Learning Neural Textual Representations for Citation Recommendation

Binh Thanh Kieu, Inigo Jauregi Unanue, Son Bao Pham, Hieu Xuan Phan, Massimo Piccardi*

*speaker; University of Technology Sydney

ICPR 2020, Milan, Italy, 10-15 January 2021

Highlights

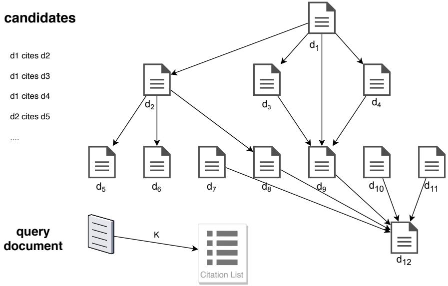
- In this paper we propose an effective method for **citation recommendation**
- The main components are:
  - a submodular scoring function to select the citations
  - a deep sequential representation for the documents using Sentence-BERT [Reimers & Gurevych EMNLP 2019]
  - a fine-tuning approach based on twin and triplet networks

Citation recommendation

- **Citation recommendation** aims to recommend references for a given document out of a pool of citable documents
- What can it be useful for? For instance, to find appropriate references for a draft you have started to write
- While we do not do this here, it can also be heavily personalised to the user's preferences, targeted venue etc

Citation recommendation: the formal task

- We are given a **query document**, q, and a **corpus** of citable documents, $C = (d_1, d_2, ..., d_N)$, which likely form a citation graph
- The task is to choose a subset $\bar{S} \subseteq C$ with $|\bar{S}| \leq K$ to be the **recommended citation list**



Citation recommendation: a baseline approach

A straightforward approach to citation recommendation could be:

- Turn the documents into some numerical representation, e.g. TF-IDF
- Compute the similarity between the query and each candidate document using some similarity function, e.g. the cosine similarity
- Recommend the top-K most similar documents

Risk? $\rightarrow$ *redundancy*!!!

Submodularity to the rescue

- When selecting the citations, one should balance **similarity to the query** and **diversity** of the recommended citations
- A scoring function that balances these two properties is typically **submodular**
- Finding the citation list that maximizes a submodular scoring function is NP-hard

Submodularity to the rescue

- However, submodular functions enjoy a key property: selecting the citations one by one with a simple, greedy algorithm is **near-optimal**[1]

- The greedy algorithm scans the corpus K times, every time adding a citation to the partial list based on a) the citations already selected and b) the rest of the corpus

- So, it is computationally heavier than a simple top K similarity search, but manageable in many cases

[1] not so "near" $\ddot\smile$ $>$ 0.632 of the actual maximum

Document representation and similarity function

- In all cases, you will need to convert your documents into a numerical representation
- Classic methods to encode a document: TF-IDF, BM-25 etc
- We instead use **Sentence-BERT** [Reimers & Gurevych EMNLP 2019]: a neural approach to embed a whole sentence/paragraph/short document into a vector using any pre-trained BERT model. It can be fine-tuned.
- A simple cosine similarity as the similarity function

Sentence-BERT fine-tuning

- We fine-tune Sentence-BERT **in a supervised manner**
- Annotated training set: the documents and their citations in the corpus (the citation graph)
- Distance between any two documents in the citation graph, $dis(d_i, d_j)$: number of nodes in the shortest path from node i to node j
- **Positive examples** for query d_i: all d_js with $dis(d_i, d_j) \leq 3$
- **Negative examples**: all the others. To limit the training time, we only use subsets of the negative examples, selected with three different strategies: *Random*, *Nearest* (to the query in similarity) and *Farthest*

Fine-tuning objectives

- **Twin aka "Siamese" network-style**: given a query document, q, and a positive or negative candidate, d, we minimize the mean squared difference between their predicted and target similarities
- **Triplet network-style**: given the query, q, a positive candidate, d^+ and a negative candidate, d^-, we impose that the predicted similarity $s(q, d^+)$ be larger than $s(q, d^-)$ by a margin:

$$\text{triplet loss} = \max[s(q, d^-) - s(q, d^+) + 1, 0] \qquad (1)$$

Experiments

- **Dataset**: *ACL Anthology Network corpus (AAN) [1]*: a dataset of **22, 085** papers in the field of computational linguistics. Papers + meta-information
- We replicate the experimental setup of [2] by excluding papers with no references and using the standard training (16, 128 papers from 1960 to 2010), dev/validation (1, 060 papers from 2011) and test (1, 161 papers from 2012) splits

Experiments

- We **fine-tune Sentence-BERT** as described (further details in the paper)
- At inference time, we use a **submodular scoring function** (details in the paper)
- **Performance evaluation**: Mean Reciprocal Rank (MRR) and F1@*k* score
- **Compared approaches**:
  - ElasticSearch with Okapi BM25 [2]
  - Citeomatic [3]
  - Our previous submodular approach, SubRef [2]
  - The proposed method with top-K inference
  - The proposed method with submodular inference

[2] https://www.elastic.co/

Main results

Method	MRR	F1@10	F1@20	F1@50	F1@100
ElasticSearch					
BM25	0.2437	0.0701	0.0632	0.0446	0.0321
Citeomatic					
Select	0.3085	0.1281	0.1339	0.0940	0.0548
Select+Rank	0.3777	0.1590	0.1472	0.0959	0.0549
SubRef (best on dev)					
BM25-QAIv2	0.3320	0.1310	0.1264	0.0911	0.0621
SBERT + top-K (best on dev)					
Siamese, d=2, farth.	0.3493	0.1424	0.1400	0.1096	0.0797
SBERT + submod (best on dev)					
Siamese+QAIv2	**0.4431**	**0.1978**	**0.1839**	**0.1327**	**0.0918**

Table: Results on the test set

Conclusions

- A novel approach to citation recommendation that leverages a deep representation of the documents
- An approach for fine-tuning Sentence-BERT with positive and negative examples derived from the citation graph
- A submodular scoring function for recommending the citations that balances their similarity to the query with their (author) diversity
- Outperformed all the compared approaches, including a state-of-the-art neural approach, Citeomatic, on the AAN dataset

Key references

- D. R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara, “The ACL anthology network corpus,” Language Resources and Evaluation, pp. 1–26, 2013.
- T. B. Kieu, B. S. Pham, X. H. Phan, and M. Piccardi, “A submodular approach for reference recommendation,” in PACLING, Hanoi, Vietnam, Oct. 2019, pp. 3–14.
- C. Bhagavatula, S. Feldman, R. Power, and W. Ammar, “Content-based citation recommendation,” in NAACL-HLT, New Orleans, Louisiana, Jun. 2018, pp. 238–251.

Any (virtual) questions?

- Thank you very much for your attention!
- Any (virtual) questions?
- Please email:
  binhkt.vnu@gmail.com, massimo.piccardi@uts.edu.au

Binh Thanh Kieu, Inigo Jauregi Unanue, Son Bao Pham, Hieu Xuan Phan, Massimo Piccardi
University of Technology Sydney, NSW, Australia
University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam