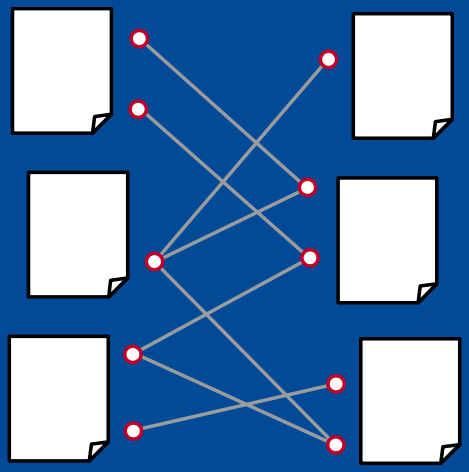
Semantic Approaches to Citation Recommendation



Albert-Ludwigs-Universität Freiburg

Tarek Saier
Master‘s Thesis

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

Citation Reommendation

- Task

"Word embeddings have been studied in information retrieval contexts such as term reweighting [x], cross-lingual retrieval [y] and short-text similarity [z]."

Find a fitting publication for [y].

Citation Reommendation

- Why?
- How?

Citation Reommendation

- Why?
- How?
  - A Data Set
  - Two Models
    - Entity based
    - Claim based
  - Evaluation
  - Discussion

Background

UNI FREIBURG

Why?

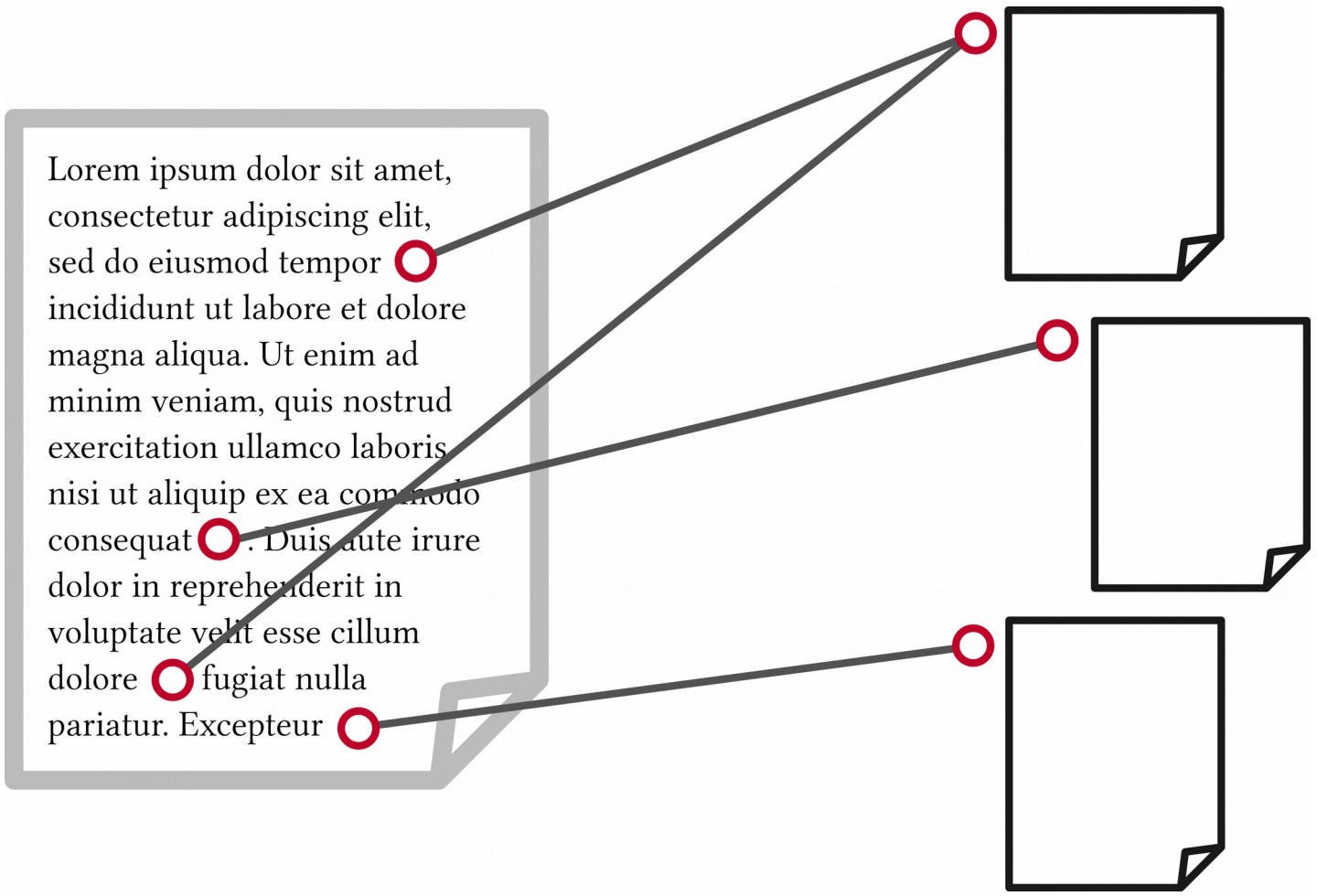
Background

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

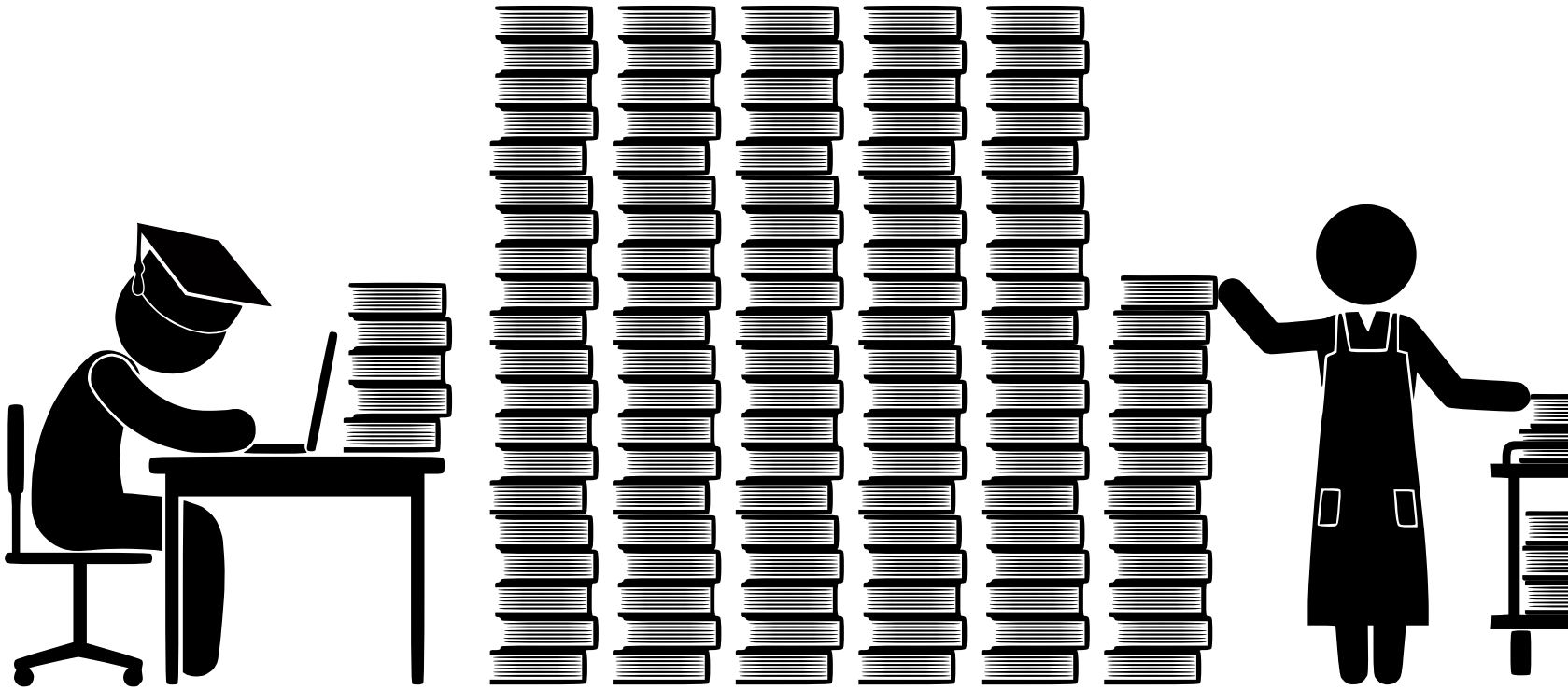
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum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

