Exposé

outlining a Master Thesis on:
Semantic approaches to scientific citation recommendation (tentative title)

Tarek Saier
Reviewer: Prof. Dr. Georg Lausen
Advisor: Dr.-Ing. Michael Färber

1. INTRODUCTION

This exposé will outline a prospective Master Thesis in the area of scientific citation recommendation and argue for its value. The approach will encompass the creation of a dataset and development of supervised learning methods with a focus on semantic analysis of citation contexts. Evaluation of the resulting implementation will follow the most prevalent methods in the field.

The remainder of this document is structured as follows. Section 2 will provide some theoretical background on relevant areas and give a quick overview of related work. A detailed description of the planned methodology and approach will be given in section 3; followed in section 4 by an outline of the planned evaluation. Section 5 and 6 conclude the exposé by listing the expected contributions of the Thesis and a proposed schedule.

2. BACKGROUND

2.1 Citation recommendation

The goal of citation recommendation is to provide adequate citations to a given input text. This can involve evaluating whether or not a given input text includes parts that are suitable to add citations to in the first place. For a given section of or position in an input text, the ouput recommendation can either be a single citation or a ranked list of multiple possible citations. A further distinction can be made concerning the granularity of text that a citation is recommended for. This can range from a complete document (global citation recommendation) to a specific point within a string of text (context aware/local citation recommendation). There are also approaches where citation markers—annotations in the text that mark the position of a citation—are left in the input text. In such a case the evaluation whether or not a citation should be recommended as well as the decision where exactly to put a citation are not necessary. In an ideal case, citation recommendation can even involve evaluating candididate documents in terms of their quality.

Given there are a lot of dimensions along which approaches can differ, section 2.3 will explain relevant terminology and section 2.4 will give an overview of these distinguishing dimensions. This will enable a more easily understandable overview of related work.

2.2 Semantic analysis

The idea of this thesis is to focus on semantic aspects of citation contexts. This means, rather than taking into account only syntactical aspects like n-grams, the analysis will go to a higher level of abstraction where the input's *meaning* is of importance. Because the focus of this analysis will most likely revolve around entities, claims and arguments, these terms will be defined in the following section.

2.3 Terminology

Citing/cited document. The former is the document making a reference while the latter the document being referenced. The contents of both can be taken into account when developing a citation recommendation approach, but in a considerable amount of approaches the cit*ed* documents' content is not.

Citation context. Within the citing document and concerning a single recommendation being made, this is the extend of text provided as input. Examples would be the citing documents abstract, a centence containing a citation marker or a whole document.

Citation marker. A citation marker is an annotation in the input text (or a data set) that marks the location of a citation. In scientific publications this could, for example, be a *[42]*. When left in the input text for a recommendation process, the marker's association to its corresponding reference entry is, of course, removed (e.g. *[42]* could be changed to *[]*, replaced by another type of annotation or the citing document's reference section could be made unavailable during the processing of the input).

Reference. For each citation marker there usually is a corresponding reference at the bottom of the page or near the end of the document. This reference identifies the cited document.

Citation function. The role of a citation or, put differently, the motivation that was behind putting a citation in a particular place. This can, for example, be just for referencing a data set that was used (by citing a data paper), backing up a claim or arguing for or against the overall proposition of a publication.

Metadata. In addition to a document's content, information *about* the document is also often taken into considera-

Winter semester 2018/19

Albert-Ludwigs Universität Freiburg
Technische Fakultät, Institut für Informatik
Lehrstuhl für Datenbanken & Informationssysteme

tion during the recommendation process. This is referred to as metadata.

Entity. A physical or abstract thing in the real world. Generally speaking entities like for example people, places, events and topics can be of interest.

Claim. In this setting a claim can be defined as an assertion which can be judged in tems of its factuality. While non-factual claims also exist (i.e. an opinion being stated), they do not need backing up by citations and are therefore not of interest for citation recommendation.

Argument. An argument can, in alignment with [1], be defined as being composed of a claim and one or more premises justifying the claim. To illustrate, this can take the form <premises> <step(s) of deduction> <claim> where the claim is the conclusion of the deduction.

