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 to a specific point within a string of text. There are also approaches where citation markers—annotations in the text that show where a citation was in the original—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 cited documents in terms of their quality.

Given there are a lot of dimensions along which approaches can differ, the 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. (add citation)

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 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 citation act of recommendation, 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 would, for example, be a *[1]*.

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, arguing for or against the overall proposition of a publication, etc.

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

Meta data. In addition to a documents content, information *about* the document is also often taken into consideration during the recommendation process. This is referred to as meta data.

Winter semester 2018/19

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

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.

Argument. An argument is 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 a conclusion.[1]

2.4 Dimensions

```
input
    citing doc
        citation context (length)
        citation marker provided
        what metadata provided
    cited doc(s) / "corpus"

output
    citation placement granularity
```

2.5 Related work

- survey stuff (take more narrow look)
- wiki page news article papers (for entitiy stuff)

3. METHODOLOGY AND APPROACH

foo bar
MAG[8][4][5][9]
entity[7]
claim[6]
argument[3]

data sets that were considered and why (benefits, drawbacks, ... (cite accodingly))

MAG start and arXiv start scenario (see wiki)

details of arXiv processing, challenges, etc. (MAG for evaluation where citation marker position is not relevant)

4. EVALUATION

foo bar

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[2] - a nice dataset like gold standard paper[2] but not restricted to CS domain

6. SCHEDULE

7. REFERENCES

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

[2] 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.

[3] 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.

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

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

[6] 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.

[7] 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).

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

[9] 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.

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	-