Investigating Convolutional Networks and Domain-Specific Embeddings for Semantic Classification of Citations

Anne Lauscher
University of Mannheim
B6, 26
Mannheim 68159
anne@informatik.uni-mannheim.de

Goran Glavaš
University of Mannheim
B6, 26
Mannheim 68159
goran@informatik.uni-mannheim.de

Simone Paolo Ponzetto
University of Mannheim
B6, 26
Mannheim 68159
simone@informatik.uni-mannheim.de

Kai Eckert
Stuttgart Media University
Nobelstraße 10
Stuttgart 70569
eckert@hdm-stuttgart.de

ABSTRACT

Citation graphs and indices underpin most bibliometric analyses. However, measures derived from citation graphs do not provide insights into qualitative aspects of scientific publications. In this work, we aim to semantically characterize citations in terms of polarity and purpose. We frame polarity and purpose detection as classification tasks and investigate the performance of convolutional networks with general and domain-specific word embeddings on these tasks. Our best performing model outperforms previously reported results on a benchmark dataset by a wide margin.

CCS CONCEPTS

•**Computing methodologies → Natural language processing;** Neural networks; •**Information storage and retrieval** → *Content analysis and indexing;*

KEYWORDS

Citation Polarity, Citation Purpose, Classification, Support Vector Machine, Convolutional Neural Network, Word Embeddings

ACM Reference format:
Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. 2017. Investigating Convolutional Networks and Domain-Specific Embeddings for Semantic Classification of Citations. In *Proceedings of WOSP 2017, Toronto, ON, Canada, June 19, 2017,* 5 pages.
DOI: 10.1145/3127526.3127531

1 INTRODUCTION

Citation graphs and indices have long been supporting various analyses in the sociology of science [6, 7]. Citation graphs are used to detect research communities and retrace the evolution of ideas over time. Various measures reflecting the impact of a publication, journal, or an author exploit only raw citation counts. For example, the widely-known *h-index* [8] is commonly used to assess the scientific impact of a researcher.

Purely quantitative measures alone, however, may often be misleading regarding the positive impact of some research. For example, a publication on widely-criticized work will still have a large number of citations. Being based on simple counts, quantitative scientometric measures reflect quantitative rather than qualitative aspects of research – we are not only interested in how often a work is cited, but also why it is being cited. Knoth and Herrmannova [15] recently introduced the term *semantometrics* to describe a new category of scientometric measures that account for qualitative aspects of citations. Automated qualitative analysis of publications is challenging, as it requires processing the textual content of all citing publications. Most existing models for qualitative analysis of citations employ a range of heavily manually-engineered features.

In this work we evaluate models that require virtually no feature engineering on tasks of citation polarity and purpose classification. Citation polarity (also known as citation sentiment classification) assigning a polarity (positive, negative or neutral) to a citation, considering the citation context [2]. Citation purpose classification (also known as citation function classification) is a more fine-grained type of analysis that aims to provide a functional characterization of a citation [22].

The contributions of this work are twofold. First, following a series of successful applications of convolutional neural networks (CNNs) [17] in short text classification [12, 14], we present the first CNN application in the area of qualitative citation analysis. Using CNNs allows us to avoid extensive feature-engineering present in existing semantometric models. Secondly, we investigate the impact of using domain-specific word embeddings[1].

Experimental results on a benchmark dataset show that our best performing CNN models outperform previously reported results by a wide margin, both for polarity and purpose classification.

WOSP 2017, Toronto, ON, Canada
© 2017 ACM. 978-1-4503-5388-5/17/06…$15.00
DOI: 10.1145/3127526.3127531

[1] The domain-specific word vector representations produced in this research are available for download at: https://github.com/anlausch/scientific-domain-embeddings.

2 RELATED WORK

A significant body of work exists both for citation polarity classification [1, 2, 10, 13] and citation purpose classification [5, 22].

