Citcom – Citation Recommendation

Melina Meyer[1] Jenny Frey[1] Tamino Laub[1] Marco Wrzalik[2] Prof. Dr. Dirk Krechel[2]

Abstract: Citation recommendation aims to predict references based on a given text. In this paper, we focus on predicting references using small passages instead of a whole document. Besides using a search engine as baseline, we introduce two further more advanced approaches that are based on neural networks. The first one aims to learn an alignment between a passage encoder and reference embeddings while using a feature engineering approach including a simple feed forward network. The second model takes advantage of BERT, a state-of-the-art language representation model, to generate context-sensitive passage embeddings. The predictions of the second model are based on inter-passage similarities between the given text and indexed sentences, each associated with a set of references. For training and evaluation of our models, we prepare a large dataset consisting of English papers from various scientific disciplines.

Keywords: citation recommendation; natural language processing; representation learning

1 Introduction

Writing scientific papers or any kind of work that requires a lot of citing can be very exhausting. A lot of research is necessary to find documents that are close to a specific topic and worth citing to prove a point. There already exist some tools to propose related work. However, there is potential for improvements in these tools and thus for research. The task focusing on this problem is called citation recommendation. It can be be divided into two categories: global approaches and context-based recommendation. While approaches of the first category pay attention on the whole document, context-aware attempts focus on sentences before and after a citation. After considering various options, three different context-related approaches to this task are pursued here, including the baseline. The first is to create a simple proof of concept application in order to get a benchmark for recommendation performance. The other models are based on neural networks. As a baseline experiment, we use Elasticsearch [3], a full-text search engine that can find documents due to their structural similarity. The second model uses feature engineering and a feed forward network. Additionally, we wanted to use a tool that learns characteristics of the language used in the documents and takes semantics into consideration, which brought us to BERT[De19] and is used in the third model. This model learns passage similarity using BERT embeddings to

[1] RheinMain University of Applied Sciences, Faculty of Design Computer Science Media, Unter den Eichen 5, 65195 Wiesbaden, Germany, <forename>.<surname>@student.hs-rm.de

[2] RheinMain University of Applied Sciences, Working Group LAVIS – Learning and Visual Systems, Unter den Eichen 5, 65195 Wiesbaden, Germany, lavis@hs-rm.de

[3] See https://www.elastic.co/

doi:10.18420/inf2020_82

predict references. For evaluation, we create a dataset which is based on a dump of arXiv documents of all scientific disciplines from recent years. In summary, our contribution is to prepare a comprehensive dataset suitable for citation recommendation, a baseline for comparison purposes and to develop two more advanced models.

2 Related Work

Citation recommendation is a complex task that has gained increasing attention in recent years. Besides some methods for proposing references for entire documents [CHY18; Kü12; Re14], many new publications refer to the context-based approach. Local recommendation, as the term was coined by the authors of [He10], uses contextual information. Based on this idea, [Hu12] proposed an approach via a machine translation system transferring keywords of the context into cited documents. They continued this work in [Hu15] by adding semantic embeddings of the words of the context as well as cited documents. Finally, a recommendation is carried out based on the semantic distance in vector space. [TWZ14] presented the first embedding-based approach for context-aware citation recommendation, which uses TF-IDF vectors to form cross-language embeddings and uses them for the proposals. Approaches without neural networks based on information retrieval techniques and metrics such as TF-IDF or BM25 also have been investigated. In their proposals, [Du16] annotate each sentence of the given documents with CoreSC classes, which are indexed in Lucene and used to determine similarity. On the other hand, [EF17] use BM25 as a baseline for their encoder-decoder framework, inspired by neural machine translation, which learns relations between text pairs of variable length. The approach is further supplemented by additionally analyzing the writing style of authors. Other procedures also include document metadata. [FS20] is a semi-genetic hybrid recommender system for citation recommendation. The authors combine embedding and information retrieval approaches using a fitness score to receive the top k recommendations. Recent approaches additionally include language models such as BERT. In [Je19], BERT is used as a context encoder for textual embeddings, supplemented by a Graph Convolutional Network (GCN) for metadata and reference relationship between papers as a citation encoder for the construction of graph embeddings. The concatenation of the output vectors is then used as input of a feed forward network.