Background


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

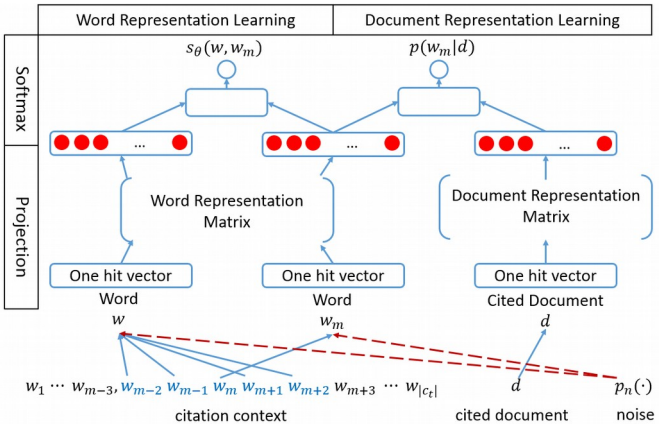


Background

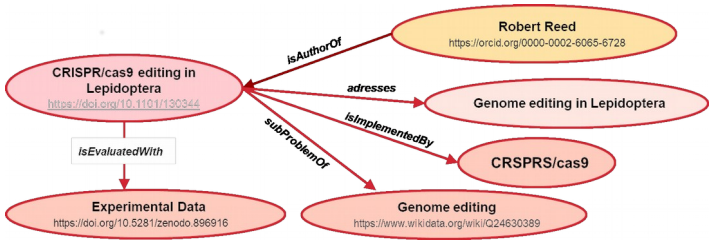
- Automatic processing

Background

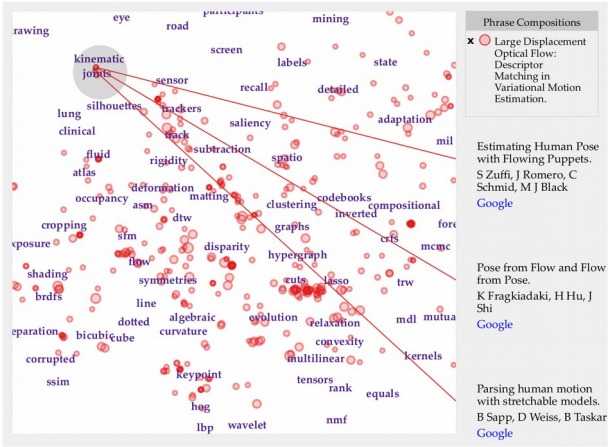
- Automatic processing
- Many approaches



(Huang et al., 2015)



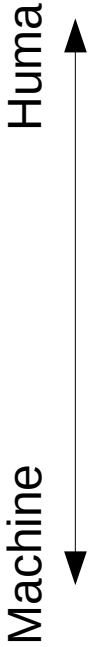
(Jaradeh et al., 2019)



(Berger et al., 2016)

Background

- Automatic processing
- Many 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)



Background

- Automatic processing
- Many approaches

Human ↑ Machine ↓

  - 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

- Task

"Word embeddings have been studied in information retrieval contexts such as term reweighting [x], cross-lingual retrieval [y] and short-text similarity [z]."

Find a fitting publication for [y], using a semantic model of its context.

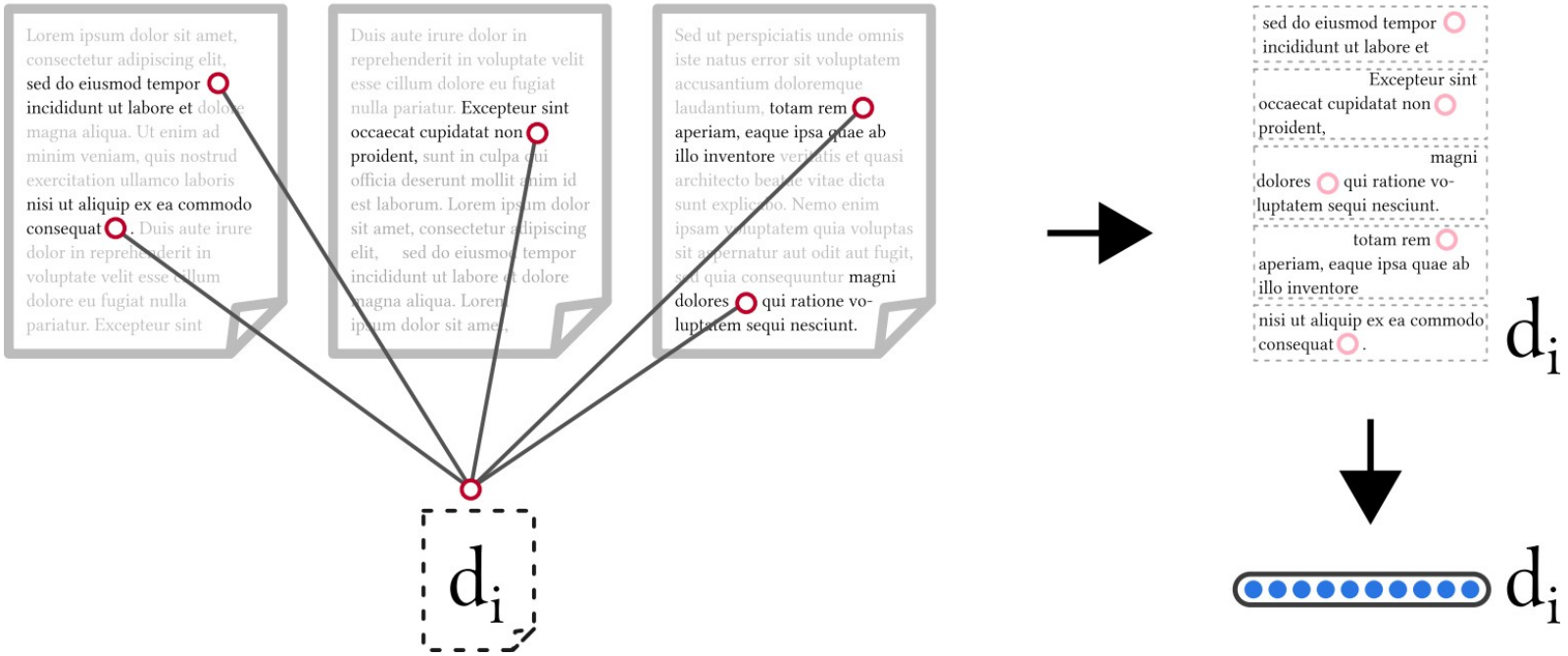
Background

UNI FREIBURG

How?

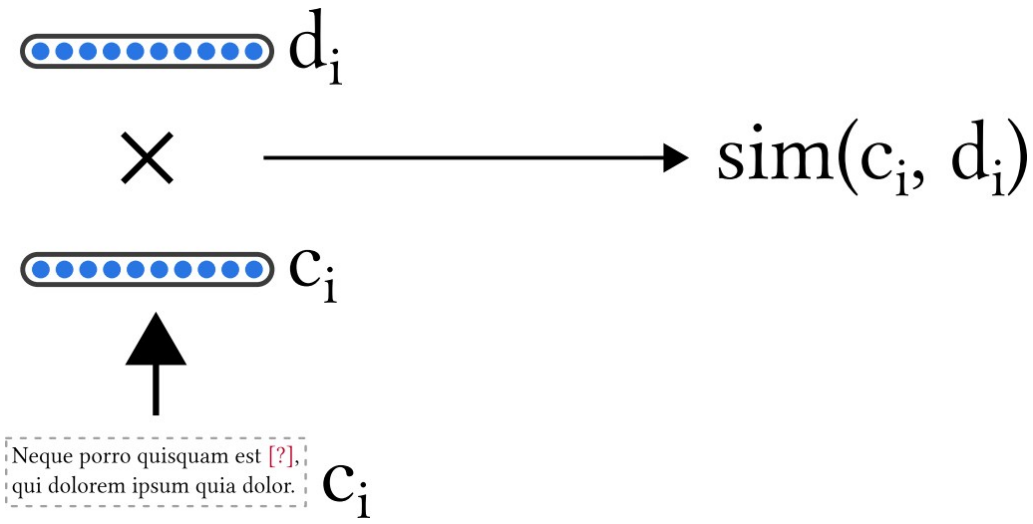
Background

- How is a paper being referred to



Background

- Find document, described similar to input



Data Set

- Existing data sets
  - CiteSeer^{x}
  - PMC-OAS
  - arXiv CS
  - Scholarly
  - ACL-AAN
  - ACL-ARC

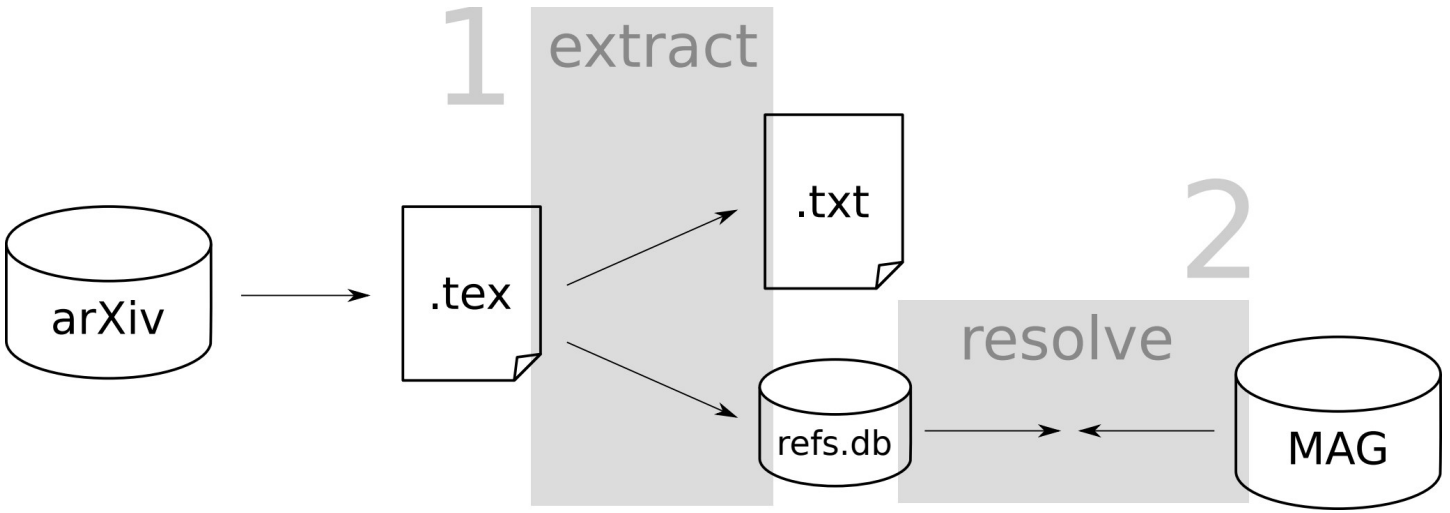
Data Set

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

- Create new data set

Data Set

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



Data Set

- arXiv data set (1991-2017)
  - large
    - 2.3M cited papers
    - 0.9M citing papers
    - 13M references
    - 25M citation contexts
  - accurate citation markers, interlinking
  - spanning multiple disciplines
  - flexible data format

Approaches

- Semantically model citation contexts

Approaches

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

Approaches

- 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

- NP model

  All NPs within a citation context.

- NPmarker model

  NP directly preceding citation marker

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.”

- Similarity: cosine similarity in VSM

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

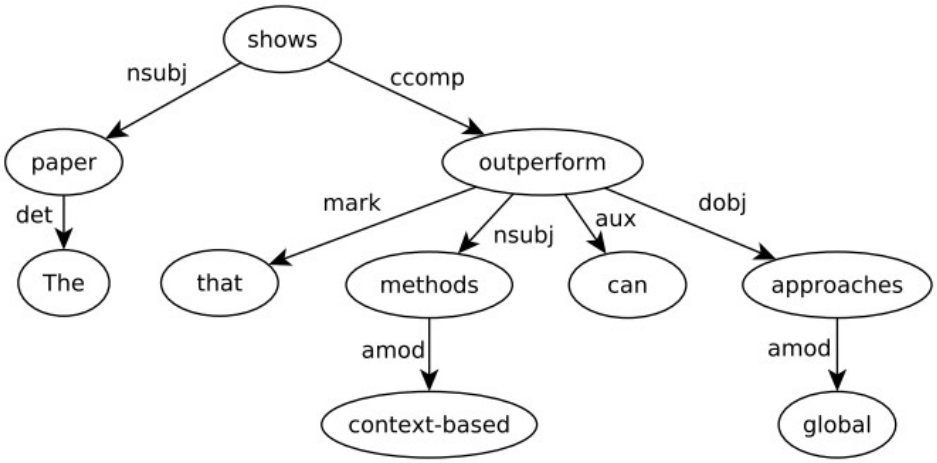
  “The paper shows that context-based methods can outperform global approaches.”