Athar [2] first worked on citation polarity classification, combining a range of lexical, dictionary-based, and syntactic features with a linear support vector machines (SVM) classifier. Similarly, Jochim and Schütze [10] fed a range of features for citation polarity classification to a maximum entropy classifier, whereas Kim and Thoma [13] trained an SVM model RBF kernel using occurrence statistics of n-grams in an annotated corpus as features.

Teufel et al. [22] classified function of citations into one of 12 categories. They employed a k-NN classifier using cue phrases, self-citation, and the position of the citing sentence as features. Dong and Schäfer [5] analyzed the effectiveness of different feature groups (e.g., positional, lexical, syntactic) for function classification over a range of classifiers, pointing to syntactic features as being most useful. Xu et al. [24] focused on discerning functional from perfunctory citations, using a combination of textual and external features. Abu-Jbara et al. [1] and Jha et al. [9] addressed both polarity and purpose classification with an SVM employing an extended set of features such as speculation cues and self-citation indicators. All of the above models rely on heavy manual feature design and feature engineering.

Jochim and Schütze [11] first applied a deep learning model to the citation polarity classification. In a domain-adaption setting, they trained a marginalized stacked denoising autoencoders (mSDA) on product reviews and used it to predict the polarity of citations. To the best of our knowledge, there have been no attempts to apply convolutional neural networks, achieving state-of-the-art performance on a range of text classification tasks [12, 14, 20, 21], to citation context analysis.

3 CLASSIFICATION MODELS

Our primary goal is to avoid tedious feature engineering for citation context classification. Here we describe two models that satisfy this criteria which we evaluated in our experiments.

3.1 Convolutional Neural Network

CNNs [17], introduced to the NLP community by Collobert and Weston [4], exhibit state-of-the-art performance on a range of text classification tasks [14, 20, 21]. CNN is a feed-forward neural network consisting of one or more convolution layers. Each convolution layer consists of a set of filters. When applied to textual data, convolutions of filters and text slices – matrices produced by sequentially sliding a window of size k over the embedding-based representation of text – are computed. Each convolution layer is followed by a pooling layer, which subsamples the output of the convolution layer (e.g., by taking N maximal values). This architecture allows the network to capture local aspects, i.e. the most informative k-grams in text for the task. We use a CNN with a single convolution and single max-pooling layer. We use rectified linear unit activation and optimize the network parameters with the RMSprop algorithm [23] to minimize the cross-entropy loss.

To be subdued to a CNN, texts must be represented as numerical vectors, which can be achieved by using word embeddings [18, 19, *inter alia*]. More precisely, each text is represented as a matrix of size $N \times L$, where N is the length of the text (in number of tokens) and L is the length of word embeddings. Because CNN expects the same number of features for all texts, all instances must be of equal length. In our experiments we set N to the length of the longest text in the dataset and pad all other sentences with a special padding token to which we assign a random embedding vector.

3.2 SVM with Embedding Features

Having in mind (1) that SVM has been widely used for citation polarity and purpose classification and (2) that by employing word embeddings we may still avoid manual feature engineering, we decided to compare CNNs performance to that of an SVM model using the semantic embedding of the text. We compute the embedding of the text as weighted continuous bag of words (WCBOW) aggregation of word embeddings Mikolov et al. [18]:

$$WCBOW(t_1, \ldots, t_k) = \frac{1}{\sum_{i=1}^{k} a_i} \sum_{i=1}^{k} a_i v(t_i)$$

where t_i is the i-th token of a k-token-long text, $v(t_i)$ is the word embedding of the token t_i, and a_i is the TF-IDF weight of the token, computed on the training set, which we use in order to reflect the relative informativeness of words. This results in a single aggregate embedding vector for each text, which we then feed to the SVM classifier with an RBF kernel.