3 Dataset

In order to recommend citations for scientific papers and to evaluate our model on a large dataset, we reuse the record provided in [SF20]. The dataset is based on an arXiv source dump including papers from 1991 to 2018 and of all scientific disciplines available on arXiv.org. It offers 29.2 million ready-made citation contexts, each consisting of three sentences: the sentence containing the citation and the two surrounding ones. We use these contexts to find the corresponding sections in the full text files, extract all references of the section and map them to the arXiv- and/or MAG-ID of the cited paper. We only consider references that could be matched to arXiv- or MAG identifiers as only these can be verified. Furthermore, we use a dump of arXiv metadata[4] to obtain basic background information about the cited papers. Since MAG metadata are difficult to obtain for research purposes, we limited our experiment to passages that reference known arXiv papers as we need the metadata for one of our approaches. Finally, training and test data are divided in the ratio 80:20. Since our models require a lot of computing power, we create another dataset to reduce the computing time. Since we want to use the scientific categories as additional meta information, we cluster the passages into 176 clusters based on those categories. From each cluster, five percent of the references are randomly selected and citing passages are part of the reduced dataset. This dataset is also randomly split in the ratio of 80:20. Table 1 illustrates the details of our datasets containing the total number of disjoint references, training passages, disjoint references contained by the training passages and test passages.

Dataset	**#Refs**	**#Train Pass.**	**#Train Refs**	**#Test Pass.**
full	718,329	14,932,722	700,220	3,733,181
reduced	115,643	1,339,920	115,643	329,130

Tab. 1: An overview of the employed datasets

4 Model Architectures

4.1 Baseline Model

In order to get an initial orientation, a baseline experiment is first performed using Elasticsearch. Elasticsearch is a search engine based on the Lucene program library for full text searches. First, several documents are added to indices so that similar documents can be found by a query using a RESTful API. For both of the datasets, a new index is created. Due to the large amount of training data, the passages are grouped according to their content. This means that passages with the same context but different target references are identified and indexed only once. Each indexed document contains a text field with `english` language analyzer[5] as well as two keyword fields for lists of all associated passage IDs and target references. This reduces the amount of training data to a total of 8,700,191 indexed documents for the full dataset and 1,199,747 for the reduced one. Especially for the full dataset, this has a significant impact on the duration of the subsequent evaluation.

4.2 Feed Forward Model with Feature Engineering

In the first model, a feature-based approach is used to achieve better results in the prediction of citations. This model consists of two sub models and the architecture is presented in

[4] Downloaded from https://archive.org/details/arXiv-metadata-dump-2019-06-18.tar.xz

[5] See https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-lang-analyzer.html#english-analyzer

Figure 1. The features of each passage that we want to use for our embedding are TF-IDF scores and the passage length. To have equally sized feature vectors for each of the passages, only the scores of the 100,000 most common words across all passages, sorted by the occurrences in the corpus, are used. So, each passage is represented as a vector of 100,001 dimensions, where 100,000 dimensions made up of the TF-IDF values of the words. The remaining dimension represents the passage length, which is calculated by dividing the length of the current passage by the average passage length of our corpus.

