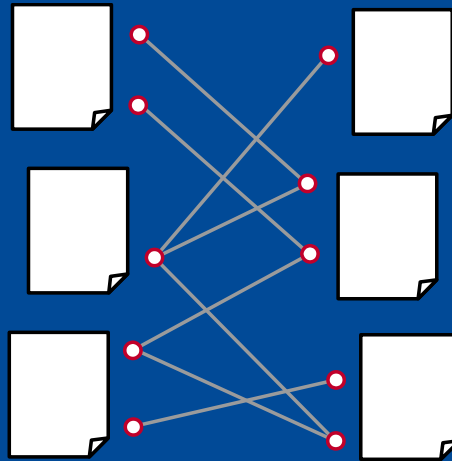


Semantic Approaches to Citation Recommendation



Albert-Ludwigs-Universität Freiburg

Tarek Saier
Master's Thesis

Examiners: Prof. Dr. Georg Lausen
Prof. Dr. Christian Schindelhauer



**UNI
FREIBURG**

Overview



- Background
- Data Set
- Approaches
 - Entity based
 - Claim based
- Evaluation
 - Offline
 - User study
- Discussion





Background



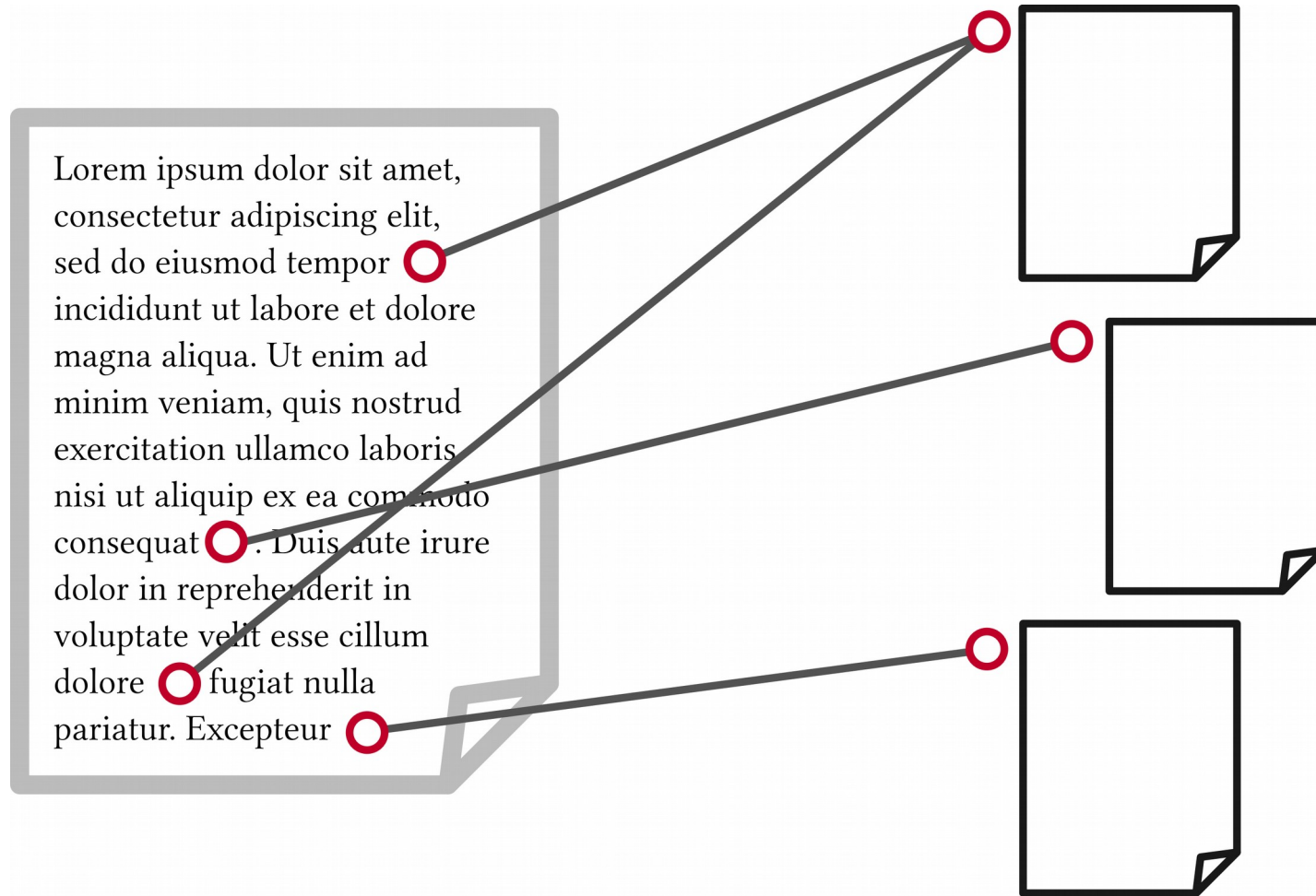
Lorem ipsum dolor sit amet,
consectetur adipiscing elit,
sed do eiusmod tempor
incididunt ut labore et dolore
magna aliqua. Ut enim ad
minim veniam, quis nostrud
exercitation ullamco laboris
nisi ut aliquip ex ea commodo
consequat. Duis aute irure
dolor in reprehenderit in
voluptate velit esse cillum
dolore fugiat nulla
pariatur. Excepteur

Background



Lorem ipsum dolor sit amet,
consectetur adipiscing elit,
sed do eiusmod tempor 
incididunt ut labore et dolore
magna aliqua. Ut enim ad
minim veniam, quis nostrud
exercitation ullamco laboris
nisi ut aliquip ex ea commodo
consequat . Duis aute irure
dolor in reprehenderit in
voluptate velit esse cillum
dolore  fugiat nulla
pariatur. Excepteur 