3.3 General vs. Domain-Specific Word Embeddings

Both above models use word embeddings – semantic vectors that capture the meaning of words. In all our experiments, we classify texts from a specific domain of scientific publications from the area of natural language processing and computational linguistics (cf. Section 4), which is a sub-domain of all scientific publications. A research question that naturally arises is whether domain-specific word embeddings, i.e., word embeddings trained on large in-domain corpus, would lead to better classification performance than general word embeddings. In order to investigate the effects of using domain-specific embeddings, we evaluate three different variants of the above two models, employing (1) general word embeddings, (2) embeddings trained on domain corpora consisting of scientific publications from various research fields, and (3) embeddings trained on the narrowly in-domain corpus of publications from the area of natural language processing and computational linguistics.

4 DATA

We briefly describe the corpora used to train different word embeddings and the classification dataset used in our experiments.

4.1 Word Embeddings Corpora

We experimented with 50-dimensional GloVe embeddings [19] trained on three different corpora: (1) general domain Wikipedia + GigaWord corpus,[2] (2) the CORE corpus of scientific publications aggregated from Open Access repositories and journals [16], and (3) the Association for Computational Linguistics (ACL) Reference

[2] http://nlp.stanford.edu/data/glove.6B.zip

Dataset	Size (in tokens)
Wikipedia + GigaWord	6,000,000,000
CORE corpus	2,530,738,678
ACL Reference Corpus	81,365,802

Table 1: Corpora used to train word embeddings.

Classification	Label	Proportion
Polarity	*positive*	32.6%
	negative	12.4%
	neutral	55.0%
Purpose	*criticizing*	16.3%
	comparison	8.1%
	use	18.0%
	substantiating	8.0%
	basis	5.3%
	neutral	44.3%

Table 2: Dataset distributions of citation labels.

Corpus[3] [3]. We compare the sizes of these three corpora in Table 1. The CORE corpus is significantly larger than the ACL Reference Corpus, as it aggregates publications over various disciplines, whereas the ACL Reference Corpus only contains publications related to computational linguistics and natural language processing.

4.2 Citation Classification Corpus

We used the dataset from [1] and [9] in our experiments. It contains 3,271 citation context instances, each consisting of four sentences: the sentence citing a given target reference, one preceding sentence, and two following sentences. All of these contexts have been annotated for citation polarity and citation purpose. Citation polarity was annotated with one of three labels –*positive*, *negative*, and *neutral*. On the other hand, one of six categories had to be chosen as a label for citation purpose: *criticism*, *comparison*, *use*, *substantiation*, *basis*, and *neutral*. The distribution of instances over the different categories for both polarity and purpose are shown in Table 2. In addition to assigning polarity and purpose labels to citation contexts, annotators labeled each sentence of the context as being informative for the polarity and polarity classification or not. We observe that the dataset is heavily skewed towards the least informative *neutral* class for both classification dimensions.

5 EVALUATION

We describe the experimental setting, the model variants and baselines we evaluate, and the performance levels they reach.

5.1 Models and Baselines

Our primary goal was to evaluate the two models from Section 3, as models that do not require any feature design effort: CNN and SVM with aggregate text embedding. For each of these two models we evaluated three variants, using word embeddings trained on different corpora: General, CORE, and ACL (cf. Section 4). We coupled our models with the following baselines:

(1) Given the heavily skewed label distributions for both classification tasks, we use the majority class baseline predicting the most frequent class in the training set (*neutral* in both cases);

(2) We also evaluate a linear SVM with discrete TF-IDF-weighted bag-of-words features. Comparing this baseline with the embedding-based SVM model provides insights into usefulness of word embeddings for citation classification tasks;

(3) Finally, we report the performance of the SVM model with a rich set of features from [9], since they evaluate their model on the same dataset [1].

Model	P	R	F_1
Majority	18.3	33.3	23.6
Jha et al. [9]	67.1	70.6	68.8
SVM TF-IDF	77.9	76.3	77.1
SVM General emb.	79.1	74.0	76.5
SVM CORE emb.	83.2	72.1	75.3
SVM ACL emb.	81.3	75.4	77.3
CNN General emb.	82.0	75.9	**78.8**
CNN CORE emb.	81.8	76.1	**78.8**
CNN ACL emb.	81.2	75.4	78.2

Table 3: Polarity classification results.