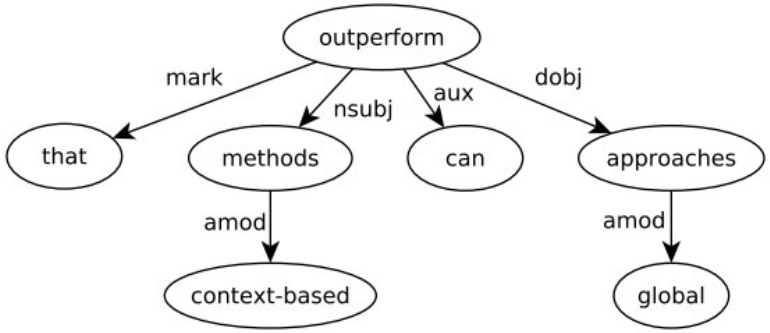
```
?a shows ?b
    ?a : The paper
    ?b : SOMETHING := context-based methods
            can outperform global approaches

?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.”

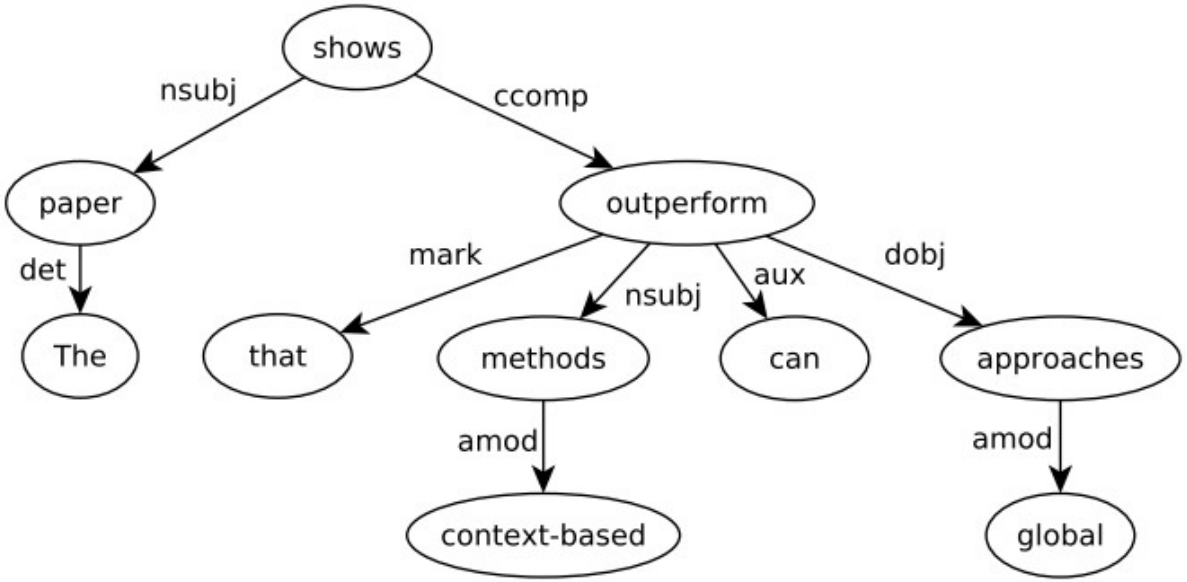




Claim Based Approach

- Claim model

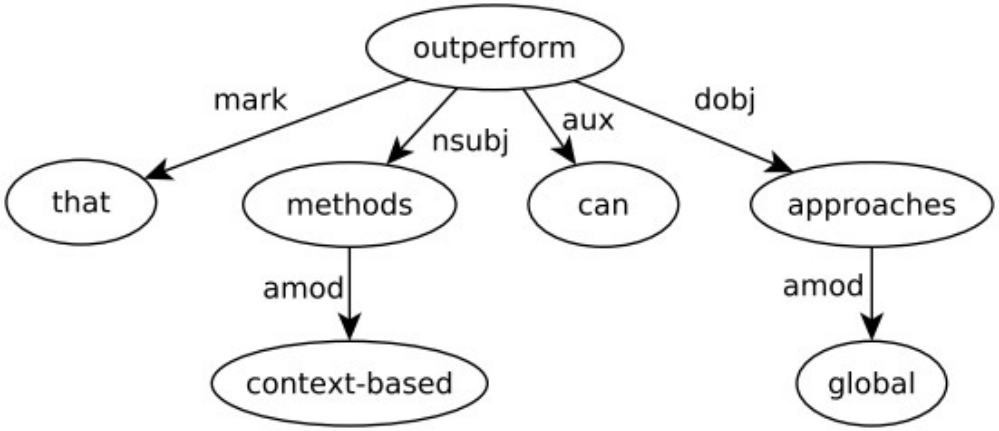
  “The paper **show**s that context-based methods can outperform global approaches.”



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
```

