GatesNLP: An Experimental Extension to Content-Based Citation Recommendation
CSE 481 N

Bryan Hanner * bryhan@cs.uw.edu
Chia-ko Wu * chiakowu@cs.uw.edu
Mitali Palekar * mitali97@cs.uw.edu
Swojit Mohapatra * swojit@cs.uw.edu

June 2019

Abstract

For our University of Washington NLP capstone, we present an extension of prior work on content-based citation recommendation. Given a query paper's text and candidate papers' text, we generate a ranked list of candidate papers in which cited papers appear first. We experiment with a variety of supervised and unsupervised learning approaches to generate rankings as well as evaluate the effectiveness of our models' rankings quantitatively and qualitatively. We find that our best model is TF-IDF, achieving a mean reciprocal rank of 0.31 on our test set. Finally, we release two new datasets of natural language processing and computer security papers to facilitate future academic research.

1 Introduction

Due to the large amount of scientific literature being published today, exploring related literature has become challenging. Content-based citation recommendation has the potential to reduce this problem by signalling related literature to authors as well as readers of different forms of scientific literature.

We extend upon the work presented by Bhagavatula et al. [2018] by implementing a variety of unsupervised and supervised learning approaches that use paper titles and abstracts as input. We then empirically analyze our results, developing insights into strengths and weaknesses of our models. Finally, we also release two new datasets of natural language processing and computer security papers on Kaggle.

Our main contributions include the following:

1. Two new datasets of natural language processing and security papers – (1) raw papers and (2) a pairwise sampling of papers from the first dataset.
2. Baselines using the Jaccard similarity index and Term Frequency-Inverse Document Frequency (TF-IDF).
3. A supervised model for ranking papers by the pairwise confidence score of a neural network trained to classify if one paper cites another.
4. An unsupervised model using GloVe for ranking papers using Word Mover's Distance (WMD).
5. A pre-trained NER model leveraging SciBERT to generate rankings using the number of shared entities.

2 Technical Ideas

In this section, we first formalize our task and then present three distinct ideas that lead to our subsequent dataset creation process as well as the exploration of our models.

*University of Washington, all authors listed in alphabetic order.

2.1 Task Overview

We formulate content based citation recommendation as a ranking problem. Given a query paper P and a list of candidate papers (our training set), we develop a ranking of all candidate papers where papers that P might cite appear higher than papers that don't. As such, our ranking suggests that papers higher up in the ranking are closer in content to the query paper than papers lower in ranking. For candidate papers, we focus on titles and abstracts as a summarization of their content due to time, computational, and licensing constraints.

2.2 Idea Formulation

Dataset Creation: Citations generally occur within a field. Through an empirical analysis of past scientific literature, we realize that paper citations often occur within the same broader *field*. We leverage this in creating a new dataset of natural language and computer security papers.

Our dataset spans the fields of natural language processing and computer security and includes abstracts, titles, and metadata [Ammar et al., 2018] filtered by venue from the Semantic Scholar corpus[1]. We specifically pick natural language processing and computer security as fields due to our familiarity with the topics and access to the data. Our dataset contains 26,495 papers (7,450 natural language processing papers and 19,045 security papers) and is available on Kaggle[2].

Supervised Approach: People cite literature for different reasons and learning this relationship would be important to understand future citation recommendations. Through a discussion with course staff and current students, we realize that scientific literature is cited due to a variety of reasons. We aim for our task-specific supervised neural model to learn textual patterns that correspond to these relationships. Our model takes as input two papers and their associated label as cited/not cited and is trained to predict if the first paper cites the second. Then, the generated model's confidence score (for whether the first paper cites the second) is used to sort the ranked list of candidate papers.

Unsupervised Approach: Content plays an important role in citations, so understanding semantic meaning is important. At the same time, we believe that content, specifically semantic meaning, plays an important role in recommending citations. As such, we implement approaches that encode semantic meaning for both baselines (Jaccard similarity and TF-IDF) as well as unsupervised approaches (GloVe and NER tagging) to rank candidate papers.

3 Empirical Section

In this section, we present a deep dive into our model approaches as well as define key evaluation metrics for the performance of our models. Across all models, unless stated otherwise, we leverage paper titles and abstracts as the paper text, and all inputs are lowercased and tokenized with the basic SpaCy tokenizer.