Model	P	R	F_1
Majority	7.4	16.7	10.3
Jha et al. [9]	54.9	62.5	58.4
SVM TF-IDF	74.3	70.9	72.6
SVM General emb.	86.8	64.7	74.1
SVM CORE emb.	81.7	66.2	73.1
SVM ACL emb.	81.7	66.0	73.0
CNN General emb.	79.9	68.2	73.6
CNN CORE emb.	80.8	68.8	**74.3**
CNN ACL emb.	76.7	68.4	72.3

Table 4: Purpose classification results.

5.2 Experimental Setting

In order to make our results comparable to those reported in [9], we evaluate the models in 10-fold cross validation (CV) setting. More precisely, for each model we execute a nested CV evaluation, where for each fold of the outer CV loop, we optimize model's hyperparameters via grid search in the inner CV. The reported performance is macro-averaged over the folds of the outer CV loop.

5.3 Results

Polarity classification results are shown in Table 3 and purpose classification results in Table 4. Surprisingly, the linear SVM with bag-of-word features is a very competitive baseline on both classification tasks. More surprisingly, it performs 8% (polarity) and 14% (purpose) better than the SVM model from Jha et al. [9], which

[3] Version 20160301, ParsCit structured XML.

Classification	Model	Context	P	R	F_1
Polarity	CNN CORE emb.	Citing Sentence	81.8	76.1	78.8
	CNN CORE emb.	Gold Standard	85.8	78.7	**82.1**
	SVM CORE emb	Citing Sentence	83.2	72.1	75.3
	SVM CORE emb.	Gold Standard	84.1	75.6	79.6
Purpose	CNN CORE emb.	Citing Sentence	80.8	68.8	74.3
	CNN CORE emb.	Gold Standard	85.2	73.3	**78.9**
	SVM CORE emb.	Citing Sentence	81.7	66.2	73.1
	SVM CORE emb.	Gold Standard	84.8	69.2	76.2

Table 5: Impact of the choice of the citation context on the classification results.

uses a much richer set of features. This is probably because Jha et al. [9], reportedly, do not optimize their model's hyperparameters. Also, the SVM models with embedding features do not outperform the linear SVM baseline, regardless of the corpus used to train the embeddings. All this suggests that citation polarity and purpose are strongly indicated by a particular set of lexical clues.

The CNN model has a slight edge over all SVM-based models, but the performance gains performance are much lower than for other text classification tasks [14, 21]. In-domain specialization of embeddings does not seem to play a significantly positive role. The best results are obtained using the super-domain CORE embeddings. The in-domain ACL embeddings are probably of lower quality due to much smaller size of the ACL Reference Corpus.

Table 5 shows the classification results of SVM and CNN with CORE embedding features when using different context sizes. As it can be seen, for all models the performance improves by around 3% to 4% when the gold standard citation context is taken into account instead of only the directly citing sentence. This suggests that a fine grained identification of the citation context is an important step that needs to precede the classification tasks at hand.

When analyzing the results in depth we noticed that for both classification tasks most errors that happened correspond to a misclassification of a citation context into the category *neutral*. This type of error occurred in 61% of all the misclassifications that happened in the purpose classification and in 59% of the errors which occurred when classifying polarity. This may be due to the skewness of the benchmark dataset we used. Another frequent error that happened in the purpose classification is the misclassification of an instance of the category *basis* as *use*, which is probably due to the high interrelation of those two purposes. Similarly, all purpose classifiers often confused the instances of the class *comparison* with instances of the class *criticizing*.