2.4 Dimensions

To systematically categorize approaches to citation recommendation, distinctions can be made concerning the input and the ouput of a mechanism. With regards to the input, the dimensions *citation context* (length/position), *citation markers* (available or not) and *metadata* (available to what extend) can be used. In part, these can be further broken down as shown in the following table.

	learning	use
citing doc	<val>	<val>
cited doc	<val>	<val>

That is, citing and cited documents can be looked at separately, and a distinction can be made as to what is available during the learning phase and what needs to be provided as input during actual use of the resulting system. Note that for citation markers and context, only the *citing doc* row is applicable[1] and for metadata most likely only the *learning* column is[2]. Because dimensions along tree axis are hard to visualize effectively, the distinction can be flattened to the following aspects:

- citation context (learning)
- citation context (use)
- citation markers (learning)
- citation markers (use)
- metadata (citing doc)
- metadata (cited doc)

[1]The term "citation context" is used to refer to the context in the citing document. One could make a point, though, to furthermore introduce the notion of a context in the cited document. This could then be used, for example, to distinguish whether or not a mechanism outputs only a recommended document or also a specific section that is relevant; or to distinguish whether or not (parts of) cited documents are used during the learning phase.

[2]Although metadata aspects like the "date of the citing doc" could also be used in the online system. That is, given a newly written text without citations, an approach could interpret the input as a "recent citing doc" and recommend citations accordingly.

To give a concrete example, an approach could be trained on input with citation markers (citation markers learning), but be able to give useful output for input without markers (citation markers use) as well.

Above example also suggests, that there is a distinction to be made concerning an approaches output. A dimension *citation placement granularity* can be used to distinguish whether citation recommendations are given for a whole document, on a sentence level or if specific points within the text are identified.

2.5 Related work

[3] (Färber et al.) offers a comprehensive overview of the field of citation recommendation as well as a comparison of concrete approaches. Focus in the following will be works with distinct similarities or differences to the proposed approach (explained in section 3) which are therefore helpful in defining it.

Mishra et al. describe in [11] an approach to recommend news articles that can be used as references for Wikipedia articles describing historical events. Their goal is to offer readers an insight into the detailed view on and reporting of an event *at the time* as an addition to the more overarching representation on Wikipedia. This approach employs named entities as a key component to identify appropriate news articles to recommend. It is therefore similar in this regard to the first step in the Master Thesis' approach where the focus also will lie on recommendation based on entities. The domain, being Wikipedia and news articles, differs from scientific publications.

In [9], Levy et al. describe a method for claim detection using a cascade of classifiers. The detection of claims will also be necessary in the proposed Thesis' second step (citation recommendation based on claims). Levy et al. do, however, restrict their detection of claims to those related to a predefined topic and include claims that are statements of an opinion, which will most likely not be the case in the Master Thesis.

In a similar fashion Goudas et al. tackle argument extraction in [5], which will need to be done in the Thesis' thrid step (citation recommendation based on arguments). The document type being social media texts is, however, different.

Tbahriti et al. semantically analyse abstracts of scientific publications in [14] to aid the retrieval of similar documents. They classify sentences into one of *purpose*, *methods*, *results* and *conclusion* and show that treating these classes of sentences in a distinct manner can help finding documents with similar references. While mentioning that a system like theirs could also be used for recommending citations, this was not part of the approach. Furthermore, the semantic classes of sentences are identified within the abstract and therefore differ in terms of the citation context from the proposed Thesis.

In [2] Duma et al. apply the rhetorical annotation scheme CoreSC[10] (including classes like *hypothesis*, *background*, *method*, etc.) to citation contexts and use this classification as part of the query when determining the citations to recommend. This determination of the function of a citation within the argumentative structure of an input text is an aspect that is likely going to be part of the Master Thesis' thrid step (citation recommendation based on arguments). The used citation context length of 3 sentences is likely to

be in the realm of what will be used in the Master Thesis as well. A difference lies in the research domain of the publications used. While Duma et al. use publications from biomedical science, the Thesis will most likely use publications form the eight fields found on arxiv.org.

Kobayashi et al. classify citations into three discourse facets (*objective*, *method*, *result*) in [8]. For their citation recommendation approach they then use facet based vector representations of their citation graph. While, as with [2], the rhetorical analysis of citation contexts is a similarity to the proposed Thesis, the focus on the citation graph will most likely not be one.

3. METHODOLOGY AND APPROACH

Proposed is a context based citation recommendation approach for scientific publications. In the following subsections considerations concerning the data set and the recommendation process will be explained.