- Similarity: cosine similarity of TFIDF weighted vectors in VSM

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

Evaluation

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

Offline Evaluation

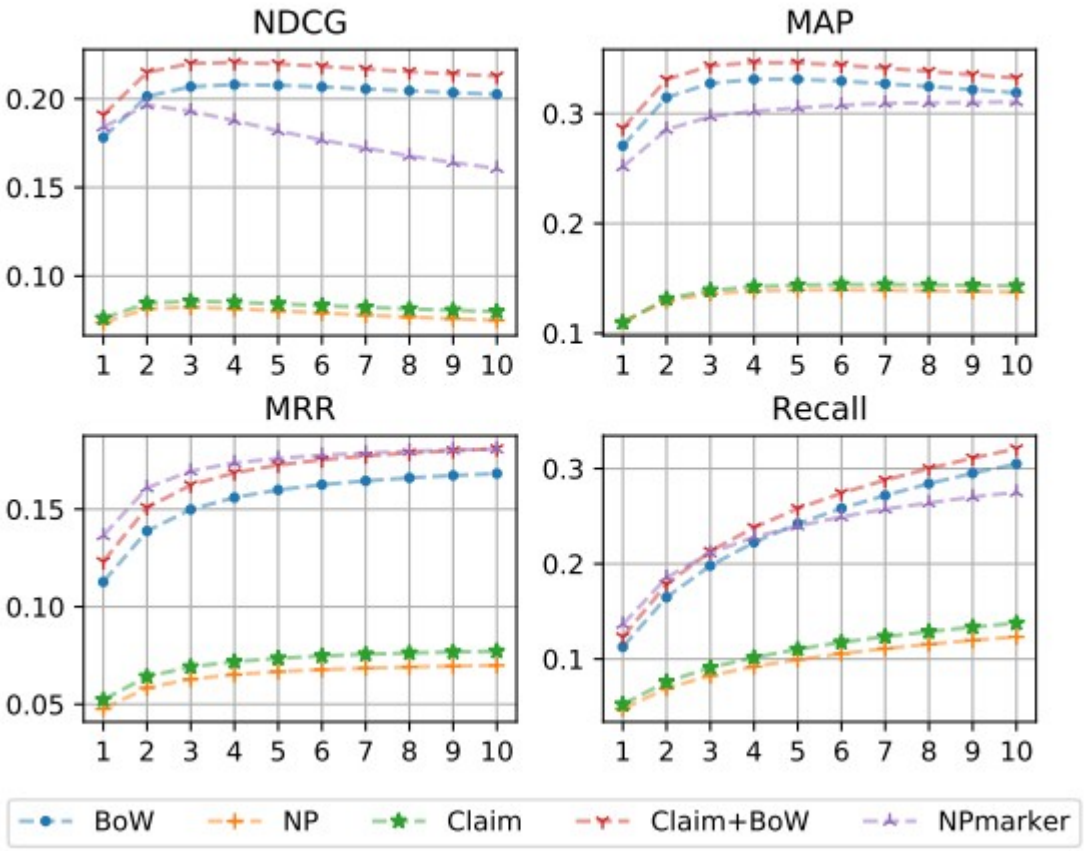
Data sets	#contexts	filter
- arXiv	1.8M	CS
- MAG	8.6M	CS, English
- RefSeer	3.6M	clean
- ACL-ARC	30k	resolved ref.

