A preliminary study to compare deep learning with rule-based approaches for citation classification

Julien Perier-Camby[1], Marc Bertin[2][0000−0003−1803−6952], Iana Atanassova[3][0000−0003−3571−4006], and Frédéric Armetta[1]

[1] Laboratoire LIRIS, Université Claude Bernard Lyon 1, France
julien.perier-camby@etu.univ-lyon.fr - frederic.armetta@univ-lyon1.fr
[2] Laboratoire ELICO, Université Claude Bernard Lyon 1, France
marc.bertin@univ-lyon1.fr
http://www.elico-recherche.eu/membres/marc-bertin
[3] CRIT, Université de Bourgogne Franche-Comté, France
iana.atanassova@univ-fcomte.fr

Abstract. Categorization of semantic relationships between scientific papers is a key to characterize the condition of a research field and to identify influential works. Recently, new approaches based on Deep Learning have demonstrated good capacities to tackle Natural Language Processing problems, such as text classification and information extraction. In this paper, we show how deep learning algorithms can automatically learn to classify citations, and could provide a relevant alternative when compared with methods based on pattern extractions from the recent state of the art. The paper discusses their appropriateness given the requirement of large datasets to train neural networks.

Keywords: Biattentive Classification Network · Citation Classification · Citation Analysis · Citation Contexts.

1 Introduction

The categorization of semantic relationships is at the very heart of bibliometrics and Natural Languages Processing research. As described by Garfield more than 50 years ago [7], understanding how scholars use and frame citations is an essential prerequisite to characterize the state of a scientific field and to identify influential works. The research on citation acts has already proposed numerous empirical studies and models, in particular through the production of ontologies such as CiTO (see [14, 5]) or studies on the analysis of sentiments applied to the context of citations [3, 11].

Most of the studies in this field rely on Rule-based Information Extraction in order to categorize and semantically annotate citation acts. The general idea of such approaches in Natural Language Processing is to propose a categorization of citation contexts through the identification of patterns or text structures [16,

CEUR-WS.org/Vol-2345/paper11.pdf

9, 2, 10, 1]. Nevertheless, the declarative nature of rule-based approaches leads to drawbacks and tends to be replaced by machine learning alternatives [4].

To our knowledge, deep learning methods have not yet been applied to categorize citations in texts, i.e. to determine a class for each of the citation acts. The reasons for this are mainly because few datasets are publicly available, and because they tend to be small and unbalanced, making them difficult to use for the development of deep learning approaches. If we consider the progress enabled by deep learning in any domains, it is nevertheless interesting to show how deep learning approaches behave within this innovative context.

In this paper, we aim to compare the most efficient rule-based approach from the state of the art used for categorization [8] to a famous deep learning approach well-known for its ability to identify sentence meanings [15]. In section 2, we describe how rule-based approaches have been applied to categorize citations and the main principles of deep learning approaches. We discuss the advantages and drawbacks for both approaches, and underline the challenges in training a neural network with a dataset that is limited in size and unbalanced between labeled categories or classes. This section introduces the Biattentive Classification Network (BCN, [12]) combined with Embeddings from Language Models (ELMo, [15]) word representations that we experiment. Section 3 introduces the corpus and the protocol that we use for evaluation. The results are presented and discussed in section 4. The conclusion is presented in section 5.

2 Categorization of Semantic Relationships

In this section, we describe and discuss the general functioning of rule-based and deep learning approaches and their requirements. Rather that giving the detailed description of each of the approaches, we present their general properties for the sake of comparison.

2.1 Rule-based information extraction

In rule-based approaches, one has to define a set of discourse features which can be relevant to characterize the sentences semantics dedicated to different scopes. A state-of-the-art method for rule-based information extraction applied to citation framing has been proposed by [8], using pattern-based features, topic-based features and prototypical argument features. As a final step, a training phase is used to weight the relevance of each of the available patterns depending on the class to predict. This is usually done through shallow machine learning models (for instance, k-nearest neighbors [17] or random forest [8]). Such models require smaller sizes of training datasets to provide satisfying results, compared to deep neural networks.

It should be noted that rule-based methods suffer only slightly from unbalanced datasets as the features are hand-crafted, and therefore inferred on wider knowledge and not limited to the sample in the training dataset. If a citation class is under represented, the classifier could still capture part of the meaning,

as the knowledge used for the capturing is provided by an expert. Thus, the lack of balance in the dataset, only slightly degrades the classifier learning.

2.2 Deep learning information extraction

Deep learning algorithms are artificial neural networks that learn to perform tasks by learning from samples. For the specific problem we address, the network takes as input some selected characteristics of the citation and learns to give as an output the appropriate prediction (citation class). The efficiency of such algorithms does not rely on any task-specific rules, but rather benefits from non linear functions dedicated to capture complex patterns during the learning phase in order to produce a model capable of categorizing new samples.

