Learning Neural Textual Representations for Citation Recommendation

Binh Thanh Kieu, Inigo Jauregi Unanue, Son Bao Pham, Hieu Xuan Phan, Massimo Piccardi
University of Technology Sydney; University of Engineering and Technology, VNU, Hanoi

Highlights

In this paper we propose an effective approach for **citation recommendation**. Its 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] [1]
- a fine-tuning approach based on twin and triplet networks

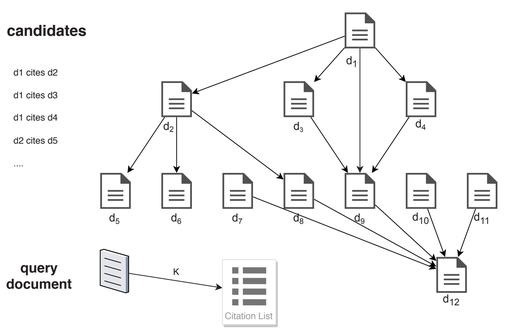
Citation Recommendation

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.

Formally, 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**.



Approach: Submodular Inference

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 a similarity function, e.g. the cosine similarity.
- Recommend the top-K most similar documents.

Risk? → redundancy!!! All the recommended citations may look the same.

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**, and finding the citation list that maximizes it is NP-hard. However, submodular functions enjoy a key property: selecting the citations one by one with a simple, greedy algorithm is **near-optimal**.[a]

[a] not so "near" ☺ > 0.632 of the actual maximum.

Approach: Document Representation

To represent the document, we 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. We fine-tune it **in a supervised manner**:

- Annotated training set: the documents and their citations in the corpus (the citation graph).
- **Positive examples** for query q: all documents d_j's with number of nodes in the shortest citation path from $q \leq 3$.
- **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*.

As fine-tuning objectives, we use:

- **Twin aka "Siamese" network-style**: given a query, 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:

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

Conclusion

- 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.

Experiments and Results

- **Dataset**: *ACL Anthology Network corpus (AAN) [4]*: a dataset of **22,085** papers in the field of computational linguistics. Papers + meta-information. Same experimental setup as [2] 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.
- **Performance evaluation**: Mean Reciprocal Rank (MRR) and F1@k score
- **Compared approaches**:
  - ElasticSearch with Okapi BM25
  - Citeomatic [3]
  - Our previous submodular approach, SubRef [2]
  - The proposed method with a simple top-K inference
  - The proposed method with submodular inference

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**

The proposed approach has outperformed all the compared approaches in all metrics on this dataset. We will extend our evaluation to other datasets in the near future.

References

[1] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*, pages 3982–3992, Hong Kong, China, November 2019.

[2] Thanh Binh Kieu, Bao Son Pham, Xuan Hieu Phan, and Massimo Piccardi. A submodular approach for reference recommendation. In *PACLING*, pages 3–14, Hanoi, Vietnam, October 2019.

[3] Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. Content-based citation recommendation. In *NAACL-HLT*, pages 238–251, New Orleans, Louisiana, June 2018.

[4] Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. The ACL anthology network corpus. *Language Resources and Evaluation*, pages 1–26, 2013.