Neural Multi-Task Learning for Citation Function and Provenance

Xuan Su, Animesh Prasad, Min-Yen Kan, Kazunari Sugiyama
School of Computing, National University of Singapore, Singapore
{suxuan,animesh,kanmy,sugiyama}@comp.nus.edu.sg

ABSTRACT

Citation function and provenance are two cornerstone tasks in citation analysis. Given a citation, the former task determines its rhetorical role, while the latter locates the text in the cited paper that contains the relevant cited information. We hypothesize that these two tasks are synergistically related, and build a model that validates this claim. For both tasks, we show that a single-layer convolutional neural network (CNN) outperforms existing state-of-the-art baselines. More importantly, we show that the two tasks are indeed synergistic: by jointly training both tasks using multi-task learning, we demonstrate additional performance gains.

CCS CONCEPTS

• **Information systems → Retrieval tasks and goals**; **Clustering and classification**.

KEYWORDS

Citation Analysis, Multi-Task Learning, Neural Networks

ACM Reference Format:
Xuan Su, Animesh Prasad, Min-Yen Kan, Kazunari Sugiyama. 2019. Neural Multi-Task Learning for Citation Function and Provenance. In *Proceedings of ACM/IEEE Joint Conference on Digital Library (JCDL'19).* ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

In academia, citations are used to acknowledge the intellectual credit to prior sources of knowledge. But why do authors make citations? What kind of rhetorical role do these citations play? In order to answer these questions, we introduce the first task of interest: citation function classification, where a computer system assigns one out of a set of predefined rhetorical roles to a given citation.

Readers are not bound by the frame of the citing paper; they may trace through the citation to the cited paper containing the purported cited information. Thus, the second task: citation provenance identification, the task of identifying the cited information in the cited paper, corresponding to a given citation.

Most existing approaches to the above two tasks employ conventional machine learning models. These methods require manual effort to obtain rules or features. Recent advances in deep learning bypass the cumbersome engineering processes: current models can represent a large class of features implicitly without laborious engineering. Our work examines the two citation analysis tasks under the auspice of deep learning. No attempt has been made to examine the pair of tasks in tandem. Intuitively, there is a correlation between the function and provenance labels for a given citation. Knowing the function of a given citation may then help determine whether a text fragment is its provenance, and vice versa. The implicit relationship between citation function and provenance motivates us to attempt multi-task learning (MTL) [2] for both tasks.

We apply a single-layer convolutional neural network (CNN) to both citation function and provenance. We then use a multi-task learning architecture on top of the CNN structure to obtain improved performance, verifying our hypothesis that the tasks are synergistic.[1]

2 DATASET

For **Citation Function**, we adopt the classification scheme in [6]. This scheme consists of the following four classes: (*Weak*)ness, Compare and Contrast (*CoCo*), (*Pos*)itive, and (*Neut*)ral. For example, a label of *CoCo* means that the citation compares the method or evaluation results in the cited paper with the current paper. To construct our own dataset, we perform manual data annotation. Citation contexts are taken from randomly-sampled articles in the ACL Anthology Reference Corpus [1]. We use the CrowdFlower platform to crowdsource annotations. Our final dataset consists of a total of 1,432 instances, out of which 1,011 are *Neut*, 295 are *Pos*, 95 are *CoCo* and 31 are *Weak*.

Citation Provenance. We use a binary classification scheme, *Prov* and $-$*Prov*. Given a citation context, a cited paper text fragment is labeled *Prov* if it contains evidence for the cited information; it is classified $-$*Prov* otherwise. In the CL-SciSumm Shared Task 2016 [3], Task 1A requires participants to develop systems to identify the spans of text in the cited paper that most accurately reflect the citation. We directly use their public dataset. Since the dataset only contains positive (+*Prov*) instances, we must source for our own $-$*Prov* instances. As the majority of sentences do not provide provenance for a citation ($-$*Prov*), we must sample such negative sentences as we aim at keeping class balance in training a model to distinguish informative fragments from trivial ones. We randomly sample three negative instances from each cited paper. The final dataset has 608 +*Prov* and 885 $-$*Prov* instances (1,493 in total).