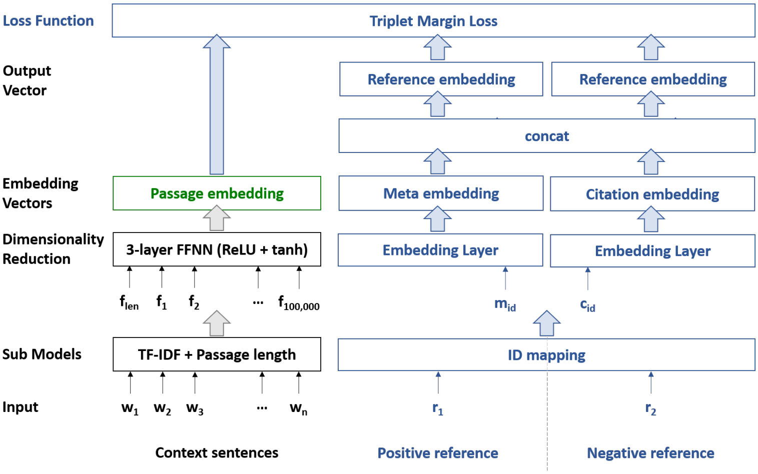

Fig. 1: Architecture of our feature based approach

We chose a three-layer Feed Forward Neural Network (FFNN) to convert the vectors down to 300 dimensions. The network consists of three linear layers, the first two are provided with a ReLU and *tanh* is utilized for the third layer. Since we use a Triplet Margin loss for training, two references per passage are needed to improve network performance. For the first, we use the correct reference featured in the original text passage and the second one is a randomly chosen reference from the corpus. The required embeddings of the references are obtained by the second sub model. The second sub model is used to encode the references. In addition, the meta-information is taken into account to improve the prediction of suitable references. As meta information, the scientific category of the cited paper is considered. For this reason, each embedding of a reference consists of a meta embedding and a citation embedding. The meta embedding refers to a feature that is shared by several references, they are grouped by this feature. The citation embedding refers to the reference itself. The reference data is used as input, which is filtered according to the required features. Based on this, identifier for metadata and the citation are determined and converted into vectors via embedding layers. Both embedding types possess their own lookup table and are concatenated for each reference. The meta embedding has a dimension of 100 and the citation embedding's dimension is 200, leading to an output embedding with a dimension of 300.

4.3 BERT Passage Model

For our third model, we use a pre-trained english BERT model [Wo19] to encode a passage. BERT (Bidirectional Encoder Representations from Transformers) is a language representation model that achieves state of the art results on different natural language processing tasks. Training a BERT model consists of two steps: pre-training and fine-tuning. In addition, BERT uses some special tokens, these are, among others, the [CLS] token and the [MASK] token. In the first step, we use a BERT tokenizer [Wo19] to convert the passage into an id sequence. As [MASK] tokens can be used to hide special tokens, we replace the references in the passages by this. Additionally, the encoded sequence has the classification token [CLS] at the beginning. Like all tokens of the sequence, this token is transformed into an embedding and corresponds to the sequence representation for different classification tasks. The converted sequence is fed into BERT to receive the word embeddings and the classification embedding of the [CLS] token of the passage. We further only use the classification embedding which is used for fine-tuning to predict references. Our first idea was to reuse the reference sub model described in the previous section. The aim was to train these sub models again by Triplet Margin Loss and fine-tune BERT, so that the encoded [CLS] embedding refers to the positive reference embedding. As this model did not perform well on a first trial dataset, we decide to use other passages for training instead of the references. The aim is to predict references using similar passages as illustrated in Figure 2.