Offline Evaluation

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

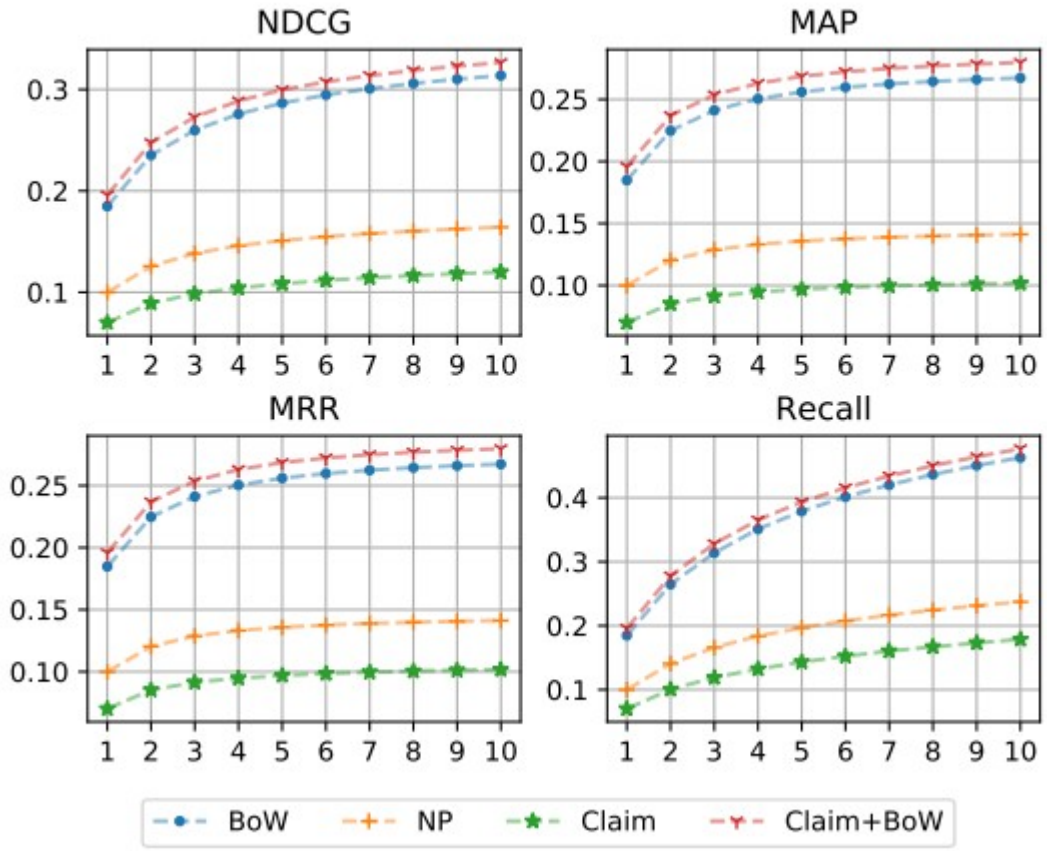
Offline Evaluation

- arXiv data



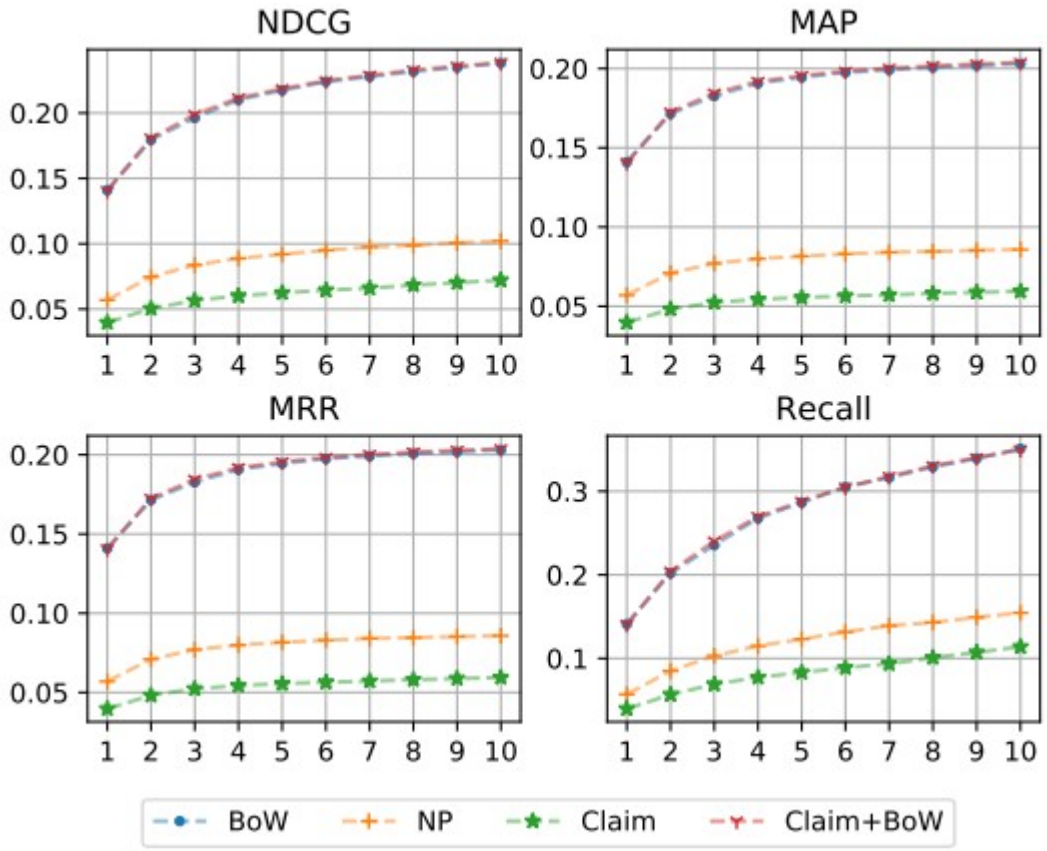
Offline Evaluation

- MAG data



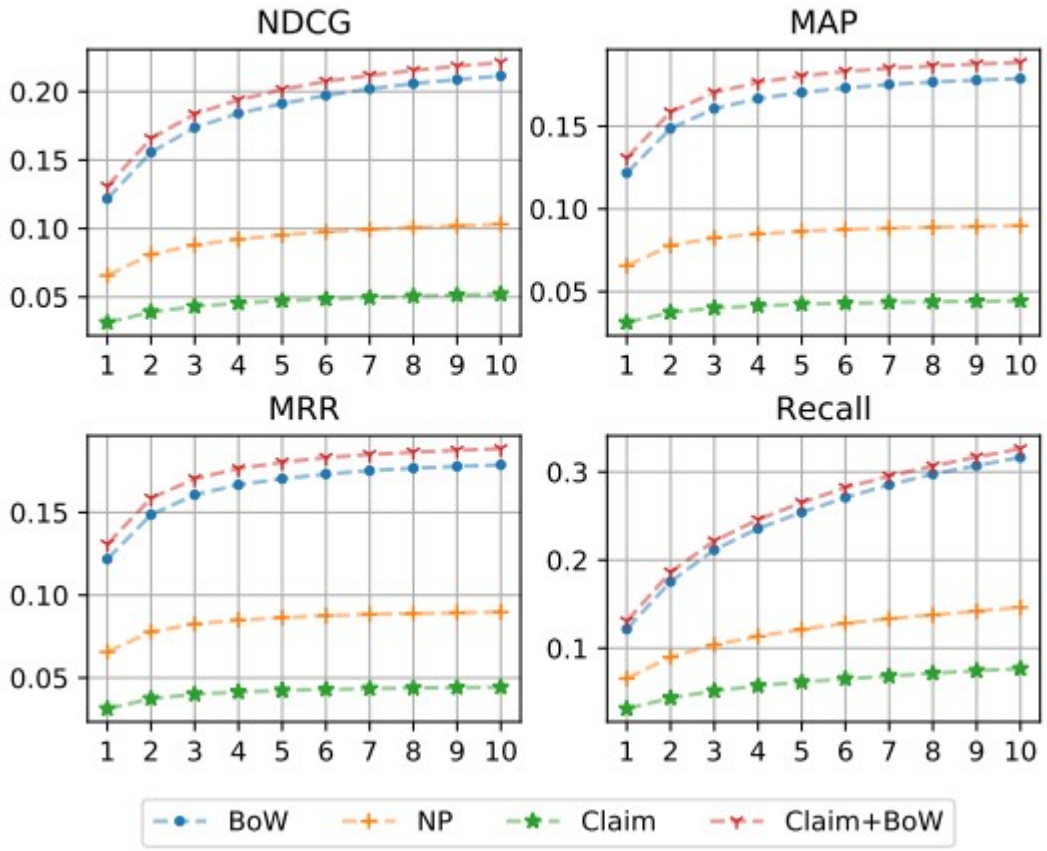
Offline Evaluation

- ACL data



Offline Evaluation

- RefSeer data



Offline Evaluation

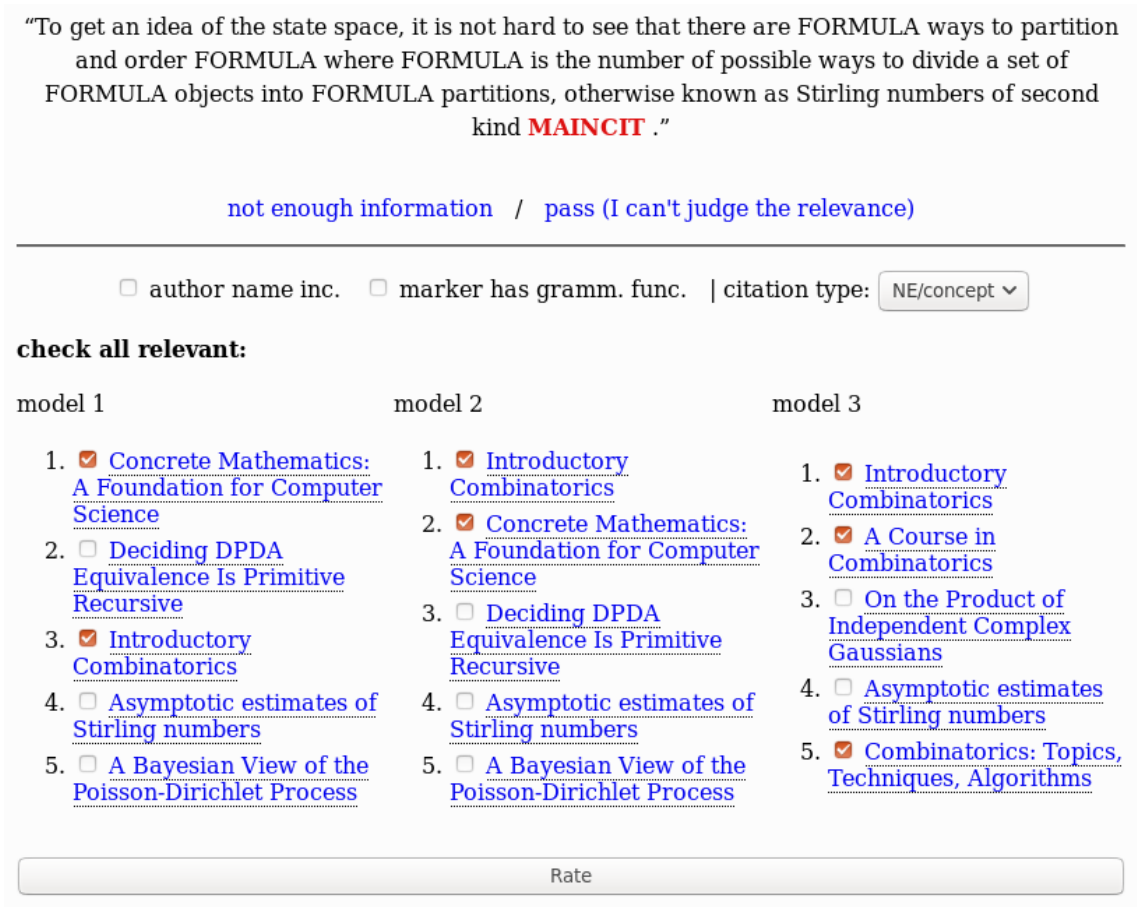
UNI FREIBURG

- Claim+BoW outperforms BoW consistently
- NPmarker outperforms BoW in some settings