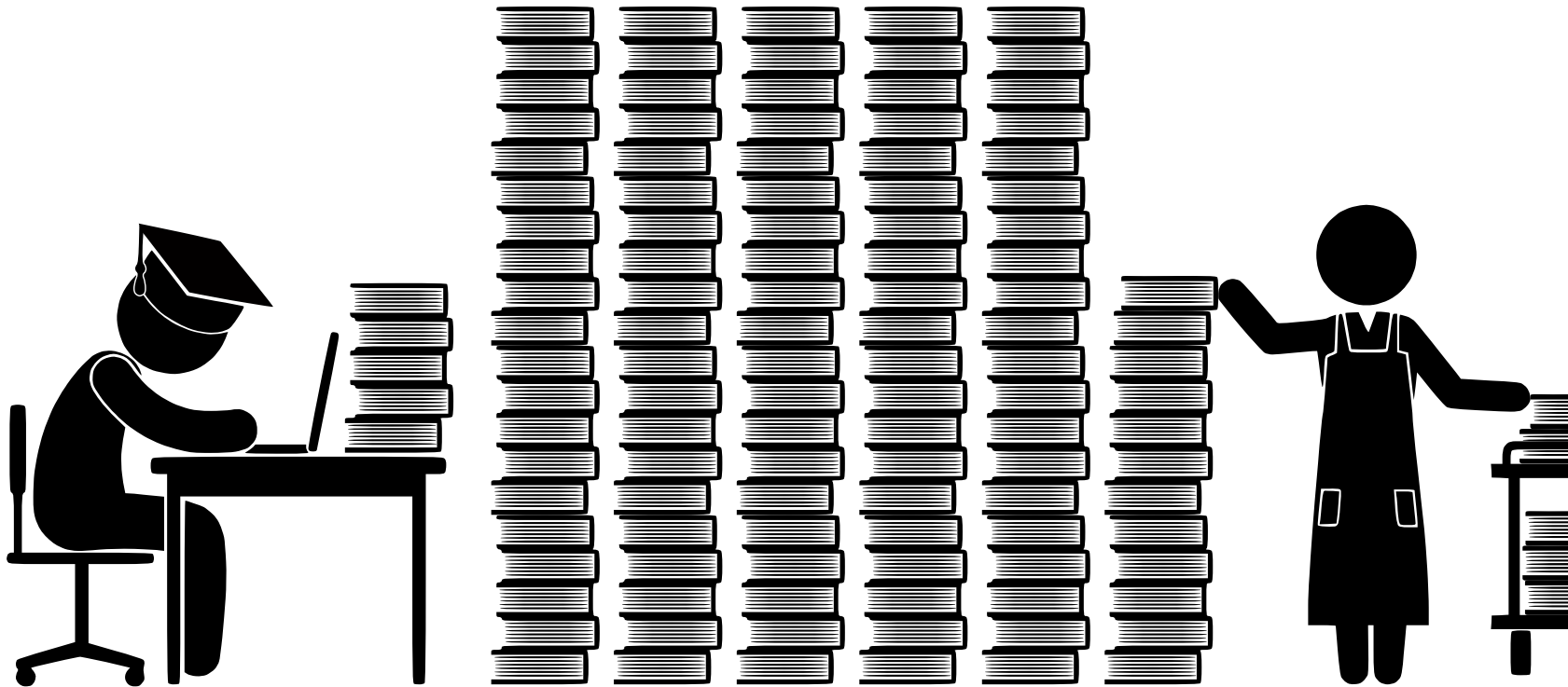
Background



Background



Background



Background



- Automatic processing
 - Processing of documents *as is*
 - Semantic modelling of documents

- Automatic processing
 - Processing of documents *as is*
 - Semantic modelling of documents
- Several approaches
 - Development of ontologies (Peroni et al., 2012)
 - Document exploration (Berger et al., 2016)
 - Recommendation for reading (Beel et al., 2016)
 - Redommendation for citing
 - Global (Galke et al., 2018)
 - Local co-citation (Kobayashi et al., 2018)
 - Local (Ebesu et al., 2017)

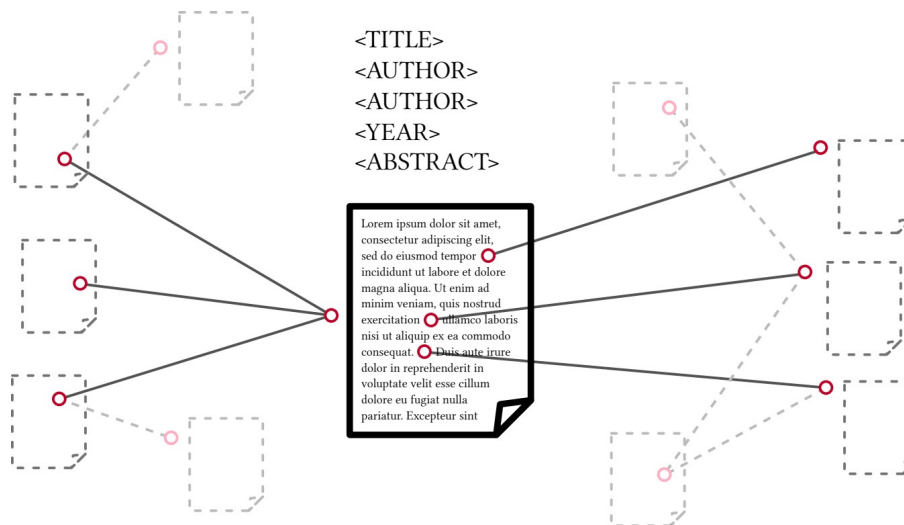
- Automatic processing
 - Processing of documents *as is*
 - Semantic modelling of documents
- Several approaches
 - Development of ontologies (Peroni et al., 2012)
 - Document exploration (Berger et al., 2016)
 - Recommendation for reading (Beel et al., 2016)
 - Redommendation for citing
 - Global (Galke et al., 2018)
 - Local co-citation (Kobayashi et al., 2018)
 - Local (Ebesu et al., 2017) **this, and also semantic**