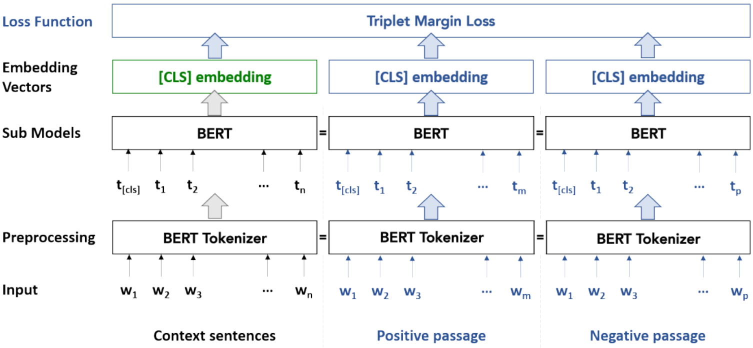

Fig. 2: Architecture of our passage-based BERT approach

For this reason, a data sample consists of a passage for which we want to predict the reference (relevant passage), a passage with the same reference (positive passage), and a passage with another reference (negative passage). As the sample requires two passages with the same reference, we only use passages if at least one other passage with the same reference exists. All passages are fed into BERT to receive the [CLS] embeddings. BERT is fine-tuned so that it learns a similarity between the relevant and the positive passage. To predict the reference of the relevant passage, the cosine similarity between the classification embedding of this passage and all other passages is calculated. The final reference is determined by the known reference of the passage that is most similar to the relevant passage.

5 Training and Evaluation

Most of citation recommendation tasks use well-known metrics such as Mean Average Precision (MAP), Recall, Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG) and hits@k for evaluation. We followed these approaches and decided to use MRR@10, hits@10 and hits@5.

5.1 Baselines

As there are many training passages, the training data was grouped by contexts and inserted one single time for each content yielding corresponding passage IDs and target references. As search query, a *More Like This Query*[6] is formed which promises a better timing performance than usual queries for large indices. The query is mapped to the content field and gets the context of a test passage as *like* statement. In addition, due to the brevity of the context, the minimum term frequency is reduced to 1 and the minimum document frequency to 3. The maximum number of query terms is decreased to 10 to further reduce response times. Using different thread workers, each query is posted separately also returning the top ten results. For each result, the proposed target references are collected and afterwards sorted by their frequency of occurrence. Finally, the topmost ten references are extracted and searched for the ground truth reference, which serves as the relevant item for calculation of MRR, hits@10 and hits@5. Table 2 illustrates the evaluation results for the reduced dataset. As the table shows, the baseline already provides respectable results.

Model	MRR@10	Hits@10	Hits@5
Elasticsearch	0.546393	0.793804	0.758596
Feature FFNN	0.000042	0.000125	0.000073
BERT Passage	0.582673	0.763704	0.713639

Tab. 2: Evaluation of the models for the reduced dataset

5.2 Feed Forward Network with TF-IDF

For training the feature embedding network we use a GeForce GTX 1080 with a batch size of 8. We train the model for three epochs using the Adam optimizer for both sub models with a learning rate of 1e−3 on the reduced dataset. We also evaluate this model by MRR@10, hits@10 and hits@5. The results of this approach are shown in Table 2. It is easy to see that the scores achieved by this approach are way lower than the other numbers. A reason for this could be that the model only takes limited context into account when looking for similar passages, mostly focusing on the reference itself. Additionally, these extremely low results may caused by the drastic reduction of the dataset due to computation time which results in having only a few passage examples per unique reference.

[6] See https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-mlt-query.html

5.3 BERT Passage Model

For fine-tuning BERT we use AdamW [LH17] optimizer with a learning rate of 1e−6. AdamW is an optimizer with decoupled weight decay is suitable for fine-tuning BERT. Due to the amount of data, we train the model on four GPUs, three GeForce RTX 2080 Ti and a GeForce GTX 1080 Ti, in parallel with a total batch size of 10 for 3 epochs on the reduced dataset. Table 2 shows the evaluation using Mean Reciprocal Rank, hits@10, and hits@5. While hits@10, and hits@5 is lower than the baseline, MRR@10 provides the best results on the dataset. When training a further epoch, it has been shown that the results of the metrics change only at the third decimal place, which shows that the model converges stably.