User Study

- Setting
  - 2 raters
  - 100 citation contexts
    - Top 5 recommendations of models
      BoW, Claim+BoW, NPmarker
    - Citation types → type specific performance

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

User Study

- Inter rater agreement
  - Overall: 87%
  - Author‘s name: 100%
  - Non-/syntactic: 100%
  - Relevance: 86%
  - Citation type: 78%

		Rater 1			
		Claim	NE/concept	Exemplific.	Other
Rater 2	Claim	12	0	1	0
	NE/concept	1	15	2	0
	Exemplific.	2	1	9	1
	Other	2	0	0	3

User Study

- Results

Model	Recall@5	MRR@5	MAP@5	NDCG@5
all contexts (138)				
Claim+BoW	**0.53**	0.44	0.41	0.46
BoW	0.51	**0.46**	**0.44**	**0.48**
NPmarker	0.35	0.35	0.33	0.34
only contexts of type "claim" (38)				
Claim+BoW	**0.63**	0.46	0.42	0.49
BoW	0.58	**0.48**	**0.46**	**0.51**
NPmarker	0.20	0.13	0.13	0.15
only contexts of type "NE/concept" (45)				
Claim+BoW	0.46	0.44	0.41	0.44
BoW	0.47	0.45	0.41	0.35
NPmarker	**0.52**	**0.53**	**0.48**	**0.51**
only contexts of type "exemplification" (38)				
Claim+BoW	**0.56**	0.52	0.47	0.52
BoW	0.54	**0.53**	**0.49**	**0.54**
NPmarker	0.21	0.24	0.24	0.24
only contexts of type "other" (17)				
Claim+BoW	0.44	0.29	0.29	0.33
BoW	0.41	0.33	0.33	0.36
NPmarker	**0.50**	**0.50**	**0.44**	**0.47**

User Study

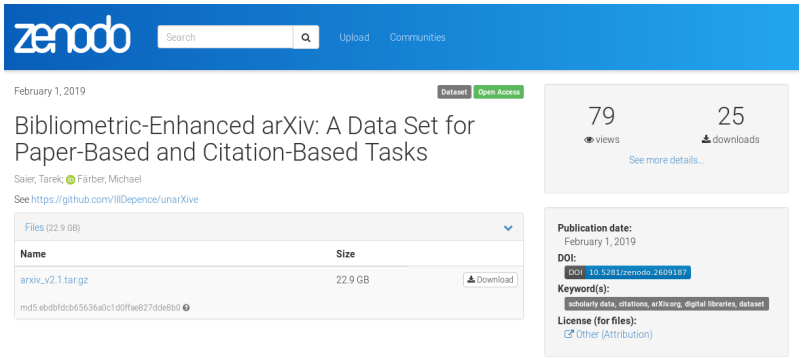
- Performance
  - In general
    - Claim+BoW only outperforms BoW in Recall metric
  - Type specific
    - NPmarker best for NE/concept type
    - Claim+BoW best for claim type

Discussion

- Data Set
- Semantic models
- Semantic citation recommendation

Discussion

- Data Set
  - Exact position of marker → especially useful for semantic models
  - Several disciplines → comparative analysis
  - Flexible format → experiments w/ e.g. context length
  - From hereon
    - Update w/ recent papers (Saier & Färber, 2019)
    - Formulas (Aizawa et al., 2014; Zanibbi et al., 2016)



Discussion

- Semantic models
  - First step towards semantic citation recommendation
  - Citations in general
    - Can enhance recommendation
  - Applied to specific citation types
    - Suitable for conceptualized type
  - From hereon
    - Lemm. predicates → relation types (Gabor et al., 2018)
    - Marker aware claim model (solve non-syntactic citations)
    - Test on more data (NPmarker)

Discussion

- Semantic citation recommendation
  - First specialized models
  - Possible immediate application apart from recomm.

    Semantic search:
    ```
    is:NP-hard
    improve:local citation recommendation
    ```

  - From hereon
    - Thorough & systematic examination of cit. types
    - Specialized models
    - Closer to LD modelling
    - Credibility of claims, argumentative structures, ...

Sources

Peroni, S. & Shotton, D. *FaBiO and CiTO: Ontologies for describing bibliographic resources and citations*, Journal of Web Semantics, 2012, 17, 33 - 43

Berger, M.; McDonough, K. & M. Seversky, L. *Cite2vec: Citation-Driven Document Exploration via Word Embeddings*, IEEE Transactions on Visualization and Computer Graphics, 2016, 23, 1-1

Beel, J.; Langer, S.; Genzmehr, M.; Gipp, B.; Breitinger, C. & Nürnberger, A. *Research Paper Recommender System Evaluation: A Quantitative Literature Survey*, Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, ACM, 2013, 15-22

Galke, L.; Mai, F.; Vagliano, I. & Scherp, A. *Multi-Modal Adversarial Autoencoders for Recommendations of Citations and Subject Labels*, Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, ACM, 2018, 197-205

Kobayashi, Y.; Shimbo, M. & Matsumoto, Y. *Citation Recommendation Using Distributed Representation of Discourse Facets in Scientific Articles*, Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, ACM, 2018, 243-251

Ebesu, T. & Fang, Y. *Neural Citation Network for Context-Aware Citation Recommendation*, Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2017, 1093-1096

Sources

Saier, T. & Färber, M. *Bibliometric-Enhanced arXiv: A Data Set for Paper-Based and Citation Based Tasks*, Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR), 2019, 14-26

Aizawa, A.; Kohlhase, M.; Ounis, I. & Schubotz, M. *NTCIR-11 Math-2 Task Overview*, Proceedings of the 11th NTCIR Conference, 2014, 11, 88-98

Zanibbi, R.; Aizawa, A.; Kohlhase, M.; Ounis, I.; Topic, G. & Davila, K. *NTCIR-12 MathIR Task Overview*, Proceedings of the 12th NTCIR Conference, 2016, 12

Jaradeh, M. Y.; Auer, S.; Prinz, M.; Kovtun, V.; Kismihók, G. & Stocker, M. *Open Research Knowledge Graph: Towards Machine Actionability in Scholarly Communication*, 2019, arXiv:1901.10816

Huang, W.; Wu, Z.; Liang, C.; Mitra, P. & Giles, C. L. A *Neural Probabilistic Model for Context Based Citation Recommendation*, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI Press, 2015, 2404-2410

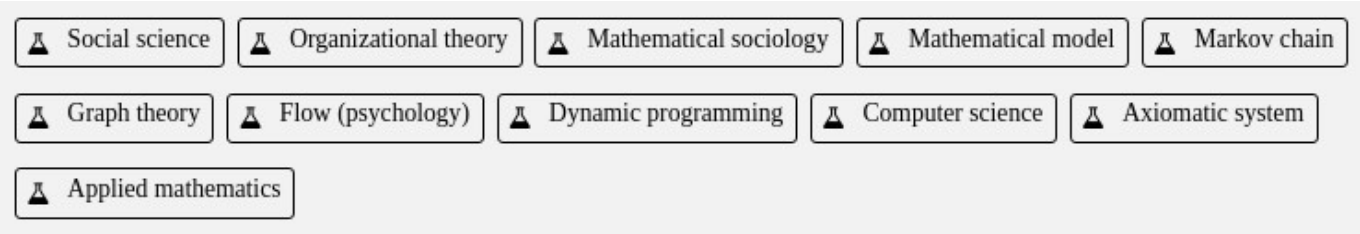
Gábor, K.; Buscaldi, D.; Schumann, A.-K.; QasemiZadeh, B.; Zargayouna, H. & Charnois, T. *SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers*, Proceedings of The 12th International Workshop on Semantic Evaluation, 2018, 679-688

UNI FREIBURG

Thank you.