3.1 Dataset

In the field of citation recommendation a lot of different datasets are used and there is no real de facto standard as of yet. Following the analysis in [3] it can be seen that each of the more widely used datasets has benefits and drawbacks. Since the proposed approach for the Master Thesis will involve semantic analysis of citation contexts, the availability of large citation contexts as well as the availability of exact citation markers/placeholders is desirable. [4] show that based on arXiv.org publication sources it is possible to generate a very clean and reliable dataset for citation recommendation. Their approach involving linking publications to DBLP restricts them, however, to only use one of the eight domains available on arXiv.org (namely computer science) resulting in only a medium sized data set. The PubMed-Central Open Access collection[3] would offer a comparably reliable input (XML files with publications' full text, following the NLM/JATS DTD[4]). `#FIXME: any good argument against the PMC Open Access collection? Even includes structured references ("bibitem strings")`

Going for size, another possible approach would be to use a matadata set like the Microsoft Academic Graph[13] (MAG), which is very lage *and* contains citations contexts—albeit only with a length of one sentence. An advantage of using the MAG would be, that the citation graph is already built. In case of generating a data set from arXiv.org sources, reference texts would first have to be linked to publications. [6] report, however, that in case of the MAG, 80 out of the 127 million papers have neither references nor citations.

Considering the above, the use of arXiv.org sources (1.4 M documents) as well as the use of the MAG (30 M connected entities of document metadata) could be considered. Advantages of the MAG are the already built citation graph[5] and its size. The citation context length of just one sentence is a bit small. It might be possible to work on an intersection of MAG and arXiv.org sources and thereby extend the citation contexts (utilizing the MAG's pre-built graph structure and the arXiv.org sources' full text availability), but this would then negate the size argument. Advantages of a dataset generated from arXiv.org sources are the freedom to decide on arbitrarily large citation contexts, the availability of citation markers and the transparency of the approach. A drawback is the necessity to match references (which in the end are just arbitrary strings input by authors) with publications to generate a citation graph.

Considering the size of citation contexts used in other approaches that semantically analyse them like [2] and [8], it could be argued that generating a dataset based on arXiv.org sources is the more promising approach for the goal at hand. The proposed generation of a dataset will therefore look as follows.

3.1.1 Generating a dataset from arXiv.org sources

arXiv.org sources are availabe for bulk download and consist, for the most part, of LaTeX source files. There are, however, publications for which there's only TeX, DVI, PostScript or HTML sources or even just a PDF. Where LaTeX source files are available they can be either converted into XML by software like LaTeXML[6] or Tralics[7] to generate plain text with citation annotations from XML, or parsed by libraries like plastex[8] or TexSoup[9], skipping the inbetween XML format. In case of a restriction to Computer Science publications, grabcite[10] could be used as a tool for the whole dataset generation process.

For the general case, after parsing the LaTeX sources and obtaining plain text with citation annotations, the reference strings corresponding to these citations still need to be resolved (i.e. matched to some kind of global ID). For citations *within* the publications on arXiv.org (i.e. matching to arXiv IDs), this can for example be done by indexing the publications' metadata with a software like Apache Lucene[11] and the querying the resulting indices with the reference strings. Because some LaTeX sources include arXiv IDs in their bibliographies this step is not necessary every time.

In a preliminary test with 10270 citations from 377 publications posted to arXiv in December 2017, 315 of the reference entries already included arXiv IDs and a further 1703 could be matched with the technique described above. While this is only close to 20%, optimistically extrapolating the numbers to the complete 1.4 M documents in arXiv this would still result in a dataset of 7.6 M citations.

3.2 Recommendation

Categorizing the proposed approach according to the dimensions defined in section 2.4, it would look as follows:

- citation context (learning): in accordance with what will turn out to be effective when using entities, claims, arguments or any combination of those. Gauging by related approaches this is probably going to be in the realm of one two a few sentences.
- citation context (use): how much input the system will need to give useful output can hardly be estimated at this point.
- citation markers (learning): will be provided at their exact position, because the dataset will be generated from LaTeX sources.

[3] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[4] https://jats.nlm.nih.gov/publishing/
[5] Although, given the MAG is generated automatically, relying on this pre-built graph could be seen as making an approach less transparent.
[6] https://github.com/brucemiller/LaTeXML
[7] https://www-sop.inria.fr/marelle/tralics/
[8] https://github.com/tiarno/plastex
[9] https://github.com/alvinwan/texsoup
[10] https://github.com/agrafix/grabcite
[11] https://lucene.apache.org/

- citation markers (use): whether citation placeholders will be needed in the input or whether the approach will include a mechanism to decide on where to put citations is still open.
- metadata (citing doc): will be available[12]
- metadata (cited doc): same as above

With regards to the actual recommendation, the machine learning approach will need to allow for the integration of comparatively explicitly modeled concepts (entities, claims, arguments). Furthermore the separate development of approaches (again, for entities, claims, arguments) and their combination afterwards has to be possible. One option would therefore be to go with engineering features that capture abovementioned semantic concepts and use these in a vector space model.