6 CONCLUSION

Existing models for semantic classification of citations rely on extensive feature engineering. In this work, we investigated two models that do not require any manual feature design – CNN and SVM with aggregate text embeddings – on citation polarity and citation purpose classification tasks. The investigated models outperform previously reported results on a benchmark dataset by a wide margin. However, only CNN models slightly outperform a simple linear SVM with lexical features. This suggests that lexical clues alone quite strongly indicate citation polarity and purpose. We also find that using domain-specific word embeddings provides no observable performance boost.

ACKNOWLEDGMENTS

This research was partly funded by the German Research Foundation (DFG), project number EC 477/5-1 (LOC-DB). We thank the NVIDIA Corporation for donating the GeForce Titan X GPU used to carry out our experiments.

REFERENCES

[1] Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics (ACL), 596–606.

[2] Awais Athar. 2011. Sentiment Analysis of Citations Using Sentence Structure-Based Features. In *Proceedings of the ACL 2011 Student Session (HLT-SS '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 81–87.

[3] Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*. 1755–1759.

[4] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML-08)*, William W. Cohen, Andrew Mccallum, and Sam T. Roweis (Eds.). 160–167.

[5] Cailing Dong and Ulrich Schäfer. 2011. Ensemble-style Self-training on Citation Classification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 623–631.

[6] Eugene Garfield. 1955. Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science* 122, 3159 (July 1955), 108–111.

[7] E Garfield, Irving Sher, and RJ Torpie. 1984. *The Use of Citation Data in Writing the History of Science.* Institute for Scientific Information Inc.

[8] J. E. Hirsch. 2005. An Index to Quantify an Individual's Scientific Research Output. *Proceedings of the National Academy of Sciences* 102, 46 (Nov. 2005), 16569–16572. arXiv:physics/0508025

[9] Rahul Jha, Amjad-Abu Jbara, Vahed Qazvinian, and Dragomir R. Radev. 2017. NLP-Driven Citation Analysis for Scientometrics. *Natural Language Engineering* 23, 1 (Jan. 2017), 93–130.

[10] Charles Jochim and Hinrich Schütze. 2012. Towards a Generic and Flexible Citation Classifier Based on a Faceted Classification Scheme. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*.

[11] Charles Jochim and Hinrich Schütze. 2014. Improving Citation Polarity Classification with Product Reviews. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*. 42–48.

[12] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A Convolutional Neural Network for Modelling Sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 655–665.

[13] I. C. Kim and G. R. Thoma. 2015. Automated Classification of Author's Sentiments in Citation Using Machine Learning Techniques: A Preliminary Study. In *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*.

[14] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751.

[15] Petr Knoth and Drahomira Herrmannova. 2014. Towards Semantometrics: A New Semantic Similarity Based Measure for Assessing a Research Publication's Contribution. *D-Lib Magazine* 20, 11/12 (Nov. 2014).

[16] Petr Knoth and Zdenek Zdrahal. 2012. CORE: Three Access Levels to Underpin Open Access. *D-Lib Magazine* 18, 11/12 (2012).

[17] Yann LeCun and Yoshua Bengio. 1998. The Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge, MA, USA, Chapter Convolutional Networks for Images, Speech, and Time Series, 255–258.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[19] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*. 1532–1543.

[20] Aliaksei Severyn and Alessandro Moschitti. 2015. Twitter Sentiment Analysis with Deep Convolutional Neural Networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 959–962.

[21] Prasha Shrestha, Sebastian Sierra, Fabio A González, Paolo Rosso, Manuel Montes-y Gómez, and Thamar Solorio. 2017. Convolutional Neural Networks for Authorship Attribution of Short Texts. In *Proceedings of the 2017 Conference of the European Chapter of the Association of Computational Linguistics (EACL 2017)*.

[22] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic Classification of Citation Function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 103–110.

[23] Tijmen Tieleman and Geoffrey Hinton. 2012. *Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude.* Technical Report 2.

[24] Han Xu, Eric Martin, and Ashesh Mahidadia. 2013. Using heterogeneous features for scientific citation classification. In *Proceedings of the 13th conference of the Pacific Association for Computational Linguistics*.