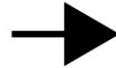
Background



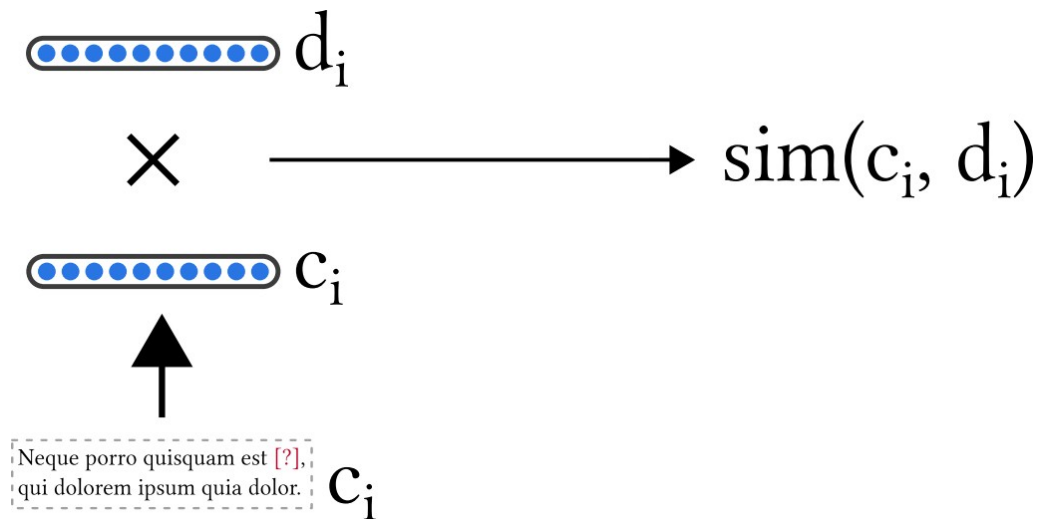
■ Redommendation for citing

- | | | | |
|---------------------|---------------|---|----------|
| - Global | paper | → | paper(s) |
| - Local co-citation | sentence+cit. | → | paper(s) |
| - Local | sentence | → | paper(s) |





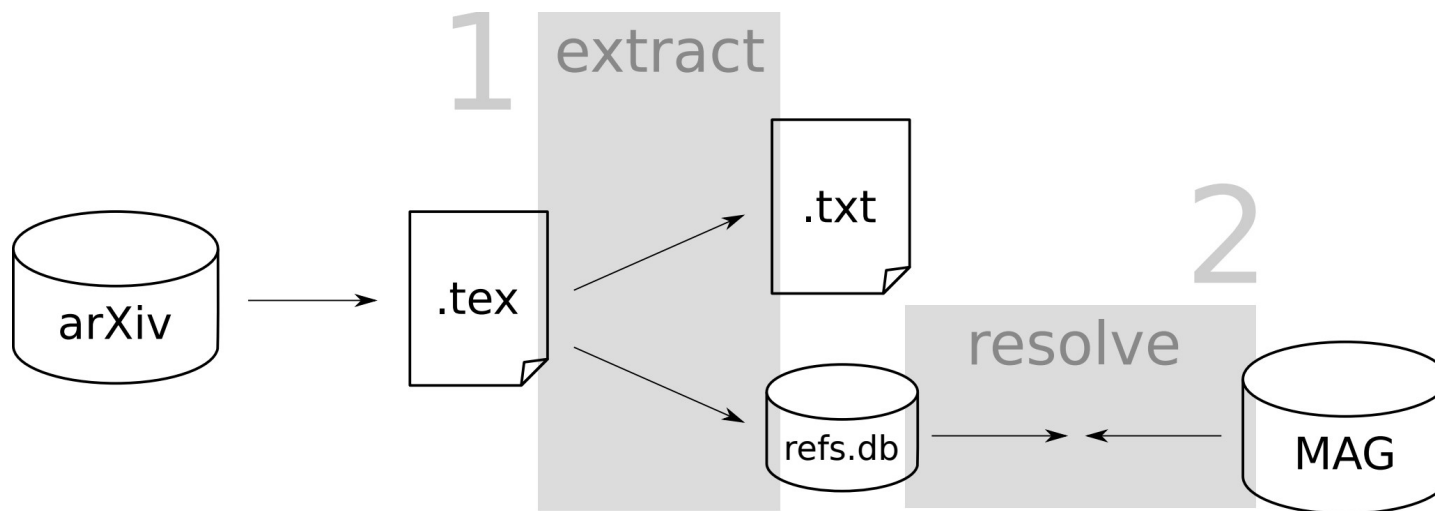
Background



- Existing data sets
 - Quality issues
 - No precise citation information (marker)
 - No citation interlinking (reference resolution)

- Create new data set

- Data sources
 - arXiv.org (LaTeX sources)
 - Microsoft Academic Graph (large)



- arXiv data set
 - large
 - 2.3M cited papers
 - 0.9M citing papers
 - 13M references
 - 25M citation contexts
 - accurate citation markers, interlinking
 - spanning multiple disciplines
 - flexible data format