Entity Based Approaches

- Fields of Study in the MAG (230k)



- Noun phrases (2.8M)
  - “*example*”
  - “noun *phrase*”
  - “context-based co-citation *recommendation*”

Entity Based Approaches

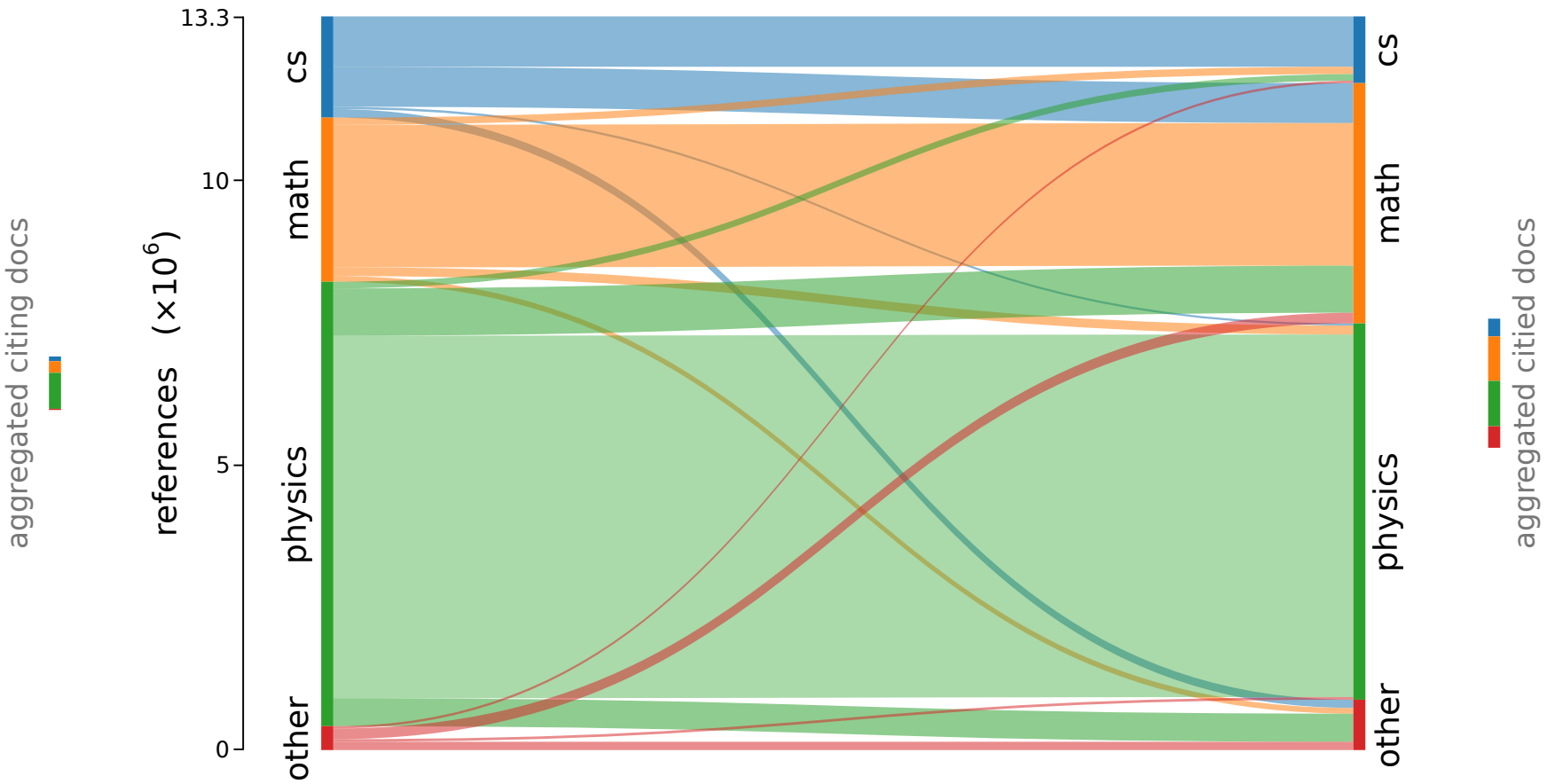
- Fields of Study in the MAG (230k)
  - parent/child structure
    - computer science$_{(0)}$
      information retrieval$_{(1)}$
      search engine$_{(2)}$
      web search query$_{(3)}$
      ranking (information retrieval)$_{(4)}$
      Okapi BM25$_{(5)}$

    - computer science$_{(0)}$
      artificial intelligence$_{(1)}$
      WordNet$_{(2)}$

Entity Based Approaches

- Fields of Study in the MAG (230k)
  - Not enough overlap between aggregated document descriptions and contexts
  - Extension (including parents of contained FoS) makes descriptions less precise
  - Problems with technical terms (e.g. motivation)

Data Set Citation Flow



Citation Types

Function	Construct	Examples (semantic construct *highlighted*)
Attribution	claim	"Berners-Lee et al. [57] argue that *structured collections of information and sets of inference rules are prerequisites for the semantic web to function*."
	NE	"A variation of this task is '*context-based co-citation recommendation*' [25]."
	-	"In [22] Duma et al. test the effectiveness of using a variety of document internal and external text inputs with a TFIDF model."
Exemplification	NE	"We looked into approaches to *local citation recommendation* such as [19–26] for our investigation."
Further reference	-	"See [58] for a comprehensive overview."
Statement of use	NE	"We use *CiteSeer*^{x} [18] for our evaluation."
Application	NE	"Using this mechanism we perform '*context-based co-citation recommendation*' [25]."
Evaluation	-	"The use of DBLP in [40] restricts their data set to the field of computer science."
Establishing links between sources	claim	"A common motivation brought forward for research on citation recommendation is that *finding proper citations is a time consuming task* [11, 19, 24, 25]."
	-	"Lamers et al. [32] base their definition on the author's name whereas Thompson [30] focusses on the grammatical role of the citation marker."
Comparison of own work with sources	claim	"Like [40] we find that, albeit written in a structured language, *parsing LaTeX sources is a non trivial task*."