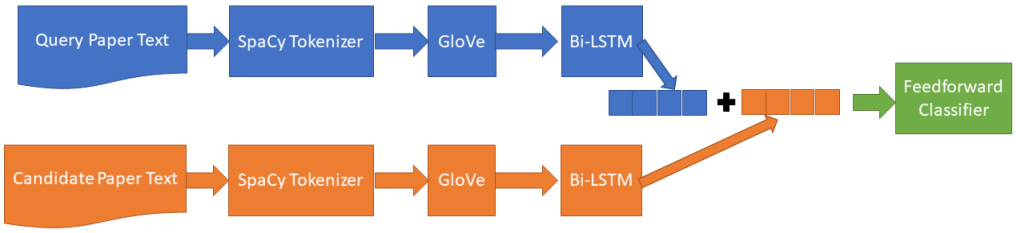
3.1 Model Approaches

Baseline Approaches: We implement Jaccard similiarity and TF-IDF as baseline measures for performance. Jaccard treats each paper as a set of words and gives us a score for each pair (known as the Jaccard index) using the portion of shared words out of all words in both documents. TF-IDF is more nuanced because it considers how often words occur in each document (term frequency) and gives more weight to words found in fewer documents (inverse document frequency). We calculate a TF-IDF embedding for each document in a pair and then take the cosine similarity of these two embeddings. For both baselines, a higher score means a pair is more similar.

[1] https://api.semanticscholar.org/corpus/

[2] https://www.kaggle.com/mitalipalekar/gatesnlp

Figure 1: Pairwise Supervised Model Architecture



Supervised Approaches [Bryan]: First, we create a pairwise dataset to train and evaluate the supervised model[3]. In this dataset, the first paper of each pair is always drawn from train, but the second paper is drawn from train, dev, or test to increase the available pairs and match our task. We experiment with different makeups of the dataset while keeping a 50/50 split of positive/negative examples in each dataset. We can change the makeup by drawing completely random positive or negative examples, or drawing *one-hop pairs* to replace some or all of the negative examples. A *one-hop pair* (A, C) is where A cites B and B cites C, but A does not cite C. In early experiments, we found the pairwise accuracy increased with the dataset size, so we draw 120,000 total pairs.

To use this model, we embed both the query paper and the candidate paper tokens with 300-dimensional GloVe embeddings and encode them using a bidirectional LSTM and 50 hidden dimensions. Both encodings are then concatenated together and passed through a feedforward network with a sigmoid and a linear layer. We use a dropout of 0.4. To train, we used the adam optimizer. The model is implemented and configured using AllenNLP [Gardner et al., 2018]. This model architecture is represented in Figure 1.

GloVe [Mitali/Swojit]: GloVe is a learning algorithm for obtaining vector representations of words [Pennington et al., 2014]. To generate word embeddings, we experiment with GloVe embeddings pre-trained on Wikipedia and Gigaword for 50, 100, 200, and 300 dimensions as well as GloVe embeddings trained on our papers' titles and abstracts. These embeddings are generated solely for titles (without abstracts) due to training time constraints. For each pair of papers, we calculate Word Mover's Distance: an algorithm for calculating the distance between two documents based on provided pre-trained embeddings, even when there are no words in common [Wu et al., 2018]. In this case, a lower score means a paper pair is more similar.

BERT [Swojit/Mitali]: We also attempt to take advantage of Bi-directional Transformers for Language Understanding (BERT) which provides contextualized word embeddings [Devlin et al., 2019]. In this case, we average over pre-trained word embeddings to generate embeddings for paper titles and then apply cosine similarity to each pair of title embeddings.

NER Tagging [Chia-ko]: Named Entity Recognition (NER) is a sequence labeling task that finds references to named entities such as topics and people in the text. We leverage SciBERT since it is trained on scientific text from the SciERC dataset in contrast to vanilla BERT [Beltagy et al., 2019, Luan et al., 2018]. Further, we extract entities from each query paper using SciBERT's pre-trained NER model and calculate how many unique entities are shared with each candidate paper's "entities" field, ignoring generic entities such as pronouns. The entities could also be extracted from the candidate text using NER if we wanted to exclusively use the candidates' text, but we use the existing field due to time constraints. At the same time, since there are many candidates sharing the same ranking, we add the cosine similarity in range [0, 1) of TF-IDF embeddings to our score (assuming no duplicate embeddings) to order the candidates first by the number of

