Citation Intent Classification

Pavan Mandava
st169661@stud.uni-stuttgart.de
mspavan04@gmail.com
Mat. Nr.: 3461015

Isaac Riley
st166211@stud.uni-stuttgart.de
isaac.r.riley@gmail.com
Mat. Nr.: 3018413

Abstract

Knowledge discovery in large scientific databases relies heavily on automatic processing of scientific writing. Because of the network structure inherent in scientific work, it is of interest to learn as much as possible about the citations themselves. In this paper, we present a method for classifying citation purpose as primarily pertaining to background, methods, or results. We use a BiLSTM model with ELMo embeddings to achieve a macro-F1 score of 0.837 on the SciCite dataset. Our code is available at https://github.com/yelircaasi/citation-analysis

1 Introduction

The pace and scale of scientific research are such that it is impossible for any individual, or even team, to read everything that is published in a given domain. For this reason, it is of interest to employ automatic methods in order to maintain an overview of the field and to identify the most relevant and interesting publications. One of the key features of scientific writing is the dense structure of direct citations between papers, which is a valuable asset for automatic knowledge discovery methods.

Many such methods exploit the network structure inherent in scientific research. Because a typical scientific paper cites many other papers, much information can be learned about a paper simplify by analyzing which papers it cites and, in turn, in which papers it is itself cited (Leydesdorff, 1998). Clearly, not all citations are identical in nature. The information about a paper gleaned from its citation profile can be augmented by taking into account the nature of each citation (Teufel et al., 2006). For example, a paper that is typically cited as a background source plays a different role within the literature than a paper that tends to be cited as a methodological standard, which is in turn qualitatively different from papers noted primarily for their results. Scientific research can be conceptualized as a graph, and edges in the graph of scientific papers are not all qualitatively the same, so the more that can be said about the nature of each, the more interesting information can be learned from this graph (Abu-Jbara et al., 2013). Thus, the ability to classify the nature of a citation is of interest for contexts for fields such as scientometrics (Li and Ho, 2008), knowledge discovery (Guo et al., 2009), information retrieval (Ritchie, 2009), and many more; this line of research has grown together from the initially independent work of many fields and has become highly interdisciplinary (White, 2004).

While much information can be gleaned from citations using rule-based methods, automatic classification of citation type is important for publication types where sections (esp. Background, Methods, Results) are not explicitly labeled; hence the need for automatic classification. This project attempts to use methods from machine learning to classify citations according to the purpose for which the paper is cited. Specifically, in the SciCite dataset, a paper is cited because it provides relevant background information, because it pioneered methods used in the citing paper, or to compare results from experiments similar to those carried out in the citing paper. This setup is advantageous because the inherent structure of scientific papers allows for the efficient creation of large datasets. Some approaches use as many as 8 distinct categories (Agarwal et al.); however, this paper deals with three, in part because these three categories provide a solid foundation on top of which other categories can eventually be learned. As will be shown, the classification even three cat-

egories presents a non-trivial learning task.

2 Methods

2.1 Baseline Perceptron Model

Our baseline is a simple linear perceptron classifier. It is used with a pre-defined set of hand-crafted features of "cue phrases", as in Pham and Hoffmann (2003). They are listed below and are generally self-explanatory:

Feature	Regular Expression?
COMPARE	
CONTRAST	
RESULT	
INCREASE	
CHANGE	
PRESENT	
IMPORTANT	
RESEARCH	
APPROACH	
PUBLIC	
BEFORE	
BETTER_SOLUTION	
PROFESSIONALS	
MEDICINE	
MATH	
COMPUTER_SCIENCE	
CITATION	
ACRONYM	✓
CONTAINS_YEAR	✓
SEQUENCE	✓
REFERENCE	✓
PERCENTAGE	✓
CONTAINS_URL	✓
ENDS_WITH_RIDE	✓
ENDS_WITH_RINE	✓
ENDS_WITH_ETHYL	✓

On the basis of these features, the perceptron performs a binary classification for each class, and the class with the highest predicted score is predicted as the label for a given sample.