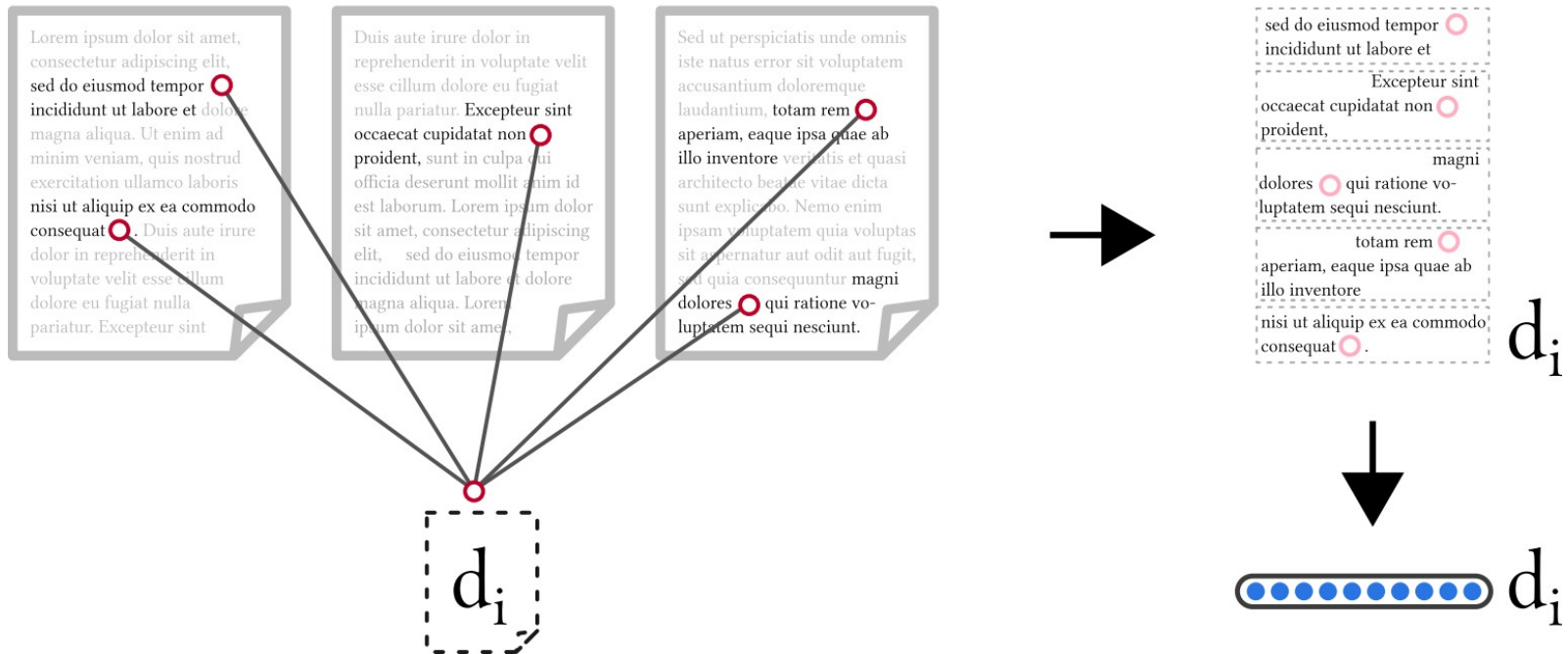
- Quality of reference resolution
 - Manually check 300
 - 3 errors → accuracy estimate: $\geq 96\%$

#		Document
1	matched	<i>"The Maunder Minimum"</i> (John A. Eddy; 1976)
	correct	<i>"The Maunder Minimum: A reappraisal"</i> (John A. Eddy; 1983)
2	matched	<i>"Support Vector Machines"</i> (Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani; 2013)
	correct	<i>"1-norm Support Vector Machines"</i> (Ji Zhu, Saharon Rosset, Robert Tibshirani, Trevor J. Hastie; 2003)
3	matched	<i>"The Putative Liquid-Liquid Transition is a Liquid-Solid Transition in Atomistic Models of Water"</i> (David Chandler, David Limmer; 2013)
	correct	<i>"The putative liquid-liquid transition is a liquid-solid transition in atomistic models of water. II"</i> (David T. Limmer, David Chandler; 2011)

Approaches



■ Semantic modelling of citation contexts



Approaches



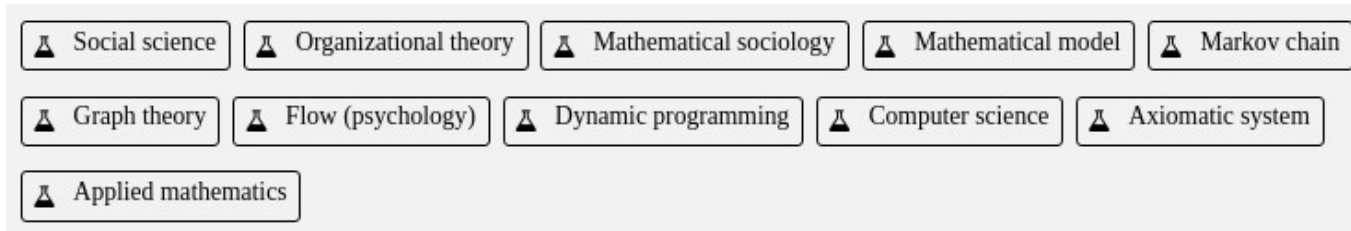
- Entities
 - Reference publications
 - Exemplifications
- Claims
 - Claims backed by citations

- Entities
 - “CiteSeer^x [18]” / “Neural ParsCit [53]”
 - “... approaches to citation recommendation [19–26]”
- Claims
 - “It has been shown, that ... [27].”
 - “A common argument for X is, that ... [3-7].”

Entity Based Approach



- Fields of Study in the MAG (230k)



- Noun phrases (2.8M)

- “*example*”
- “noun *phrase*”
- “context-based co-citation *recommendation*”

Entity Based Approach



- NP model

“We implement our M-CNN in the Caffe framework [1], with the proposed label prediction step as a new layer.”

- NPmarker model

“We implement our M-CNN in the Caffe framework [1], with the proposed label prediction step as a new layer.”

Claim Based Approach



- Identify claims with PredPatt
- Traverse parse trees
- Build predicate-argument tuples

Claim Based Approach



- Claim model
 - “The paper shows that context-based methods can outperform global approaches.”

Claim Based Approach



- Claim model

“The paper shows that context-based methods can outperform global approaches.”