Deep learning algorithms are highly sensitive to the quality of the training data as they do not rely on any external knowledge. As for any machine learning algorithm, the training data should be as balanced as possible, i.e. the variables have to be independent and identically distributed, and the training dataset should be large enough for the system to learn. For the so addressed problem, we need a dataset that is large and balanced across the different citation classes. In fact, if a citation class is underrepresented in the dataset, its characteristics will need to be extracted from a smaller number of samples and the inference mechanism will provide sub-optimal results.

For the purpose of comparison, we have selected the BCN model (Biattentive Classification Network, [12]) designed to handle sentence classification tasks. ELMo (Embeddings from Language Models, [15]) is designed to extract word representations, and can be used to encode sentences to pass through classifiers. BCN complemented by ELMo is the current state of the art on fine-grained (five-class) sentiment classification (SST-5, Stanford Sentiment Treebank). It is one of the best available algorithms from the state of the art for inference in text understanding.

3 Method and experimental setup

The dataset that we use for the training and the evaluation of the BCN model is the one used in [8]. This dataset has been fully annotated manually, which makes it particularly accurate to study the ability of an automatic classifier to imitate human performances. Table 1 presents the six classes used for the labelling of citations.

In order to underline the citation act to classify, every in-text reference is replaced in turn by a marker ('[X]'). The so formatted sentences (one marker for each sentence) are passed through the neural network for inference. We used in this paper the BCN model implemented by the AllenNLP library [6], which is a high-level framework built on PyTorch[13].

The evaluation has been done using k-fold cross-validation, with $k = 10$, for the learning and testing of the network to provide statistically significant results.

Table 1. Scheme used for the labelling of citations, extracted from [8]

Class	Description
BACKGROUND	Provides relevant information for this domain. e.g. "This is often referred to as incorporating deterministic closure (Dörre, 1993)."
MOTIVATION	Illustrates need for data, goals, methods, etc. e.g. "As shown in Meurers (1994), this is a well-motivated convention [...]"
USES	Uses data, methods, etc. e.g. "The head words can be automatically extracted [...] in the manner described by Magerman (1994)."
EXTENSION	Extends data, methods, etc. e.g. "[...] we improve a two-dimensional multimodal version of LDA (Andrews et al, 2009) [...]
COMPARISON OR CONTRAST	Expresses similarity/differences. e.g. "Other approaches use less deep linguistic resources (e.g., POS-tags Stymne (2008)) [...]"
FUTURE	Is a potential avenue for future work. e.g. "[...] but we plan to do so in the near future using the algorithm of Littlestone and Warmuth (1992)."

The original samples have been randomly partitioned into 10 equally sized subsamples. The learning has been performed on 9 subsamples and tested on the remaining one for each of the combinations. The reported results correspond to the average results over the 10 training sessions[4].

$Micro-F1$ and $Macro-F1$ scores are used to report the global efficiency of the network for each class, where $Micro-F1$ stands for the weighted arithmetic average and $Macro-F1$ stands for the non-weighted arithmetic average of the $F1$ score for each class.

4 Results and discussion

4.1 Global results

The selected deep learning and rule-based approaches performances are presented on table 2. Jurgens et al. [8] only reports the $Macro-F1$ metric as a base for comparison. Because of their rarity, accurate samples of significant size for such a study are difficult to acquire and this can be a major obstacle to clearly identifying the potential of deep learning approaches for citation categorization.

[4] The source code of the approach presented here is available on GitHub : https://github.com/jperier/BIR2019_citationBCN

Table 2. Experimental results

Approach	$Macro - F1$	$Micro - F1$
BCN + ELMo (2018)	0.405	0.588
Jurgens et al. (2018)	0.530	-

4.2 Results by class

Table 3 presents the results of the $F1$ score and the sample sizes for the different classes. One can note that some classes are more challenging to predict than others with the BCN model.

The classes are highly imbalanced in the dataset. For this reason, we consider the $Micro - average$ metric, which aggregates the contributions of all classes in an average value.

For each class, the reported efficiency clearly correlates with the size of the available data. While rule-based approaches become effective mainly thanks to expert knowledge, neural networks are completely dependant on samples for their learning. As a result, it is not surprising that the model performs poorly for small samples, as in the case of the class "FUTURE". Under-represented classes pull down the $Macro - F1$ value, and sightly influence the $Micro - F1$ value. On the other hand, the classifier performs well for larger samples, e.g. the classes "BACKGROUND" and "USES" with $F1$ scores of 0.720 and 0.640.