2.2 Simple Feed-forward Neural Network

The next experiment was a feed-forward neural classifier with a single hidden layer containing 9 units. While a feed-forward neural network is clearly not the ideal architecture for sequential text data, it was of interest to add a sort of second baseline and examine the added gains (if any) relative to a single perceptron. The input to the feed-forward network remained the same; only the final model was suitable for more complex inputs such as word embeddings. The optimizer used during training is the stochastic gradient descent, with batch size 16.

2.3 BiLSTM-Attention with ELMo

The most advanced and best-performing model was a singe-layer Bi-directional LSTM neural network (Hochreiter and Schmidhuber, 1997) with a hidden layer of dimension 50 for each direction, built using the `allennlp` library (Gardner et al., 2017). For word representations, it uses the 100-dimensional GLoVe embeddings developed by Pennington et al. (2014). To represent word contexts, it makes use of the ELMO embeddings developed at the Allen Institute (Peters et al., 2018), which contain "deep contextualized word representations". This model uses the entire input text, as opposed to selected features in the text, as in the first two models. The optimizer used during training is AdaGrad, and as in the second model, batch size is 16.

3 Experiments

3.1 Experimental Setting

We train and test our models on the SciCite dataset (Cohan et al., 2019). Each data sample consists of a sentence from a scientific paper containing a citation. The Background class contains citations pointing to papers that provide context information for the citing paper, e.g.:

"However, how frataxin interacts with the Fe-S cluster biosynthesis components remains unclear as direct one-to-one interactions with each component were reported (IscS [12,22], IscU/Isu1 [6,11,16] or ISD11/Isd11 [14,15])."

The Methods class contains citations of papers that pioneered or established a key method, tool, or dataset that is used or expanded upon in the citing paper. For example:

Mouse embryonic fibroblasts (MEFs) were also infected with EOS (early transposon promoter and Oct-4 and Sox2 enhancers) lentiviral vector selection system.

Finally, the Results class contains citations of papers whose results are compared with those in the citing paper, such as the following:

In addition, the result of the present study supports previous studies, which did not find increased rates of first-born children among individual with OCD (20,31,34).

The following table shows the class distribution in each dataset (in thousands):

	Training	Development	Testing
Background	4.8	0.5	1.0
Methods	2.3	0.3	0.6
Results	1.1	0.1	0.2
Total	8.2	0.9	1.8

3.2 Results

	Acc	F1	Rec	Prec	Class
Perc.	0.62	0.49	0.76	0.66	B
			0.57	0.53	M
			0.14	0.62	R
FFNN	0.60	0.59	0.56	0.75	B
			0.58	0.64	M
			0.82	0.38	R
BiLSTM	0.85	0.84	0.88	0.87	B
			0.80	0.87	M
			0.85	0.75	R

F1-scores reported are obtained through macro-averaging across classes. Accuracy and F1-score are for the model as a whole, while precision and recall are reported at class-level.

3.3 Error Analysis

3.3.1 Perceptron

		Predicted		
		B	M	R
Actual	Background	762	221	14
	Methods	250	347	8
	Results	138	85	36

The Results class was strongly underpredicted, with the Background class being strongly overpredicted. Manual inspection of misclassified samples revealed that many of them can be correctly identified by a human. However, there were some samples which were difficult even for humans to identify correctly. In general, while humans would be likely to outperform even the state-of-the-art models, they would be unlikely to achieve an accuracy of 100%.

3.3.2 Feed-forward network

		Predicted		
		B	M	R
Actual	Background	558	184	255
	Methods	157	352	96
	Results	32	15	212

Among misclassified samples, the most conspicuous are those belonging to Results and incorrectly classified as Background. This model predicts far more samples as Results and fewer as Background compared to the first model. Manual inspection of these misclassified samples reveals little in addition to what was already reported.

3.3.3 BiLSTM

		Predicted		
		B	M	R
Actual	Background	882	63	52
	Methods	101	483	21
	Results	32	6	221

The distribution of misclassified samples does does reveal any surprising patterns; the largest class has the most false positives and the most false negatives, but not proportionately so. Confusion between Background and methods accounts for a majority of misclassifications, but once again, this is not proportional. More interestingly, manual inspection of misclassified samples revealed that some of them are still easily classified by humans, while some are essentially indistinguishable.