?a shows ?b

?a : The paper

?b : SOMETHING := context-based methods
can outperform global approaches

?a can outperform ?b

?a : context - based methods

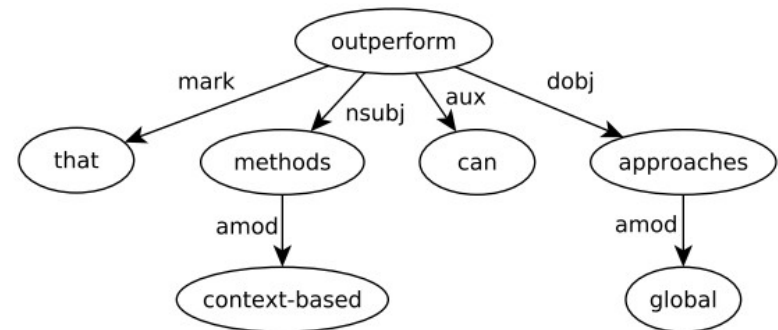
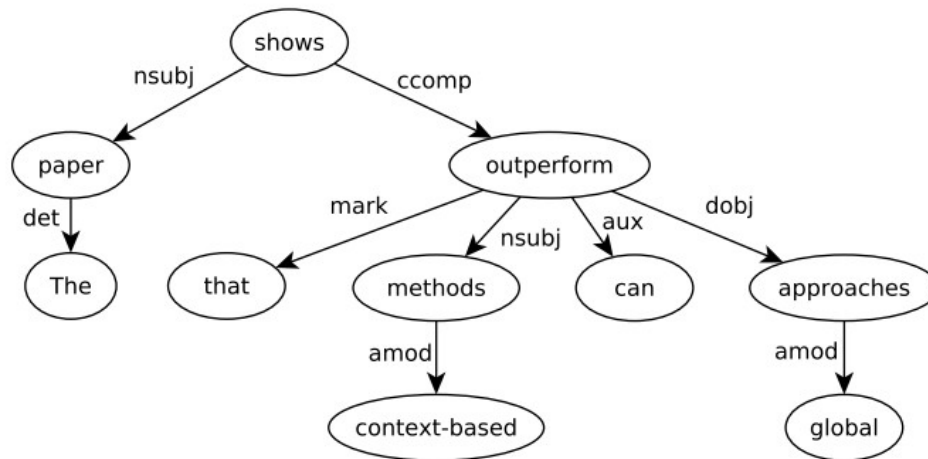
?b : global approaches

Claim Based Approach



- Claim model

“The paper shows that context-based methods can outperform global approaches.”

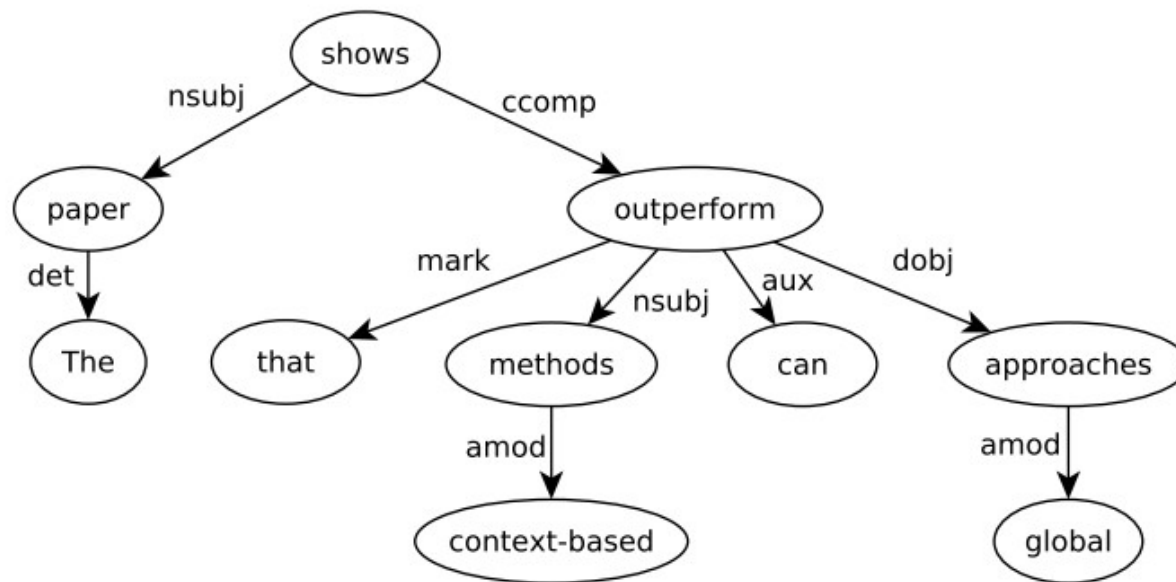


Claim Based Approach



- Claim model

“The **paper** **shows** that **context-based methods** can outperform **global approaches**.”

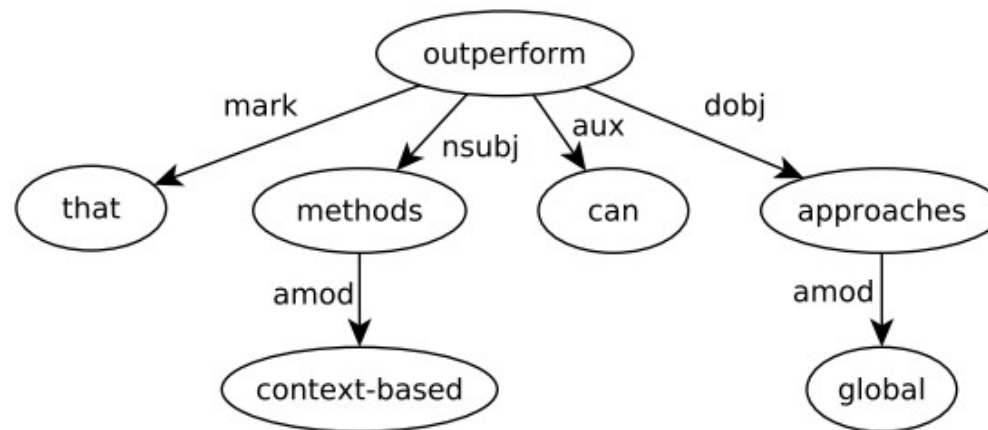


Claim Based Approach



- Claim model

“The paper shows that **context-based methods** can **outperform** **global approaches**.”