4. EVALUATION

A straightforward approach to evaluating the resulting recommender system is re-prediction of citations. As is the norm with evaluating machine learning approaches the data is split into a training and testing set. While the training set is left "intact" the citations are removed from the test set (citation placeholders might be left in place). The system then trains on the training set and tries to re-predict the citations in the test set that were originally there (or rather, recommend a list of fitting citations). To get more reliable results, above procedure is not just performed with one random split but in the manner of a k-fold cross-validation.

To measure the quality of the resulting recommendation a combination of a "hard" measure like *top-1 accuracy* and a more "soft" measure like *normalized discounted cumulative gain* (NDCG)[13] can be used.

5. CONTRIBUTIONS

- apparently semantic stuff not very explored (cite survey if possible, look at tables) - creation of another nice (exact citation markers, large citation context, etc.) dataset like gold standard paper[4] - a nice dataset like gold standard paper[4] but not restricted to CS domain

6. SCHEDULE

The proposed schedule can be seen in table 1.

7. REFERENCES

[1] P. Besnard and A. Hunter. *Elements of Argumentation*. The MIT Press, 2008.

[2] D. Duma, E. Klein, M. Liakata, J. Ravenscroft, and A. Clare. Rhetorical classification of anchor text for citation recommendation. *D-Lib Magazine*, 22, 2016.

[3] M. Färber and A. Jatowt. Citation Recommendation for Scientific Publications.

[4] M. Färber, A. Thiemann, and A. Jatowt. A High-Quality Gold Standard for Citation-based Tasks. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC 2018, 2018. r.

[5] T. Goudas, C. Louizos, G. Petasis, and V. Karkaletsis. Argument extraction from news, blogs, and social media. In A. Likas, K. Blekas, and D. Kalles, editors, *Artificial Intelligence: Methods and Applications*, pages 287–299, Cham, 2014. Springer International Publishing. r.

[6] D. Herrmannova and P. Knoth. An analysis of the microsoft academic graph. *D-Lib Magazine*, 22(9/10), 2016. r.

[7] S. E. Hug, M. Ochsner, and M. P. Brändle. Citation analysis with microsoft academic. *Scientometrics*, 111(1):371–378, Apr 2017. r.

[8] Y. Kobayashi, M. Shimbo, and Y. Matsumoto. Citation recommendation using distributed representation of discourse facets in scientific articles. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, pages 243–251, New York, NY, USA, 2018. ACM. READ!

[9] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. r.

[10] M. Liakata, S. Teufel, A. Siddharthan, and C. R. Batchelor. Corpora for the conceptualisation and zoning of scientific papers. In *LREC*, 2010. READ!

[11] A. Mishra and K. Berberich. Leveraging semantic annotations to link wikipedia and news archives. In N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, and G. Silvello, editors, *Advances in Information Retrieval*, pages 30–42, Cham, 2016. Springer International Publishing. r (ch 1-3).

[12] B. Paszcza. Comparison of microsoft academic graph with other scholarly citation databases, 11 2016. r (ch 1,""3"").

[13] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 243–246, New York, NY, USA, 2015. ACM. r.

[14] I. Tbahriti, C. Chichester, F. Lisacek, and P. Ruch. Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the medline digital library. *International Journal of Medical Informatics*, 75(6):488 – 495, 2006.

[12] https://arxiv.org/help/oa/arXiv_meta_format.html

[13] Making good use of NDCG might necessitate defining a relevance measure for recommendations that differ from the original one.

Table 1: Schedule

Time frame	Task	Results
Oct 1 – Oct 21	Develop mechanism to generate dataset with citation markers from arXiv source dump	Dataset boilerplate (i.e. with citation markers but no semantic annotation)
Oct 22 – Oct 28	Write exposé	Thesis approval
Oct 29 – Nov 04	Addd entity annotations to dataset	Dataset usable for supervised learning
Nov 05 – Nov 18	Develop entity based recommendation approach	-
Nov 19 – Nov 25	Addd claim annotations to dataset	-
Nov 26 – Dec 09	Develop claim based recommendation approach	-
Dec 10 – Dec 22	Coordination with simultaneous tangential theses and integration into CiteRec system	-
Dec 23 – Jan 06	break/buffer	-
Jan 07 – Jan 13	Addd argument annotations to dataset	-
Jan 14 – Jan 27	Develop argument based recommendation approach and start offline evaluation	-
Jan 28 – Feb 10	Offline evaluation	-
Feb 11 – Feb 24	Online evaluation	-
Feb 25 – Mar 17	Thesis writing	-
Mar 18 – Mar 31	buffer/paper writing	-