4 Conclusion

The single most striking aspect of our results is the clear superiority of a BiLSTM model. That it should achieve superior results is perhaps unsurprising, given that its structure is much better suited to the problem and it allows for the use of the entire text with additional semantic information in the form of word embeddings. However, the magnitude by which it outperforms the other models is striking.

Between the two baseline models, many of the gains made in some metrics are offset by losses in others; in particular, the trade-off between precision and recall is apparent. For example, significantly higher recall for the Results class comes at

the expense of its precision, while the reverse is true for the Backgrounds class. Interestingly, the second model obtains higher results for the Methods class in both precision and recall. These differences can be largely explained by the differences in how these models were trained. The perceptron used binary classification for each class and used all of the data for training, causing the model to over-predict the majority class and under-predict the minority classes. The feed-forward network uses a balanced subset of the data and predicted classes simultaneously, leading to higher precision and lower recall for the majority class and higher recall and lower precision for the smallest class.

There a few things we would have liked to investigate but did not due to time constraints. One would have been a progressive sequence of models allowing us to attribute improvements in performance to specific alterations in the model. For example, such a sequence might be:

perceptron →
simple feed-forward network →
feed-forward network with hidden layer →
LSTM →
BiLSTM →
BiLSTM with embeddings.

Such an approach would provide useful insights into the relative value added due to each aspect of the final model and may even provide hints as to the most promising directions for future improvements.

Another interesting question that is raised by our results regards the effect of class sampling during training. Balanced classes in training the feed-forward network brought some improvements, but was also sub-optimal. It is therefore of interest to further investigate the effect of class representation during training to determine an optimal sampling approach.

Additionally, we would have liked to add more complex features to our feature set. For example, features derved from syntactic/dependency parsing would allow us to generate interesting complex features taking into account relationships between words. While it seems plausible that they could have improved the model, this is something that needs to be examined empirically.

There are some interesting possibilities for future research in this field. One obvious extension would be to add more fine-grained categories or learning the valence (degree of positivity/negativity) of a citation. This would add some very important information to the citation networks generated, since some citations imply a higher degree of approval than others, and some are even cited for purposes of criticism.

Another valuable line of inquiry would regard the degree to which these methods generalize to other scientific fields, or even to other types of writing. For example, texts in the social sciences are likely to follow somewhat different conventions of citation and discussion, and naturally the vocabulary will differ as well. Other types of writing, such as journalism or essays, will also involve citations, but of a different kind, at it would be of interest to see how much modification to the methods presented here is required.

5 Contributions

Each of us contributed to the perceptron. Pavan coded the metrics, with some contributions from Isaac. Pavan created the lexicons and Isaac worked on the regex features. Isaac worked on the feed-forward neural network. Pavan created the basic network structure for the AllenNLP model and ran the experiments on GPU. Both worked together on code documentation, and Pavan worked on the README documentation. Isaac performed error analysis on the results. Both worked together closely on writing and editing the paper; Isaac was in charge of LaTeX typesetting.

Acknowledgments

Many thanks to Laura Bostan for her patient mentoring, kind encouragement, and thoughtful feedback. Thanks also to Roman Klinger for his clear instructions and helpful guidance.

References

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *NAACL-HLT*.

Shashank Agarwal, Lisha Choubey, and Hong Yu. Automatically classifying the role of citations in biomedical articles.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Z. Guo, Z. Zhang, S. Zhu, Y. Chi, and Y. Gong. 2009. Knowledge discovery from citation networks. In *2009 Ninth IEEE International Conference on Data Mining*, pages 800–805.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*.

Loet Leydesdorff. 1998. Theories of citation? *Scientometrics*.

Zhi Li and Yuh-Shan Ho. 2008. Use of citation per publication as an indicator to evaluate contingent valuation research. *Scientometrics*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Son Bao Pham and Achim Hoffmann. 2003. A new approach for scientific citation classification using cue phrases. pages 759–771. Springer.

Anna Ritchie. 2009. Citation context analysis for information retrieval. *Technical report, University of Cambridge, Computer Laboratory*.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. *EMNLP '06: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103—-110.

Howard D. White. 2004. Citation analysis and discourse analysis revisited. *Applied linguistics*, 25(1):89–116.