Claim Based Approach



- Claim model

“The paper shows that context-based methods can outperform global approaches.”

show:paper

show:context based methods

show:global approaches

outperform:context based methods

outperform:global approaches

→ sequential invariance

Approaches - Recommendation



- Similarity measure
 - Entities: cosine similarity
 - Claims: cosine similarity of TFIDF weighted vectors

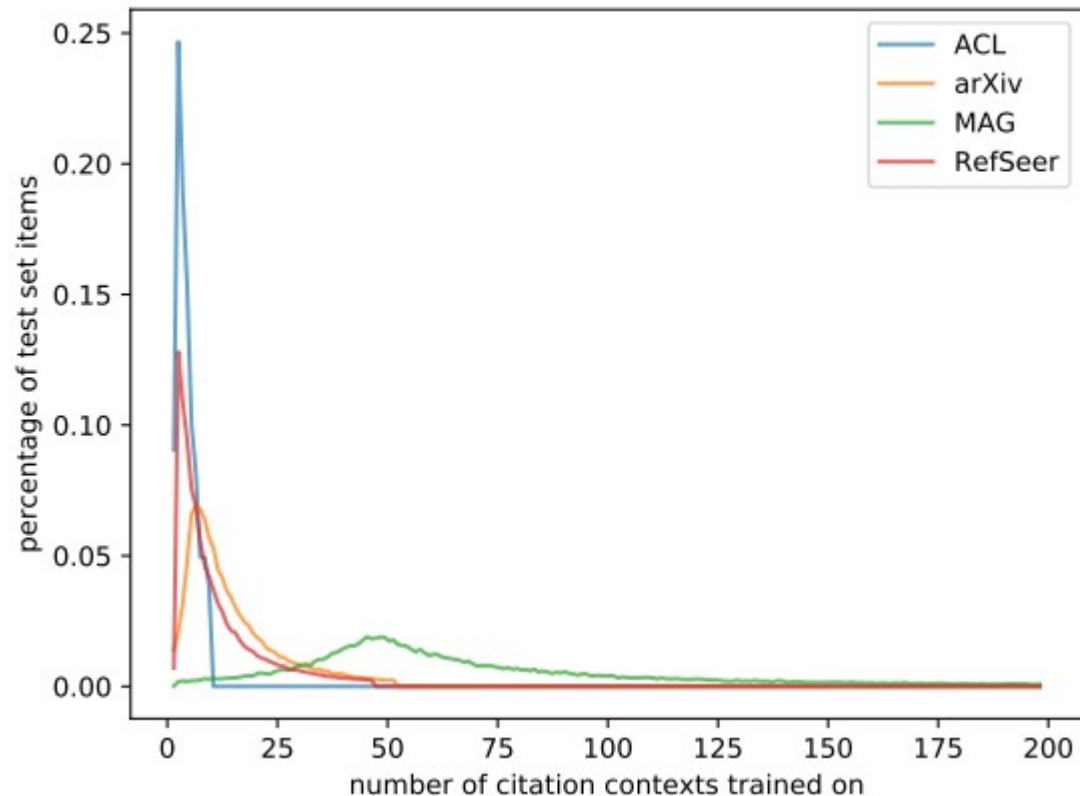
- In both cases
 - Candidates: aggregated contexts of cited docs
 - Input: single citation context

- Offline evaluation
 - Large scale
 - Limited assessment of relevance
- User study
 - Thorough assessment of relevance
 - Limited in scale

Offline Evaluation



- Data sets
 - arXiv
 - MAG
 - RefSeer
 - ACL-ARC

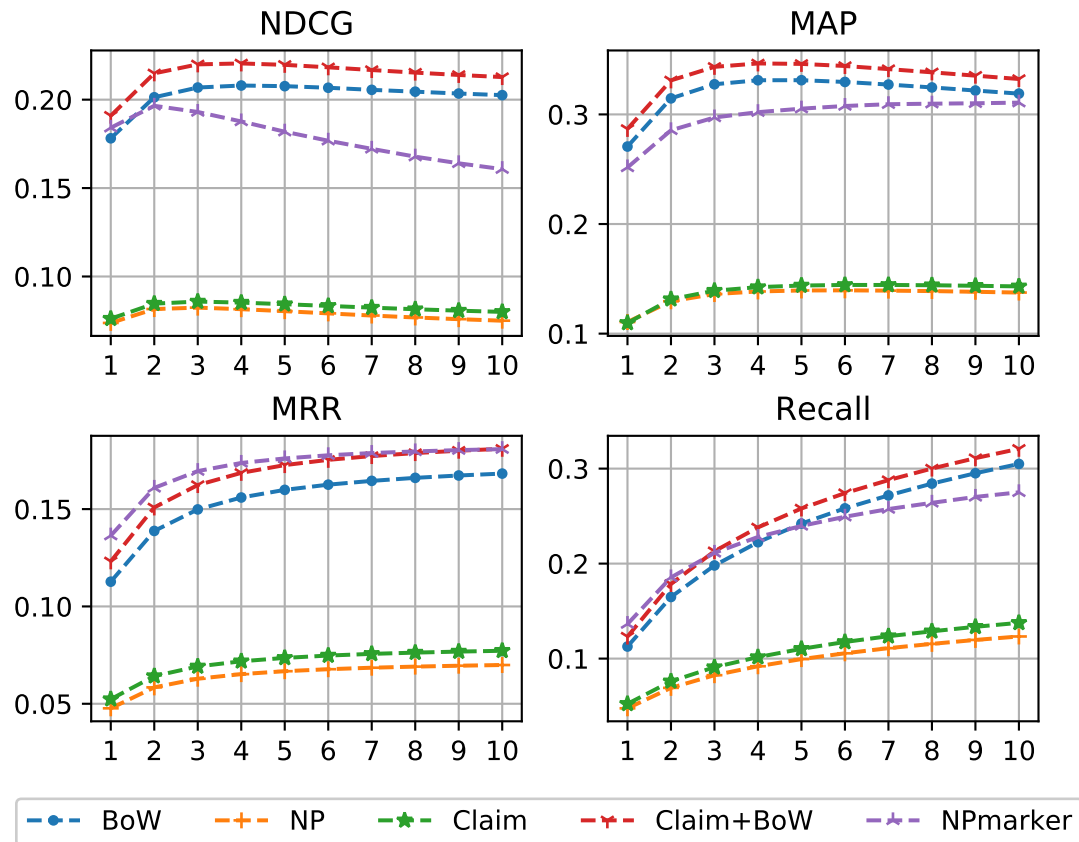


- Models
 - Bag-of-Words baseline
(punctuation, stop words, TFIDF)
 - NP
 - NPmarker
 - Claim
 - Claim+BoW
(combination of similarity scores)