6 Conclusion

Recommending citations for a given query is a complex task that will keep researchers engaged for a long time. In our paper we proposed some context-based attempts that presented us with many challenges. These models are based on different approaches, which all have shown their advantages and disadvantages. We presented them in the context of our work and attempted to improve them continuously. The baseline evaluation was particularly striking, since even a simple term-based similarity already yields respectable results. Based on this, we have tried to improve these results with our models, which are based on different approaches, and to compare the results. The results of the presented models provided some surprises, especially the model using TF-IDF. The model based on learning citation embedding and meta embedding gives worse results in comparison to the Elasticsearch baseline. As mentioned in the evaluation, the worse results can be caused by different reasons and should be examined in detail in some future work. On the contrary, reference prediction by learning passage similarities using BERT embeddings finally gives the best results on MRR but provides poorer results on hits@10 and hits@5 than the baseline calculation. In general, the models using passage similarity for reference prediction provide the best results for our use case and a well-functioning citation recommendation solution.

References

[CHY18] Cai, X.; Han, J.; Yang, L.: Generative Adversarial Network Based Heterogeneous Bibliographic Network Representation for Personalized Citation Recommendation. In: Proc. 32nd AAAI Conf. on Artificial Intelligence. AAAI Press, pp. 5747–5754, 2018.

[De19] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT. 2019.

[Du16] Duma, D.; Liakata, M.; Clare, A.; Ravenscroft, J.; Klein, E.: Rhetorical Classification of Anchor Text for Citation Recommendation. D-Lib Magazine 22/, Sept. 2016.

[EF17] Ebesu, T.; Fang, Y.: Neural Citation Network for Context-aware Citation Recommendation. In: Proc. 40th Int. ACM SIGIR Conf. on Research & Development in Information Retrieval. Pp. 1093–1096, 2017.

[FS20] Färber, M.; Sampath, A.: HybridCite: A Hybrid Model for Context-Aware Citation Recommendation. arXiv preprint arXiv:2002.06406/, 2020.

[He10] He, Q.; Pei, J.; Kifer, D.; Mitra, P.; Giles, L.: Context-aware Citation Recommendation. In: Proc. 19th Int. Conf. WWW. Pp. 421–430, 2010.

[Hu12] Huang, W.; Kataria, S.; Caragea, C.; Mitra, P.; Giles, C. L.; Rokach, L.: Recommending Citations: Translating Papers into References. In: Proc. 21st ACM Int. Conf. on Information & Knowledge Management. Pp. 1910–1914, 2012.

[Hu15] Huang, W.; Wu, Z.; Liang, C.; Mitra, P.; Giles, C. L.: A Neural Probabilistic Model for Context based Citation Recommendation. In: 29th AAAI Conf. on Artificial Intelligence. 2015.

[Je19] Jeong, C.; Jang, S.; Shin, H.; Park, E.; Choi, S.: A Context-Aware Citation Recommendation Model with BERT and Graph Convolutional Networks. arXiv preprint arXiv:1903.06464/, 2019.

[Kü12] Küçüktunç, O.; Kaya, K.; Saule, E.; Çatalyürek, Ü. V.: Fast Recommendation on Bibliographic Networks. In: 2012 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining. Pp. 480–487, 2012.

[LH17] Loshchilov, I.; Hutter, F.: Decoupled Weight Decay Regularization. arXiv preprint arXiv:1711.05101/, 2017.

[Re14] Ren, X.; Liu, J.; Yu, X.; Khandelwal, U.; Gu, Q.; Wang, L.; Han, J.: ClusCite: Effective Citation Recommendation by Information Network-based Clustering./, Aug. 2014.

[SF20] Saier, T.; Färber, M.: unarXive: A Large Scholarly Data Set with Publications' Full-text, Annotated in-text Citations, and Links to Metadata. Scientometrics/, pp. 1–24, 2020.

[TWZ14] Tang, X.; Wan, X.; Zhang, X.: Cross-language Context-aware Citation Recommendation in Scientific Articles. In: Proc. 37th Int. ACM SIGIR Conf. on Research & Development in Information Retrieval. Pp. 817–826, 2014.

[Wo19] Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Brew, J.: HuggingFace's Transformers: State-of-the-art Natural Language Processing. ArXiv abs/1910.03771/, 2019.