[3]available on Kaggle: https://www.kaggle.com/mitalipalekar/gatesnlp#pairs_*.txt

Table 1: Experimental Results

Model	Dev MRR	Test MRR
Jaccard	0.23	0.17
TF-IDF	**0.37**	**0.31**
Supervised $\frac{1}{2}, \frac{1}{2}, 0$	0.09	0.09
Supervised $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$	0.16	0.23
BERT	0.004	0.03
200-dim pre-trained GloVe	0.18	0.22
NER with TF-IDF	0.28	0.23

(a) Ranking Evaluation Results (Supervised lists the fraction of random positives, random negatives, one-hops used for training and evaluation)

Model	Train	Dev	Test
Supervised $\frac{1}{2}, \frac{1}{2}, 0$	0.82	0.83	0.82
Supervised $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$	0.67	0.67	0.67

(b) Pairwise accuracy for the supervised models.

Dim	Pre-trained	Dev MRR	Test MRR
50	yes	0.18	0.21
100	yes	0.18	0.21
200	yes	0.18	0.22
300	yes	0.18	0.21
50	no	0.18	0.22

(c) GloVe experiments with different dimensions.

shared entities and then by TF-IDF.

3.2 Evaluation & Results

In this section, we present our results and evaluate our models using the mean reciprocal rank (MRR) metric. MRR assigns a rank to be the reciprocal of the first index where a paper is correctly cited. For example, if the first correct citation in ranking R is ranked 3rd, then its reciprocal rank $= 1/3 = 0.33$. We take the mean over all such R to get the MRR for the entire dataset. There are also other relevant metrics such as F1 which we leave out due to time constraints. The results have been presented in Table 1.

From Table 1 we see that the TF-IDF baseline produces the best MRR results for our dataset. We also notice that the supervised model trained on one-hops has lower pairwise accuracy (albeit on a different dataset) but creates rankings with a higher MRR than the model trained only on randomly sampled examples. We also note that the dimensionality of the GloVe embeddings does not significantly affect our ranking scores. Finally, we see that the combination of shared entities and TF-IDF performed significantly worse than TF-IDF on its own. We also note that the BERT scores are particularly low compared to the others (we reason that this might be because BERT does not produce meaningful sentence level embeddings).

Empirical Analysis: As we empirically evaluate our models, we notice that our models capture three tiers of meaning: *field* (distinct domain of study such as natural language processing), *topic* (particular idea within a field such as machine translation), and *word*. Using this framework, we see four main trends within our models' rankings:

1. The supervised model develops field-level rankings (papers within the same field have similiar rankings), but fails to understand different topics within a field.
2. For TF-IDF, the rankings seem to be developed at a topic level. This seems to be a useful granularity for the semantic similarity needed for content-based citation recommendation.
3. For Jaccard similarity and GloVe, papers that have a higher word overlap consistently have a higher score, which suggests a word-level interpretation of the papers.
4. For NER/TF-IDF, the simple count of shared entities is a crude estimation of semantic similarity and as such, the primary sort by shared entities reduces the overall ranking quality.

To make these above analyses more concrete, we consider the following example query paper titled *Code-Mixed Question Answering Challenge: Crowd-sourcing Data and Techniques*. We notice that the

highest rank for a cited paper in the generated rankings for TF-IDF is *I am borrowing ya mixing ?' An Analysis of English-Hindi Code Mixing in Facebook* (in 9th place) and the NER model is even better at 2nd. In contrast, the one-hop supervised model's highest rank for a cited paper is 74 and GloVe's is 477. This shows the ability of TF-IDF to generally outperform our other models by estimating topic-level reasoning.

Finally, looking at the broader generated rankings by different models, we noticed that in the supervised case, the first ranked paper is the GloVe paper by Pennington et al. [2014], which is one of the most commonly cited papers in NLP, displaying the field-level information extraction. The GloVe model's first ranked paper is titled *Question Answering Using Maximum-Entropy Components*, which is one of the papers that includes "question answering" in the title, highlighting the word-level matching as described above. In this example, the NER model does better than TF-IDF by ranking the same paper 2nd instead of 9th because they share the entity "code - mixing", suggesting that some entities are worth giving a larger weight than TF-IDF would give to a general rare word.