Offline Evaluation



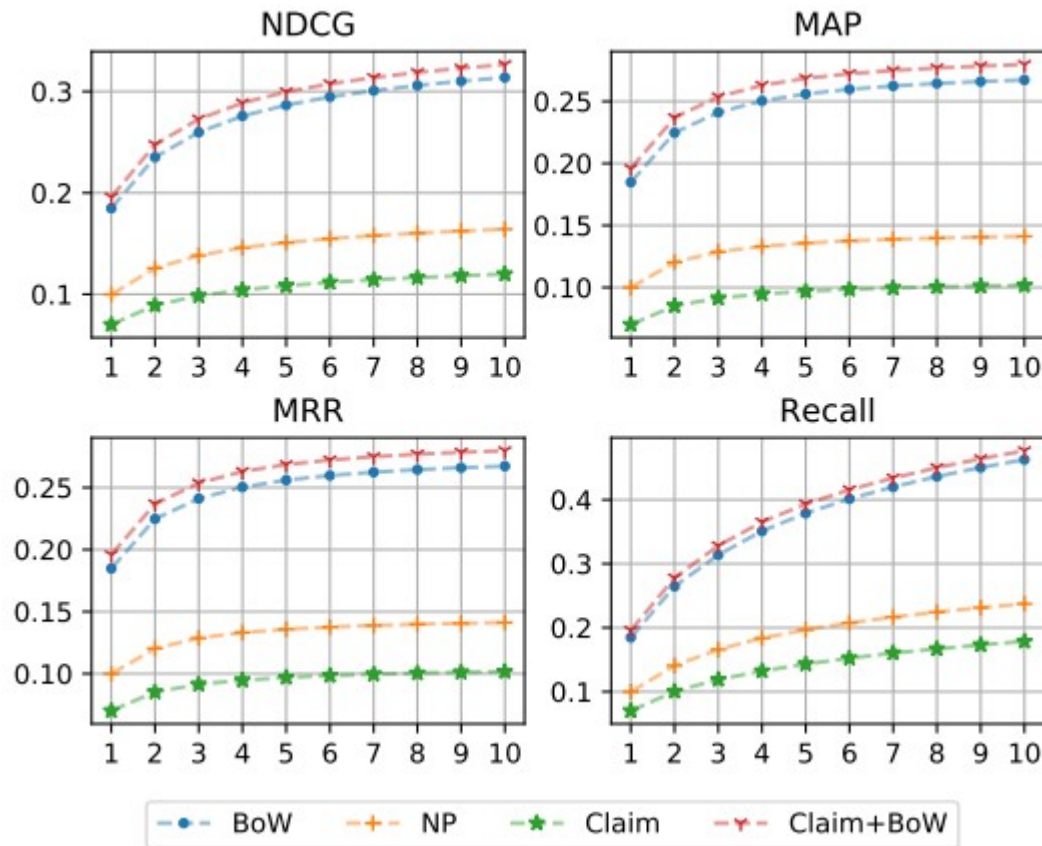
■ arXiv data



Offline Evaluation



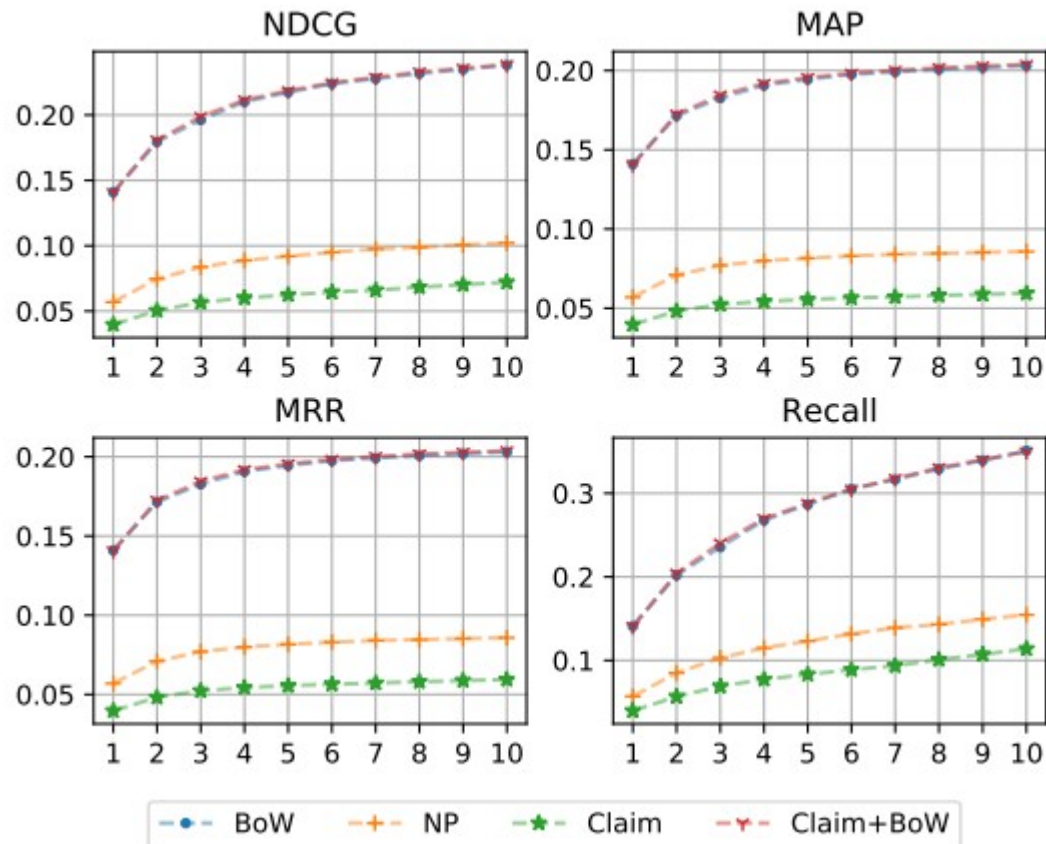
- MAG data



Offline Evaluation



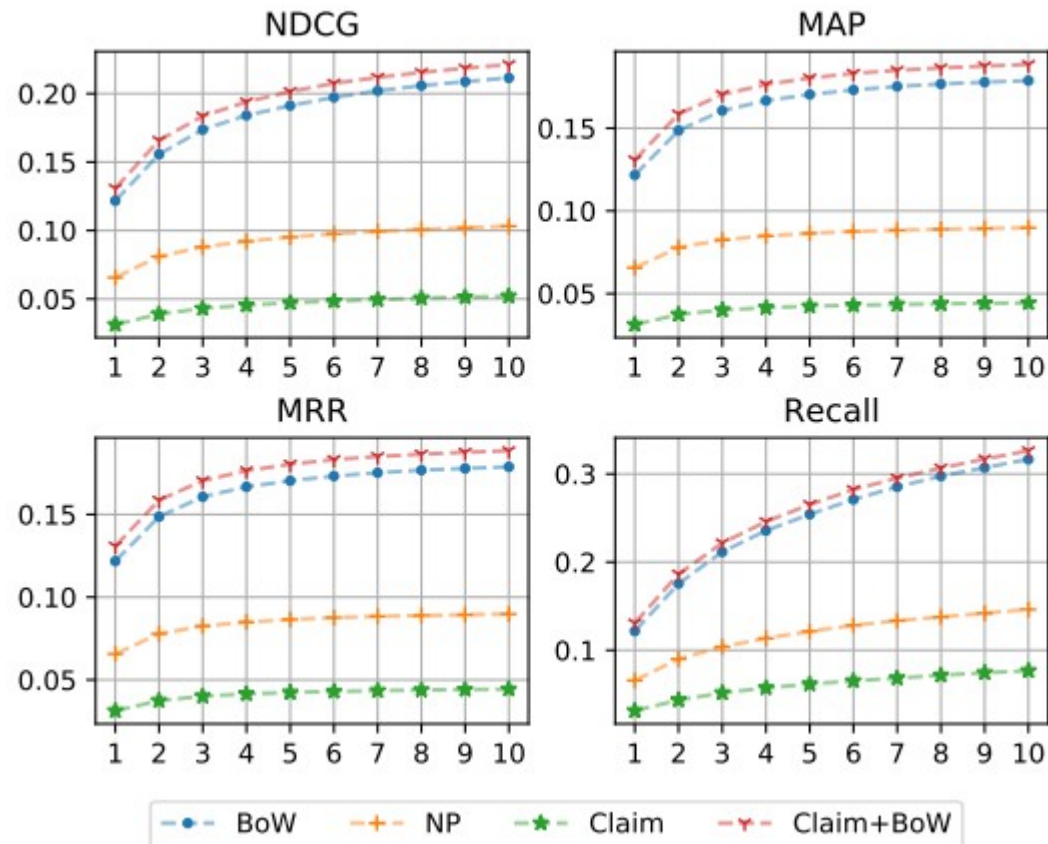
- ACL data



Offline Evaluation



- RefSeer data



- Motivation
 - Insight into offline evaluation data
 - Human judgement of recommendations
- Setting
 - 2 raters
 - 100 citation contexts
 - Top 5 recommendations of models
BoW, Claim+BoW, NPmarker
 - Citation characteristics (type / author's name / ...)

User Study



“To get an idea of the state space, it is not hard to see that there are FORMULA ways to partition and order FORMULA where FORMULA is the number of possible ways to divide a set of FORMULA objects into FORMULA partitions, otherwise known as Stirling numbers of second kind **MAINCIT** .”

not enough information / pass (I can't judge the relevance)

☐ author name inc. ☐ marker has gramm. func. | citation type: NE/concept ▼

check all relevant:

model 1

1. ☒ [Concrete Mathematics: A Foundation for Computer Science](#)
2. ☐ [Deciding DPDA Equivalence Is Primitive Recursive](#)
3. ☒ [Introductory Combinatorics](#)
4. ☐ [Asymptotic estimates of Stirling numbers](#)
5. ☐ [A Bayesian View of the Poisson-Dirichlet Process](#)

model 2

1. ☒ [Introductory Combinatorics](#)
2. ☒ [Concrete Mathematics: A Foundation for Computer Science](#)
3. ☐ [Deciding DPDA Equivalence Is Primitive Recursive](#)
4. ☐ [Asymptotic estimates of Stirling numbers](#)