Table 3. F1 score reported by class for BCM and ELMo

Class	F1 score	Sample size
BACKGROUND	0.720	1000
MOTIVATION	0.306	185
USES	0.640	823
EXTENSION	0.103	152
COMPARISON OR CONTRAST	0.570	857
FUTURE	0.093	70
$Micro - average$	0.588	3087
$Macro - average$	0.405	3087
Random	0.138	3087

5 Conclusion

The paper describes the citation classification which is a central problem leading to many applications in bibliometrics. In this work, we are interested in studying deep learning abilities to capture the semantics of citations when compared with rule-based approaches. To do so, we compare two approaches from the recent state of the art.

We still can not define an upper bound for the application of deep learning approaches to citation classification because the experiment is based on a limited dataset compared to the datasets generally used in deep learning. New datasets need to be created to delineate more precisely the $F1$ score that can be reached by such approaches. The results encourage the use of neural networks for the cases where large samples are available. In the cases when large samples are not available, it is clear that efforts invested into rule-based approaches prove reliable and can guarantee more accurate output.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

1. Abu-Jbara, A., Ezra, J., Radev, D.: Purpose and polarity of citation: Towards nlp-based bibliometrics. In: The 2013 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies (NAACL-HLT 2013). pp. 596–606. Westin Peachtree Plaza Hotel Atlanta, Georgia, USA (Jun 2013)
2. Aljohani, N.R., Nawaz, R.: Mining the context of citations in scientific publications. In: Maturity and Innovation in Digital Libraries: 20$^{\text{th}}$ International Conference on Asia-Pacific Digital Libraries, Hamilton, New Zealand, Nov. 19–22. p. 316 (2018)
3. Chali, Y., Hasan, S.A.: Towards automatic topical question generation. In: Proceedings of COLING 2012. pp. 475–492. The COLING 2012 Organizing Committee (2012), http://aclweb.org/anthology/C12-1030
4. Chiticariu, L., Li, Y., Reiss, F.R.: Rule-based information extraction is dead! long live rule-based information extraction systems! In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2013). pp. 827–832. Association for Computational Linguistics, Grand Hyatt SeattleSeattle, Washington, USA (Oct 2013)
5. Ciancarini, P., Di Iorio, A., Nuzzolese, A.G., Peroni, S., Vitali, F.: Evaluating citation functions in cito: Cognitive issues. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) The 11$^{\text{th}}$ conference proceedings for Semantic Evaluation Challenge 2014 (ESWC2014) - The Semantic Web: Trends and Challenges. pp. 580–594. Springer International Publishing, Anissaras, Crete, Greece. (May 2014)
6. Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., F. Liu, N., Peters, M., Schmitz, M., Zettlemoyer, L.: Allennlp: A deep semantic natural language processing platform pp. 1–6 (2018), http://aclweb.org/anthology/W18-2501
7. Garfield, E., et al.: Can citation indexing be automated ? In: Statistical association methods for mechanized documentation, symposium proceedings. vol. 269, pp. 189–192. National Bureau of Standards (1965)
8. Jurgens, D., Kumar, S., Hoover, R., McFarland, D., Jurafsky, D.: Measuring the evolution of a scientific field through citation frames. Transactions of the Association for Computational Linguistics **6**, 391–406 (2018), https://transacl.org/ojs/index.php/tacl/article/view/1266

9. Kim, I.C., Thoma, G.R.: 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). pp. 1–7. Niagara Falls, Canada (Aug 2015). https://doi.org/10.1109/CIBCB.2015.7300319
10. Lamers, W., van Eck, N.J., Waltman, L., Hoos, H.: Patterns in citation context: the case of the field of scientometrics. In: 23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, 2018, Leiden, The Netherlands. Centre for Science and Technology Studies (CWTS) (2018)
11. Ma, Z., Nam, J., Weihe, K.: Improve sentiment analysis of citations with author modelling. In: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 122–127. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/W16-0420, http://aclweb.org/anthology/W16-0420
12. McCann, B., Bradbury, J., Xiong, C., Socher, R.: Learned in Translation: Contextualized Word Vectors. The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS) (Aug 2017)
13. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: The Thirty-first Annual Conference on Neural Information Processing Systems (NeurIPS). The Neural Information Processing Systems Foundation, Long Beach Convention Center, CA, USA (Dec 2017), https://openreview.net/pdf?id=BJJsrmfCZ
14. Peroni, S., Shotton, D.: Fabio and cito: ontologies for describing bibliographic resources and citations. Web Semantics: Science, Services and Agents on the World Wide Web **17**, 33–43 (2012)
15. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018) (2018)
16. Pham, S.B., Hoffmann, A.: A new approach for scientific citation classification using cue phrases. In: Gedeon, T.T.D., Fung, L.C.C. (eds.) AI 2003: Advances in Artificial Intelligence. pp. 759–771. Springer Berlin Heidelberg, Berlin, Heidelberg (2003)
17. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2006). pp. 103–110. Association for Computational Linguistics (2006)