4 Related Work

Prior work suggests two types of citation based recommendation systems – local (uses only a few sentences as input) and global (uses the entire paper as input). We take a mix methods approach, using titles and abstracts, lying at the interaction of global and local recommendation systems. Moreover, much of the prior work (McNee et al. [2002], Jia and Saule [2017], Liu et al. [2015], Ren et al. [2014], Yu et al. [2012]) requires additional input such as venue, author, an initial list of citations etc. We distinguish our work by using only paper titles, abstracts, and extract-able entities. Finally, our work extends upon the work carried out by Bhagavatula et al. [2018]. While we reference this work as a base, we focus on a breadth-based approach to model development, emphasizing the strengths and weaknesses of different information extraction and retrieval techniques.

5 Future Work

We outline concrete future directions both in the short-term and long-term. In the short term, we suggest further exploration of our models. This includes experimenting with Word Mover's Distance in conjunction with BERT as well as implementing a bag-of-embeddings encoder in the supervised model case. In the long-term, we suggest improving semantic meaning encoding by interpolating our models as well as exploring methods to select the most content-dense text from papers.

6 Conclusion

In this paper, we present an extension of prior work on content-based citation recommendation. We implement several different supervised and unsupervised approaches as well as evaluate the strengths and weaknesses of our different approaches. In addition, we release two new datasets of natural language processing and computer security papers to facilitate future academic research. By our exploration of citation recommendation models, we hope to set the stage for a substantial field of future work.

7 Acknowledgments

We would like to thank the NLP capstone course staff Elizabeth Clark, Lucy Lin, and Noah Smith as well as Iz Beltagy (AI2) for insightful discussions and continuous feedback. We would also like to thank to our fellow students for their helpful ideas and feedback in class and on the discussion board: Amol Sharma, Dan Tran, Deric Pang, Divye Pratap Jain, Emma Casper, Ethan Chau, Jeff Da, Kaitlyn Zhou, Nelson Liu, Ravi S Patel, and Shobhit Ketanbhai Hathi.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew E. Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. Construction of the literature graph in semantic scholar. In *NAACL-HLT*, 2018. URL https://arxiv.org/abs/1805.02262.

Iz Beltagy, Arman Cohan, and Kyle Lo. Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676, 2019. URL http://arxiv.org/abs/1903.10676.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. Content-based citation recommendation. In *NAACL-HLT*, 2018. URL http://arxiv.org/abs/1802.08301.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. Allennlp: A deep semantic natural language processing platform. *CoRR*, abs/1803.07640, 2018. URL http://arxiv.org/abs/1803.07640.

Haofeng Jia and Erik Saule. An analysis of citation recommender systems: Beyond the obvious. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, pages 216–223, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4993-2. doi: 10.1145/3110025.3110150. URL http://doi.acm.org/10.1145/3110025.3110150.

Haifeng Liu, Xiangjie Kong, Xiaomei Bai, Wei Wang, Teshome Megersa Bekele, and Feng Xia. Context-based collaborative filtering for citation recommendation. *IEEE Access*, 3:1695–1703, 2015. URL https://ieeexplore.ieee.org/document/7279056.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *EMNLP*, 2018. URL http://arxiv.org/abs/1808.09602.

Sean M. McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. On the recommending of citations for research papers. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, pages 116–125, New York, NY, USA, 2002. ACM. ISBN 1-58113-560-2. doi: 10.1145/587078.587096. URL http://doi.acm.org/10.1145/587078.587096.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Xiang Ren, Jialu Liu, Xiao Yu, Urvashi Khandelwal, Quanquan Gu, Lidan Wang, and Jiawei Han. Cluscite: Effective citation recommendation by information network-based clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 821–830, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2956-9. doi: 10.1145/2623330.2623630. URL http://doi.acm.org/10.1145/2623330.2623630.

Lingfei Wu, Ian En-Hsu Yen, Kun Xu, Fangli Xu, Avinash Balakrishnan, Pin-Yu Chen, Pradeep Ravikumar, and Michael J. Witbrock. Word mover's embedding: From Word2Vec to document embedding. In *EMNLP*, 2018. URL https://www.aclweb.org/anthology/D18-1482.

Xiao Yu, Quanquan Gu, Mianwei Zhou, and Jiawei Han. *Citation Prediction in Heterogeneous Bibliographic Networks*, pages 1119–1130. 04 2012. ISBN 978-1-61197-232-0. doi: 10.1137/1.9781611972825.96.