5. ☐ [A Bayesian View of the Poisson-Dirichlet Process](#)

model 3

1. ☒ [Introductory Combinatorics](#)
2. ☒ [A Course in Combinatorics](#)
3. ☐ [On the Product of Independent Complex Gaussians](#)
4. ☐ [Asymptotic estimates of Stirling numbers](#)
5. ☒ [Combinatorics: Topics, Techniques, Algorithms](#)

Rate

- Inter rater agreement
 - 87.3%
- Results
 - Claim+BoW only outperforms BoW in Recall metric
 - NPmarker best for NE/concept type
 - Claim+BoW best for claim type

Discussion



- M

■ Format

1412.3684.txt

[...] It has over 79 million images stored at the resolution of FORMULA . Each image is labeled with one of the 75,062 non-abstract nouns in English, as listed in the Wordnet{{cite:9ad20b7d-87d1-47f5-aeed-10a1cf89a2e2}} {{cite:298db7f5-9ebb-4e98-9ecf-0bdda28a42cb}} lexical database. [...]

refs.db

uuid	in_doc	mag_id	reference_string
9ad20b...	1412.3684	2081580037	George A. Miller (1995). Wo...
298db7...	1412.3684	2038721957	Christiane Fellbaum (1998),...

Data Set Format



■ Format

MAG

paperid	originaltitle	publisher	
2081580037	WordNet : an electronic lexical database	MIT Press	
2038721957	WordNet: a lexical database for English	ACM	...

extracted_contexts.csv

2038721957 | 2081580037 | 1412.3684 | It has over 79 million images stored at the resolution of FORMULA . Each image is labeled with one of the 75,062 non-abstract nouns in English, as listed in the Wordnet CIT MAINCIT lexical database. It has been noted that many of the labels are not reliable CIT .

Data Set Citation Flow