3 MODELS

For each individual task, we use a simple convolutional neural network (CNN) layer as our learning model (*cf.* individual halves in Figure 1). Our CNN architecture is a standard layered pipeline: word

[1] For a full version of this work, refer to https://arxiv.org/abs/1811.07351. It contains details on dataset collection and experiments.

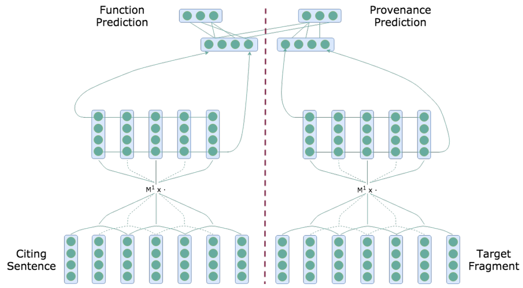

Figure 1: Our proposed MTL architecture.

Table 1: Performance of citation function and provenance models. "*" and "" indicate significant improvements at the $p < 0.01$ and $p < 0.001$ on paired significance t-test compared to the baseline ("BL"), respectively.**

(a) Citation function models

Model	BL	CNN	MTL
Precision	68.28%	68.78% ± 0.51%	69.55% ± 0.61%
Recall	69.40%	68.65% ± 0.68%	72.33% ± 0.36%
F_1	68.70%	68.31% ± 0.52%	69.63% ± 0.47%**

(b) Citation provenance models

Model	BL	dCNN	MTL
Precision	71.82%	79.36% ± 1.71%	79.47% ± 1.37%
Recall	72.13%	79.07% ± 1.84%	79.53% ± 1.36%
F_1	71.68%	78.55% ± 1.67%	79.38% ± 1.36%*

embedding, convolution, max pooling, ending with a fully connected layer. This CNN structure performs classification on a single text sequence. For citation provenance, both the citation context and the cited paper fragment need to be considered to make a classification decision. To this end, we use a double CNN (hereafter, "dCNN") architecture (similar to [5], *cf.* Figure 1), where two CNNs accept and process the two inputs separately, but combine at the fully-connected layer to generate class predictions.

A central claim of our work is that citation function and provenance are related; that there is a relationship between the function and provenance labels for a given citation. For example, it is unlikely that a *Pos* citation refers to texts in the target paper that in fact reveal a drawback. This motivates us to apply multi-task learning (MTL) which improves learning efficiency and prediction accuracy [2]. Our work requires the citing sentence as input to both tasks. Thus, we employ the dCNN model that addresses both citation function data and the citing sentences from provenance data (*cf.* Figure 1).

4 EXPERIMENTS

We implement the neural models in Keras. GloVe word embedding are taken as feature representations of the tokens. We use typical CNN settings: a window size of 5, a filter size of 256, a training minibatch size of 256, and a categorical cross entropy loss. For all neural models, we use RMSProp as the optimizer for 30 epochs.

We perform five-fold cross validation to evaluate all models[2]. For both tasks, we report all models for precision, recall, and F_1 scores weighted over all classes in Table 1. Supervised models in [4] and [6] are also included as baselines ("BL") for comparative evaluation. For citation function, the simple CNN model achieves similar performance as the supervised baseline, while MTL improves the supervised baseline by about 1%, reaching around 70%. For provenance, dCNN yields better results than the supervised baseline in all three evaluation measures. The MTL-learned dCNN model further improves over the plain CNN model by 1%, reaching scores of about 79.4% in all evaluation measures.

[2]Code and data available at https://github.com/WING-NUS/citation_func_n_prov

Table 2: Examples of (a) citation function and (b) citation provenance classifications.

(a) Citation function

Citing Sentence	**Actual**	**BL**	**MTL**
(a1) We show that the performance of our approach (using simple lexical features) is comparable to that of the state-of-art statistical MT system (Koehn et al., 2007).	*CoCo*	*Pos*	*CoCo*
(a2) Errors have been shown to have a significant impact on predicting learner level (Yannakoudakis et al., 2011).	*Neut*	*Weak*	*Neut*

(b) Citation provenance

Citing Sentence	**Target Fragment**	**Actual**	**BL**	**dCNN**
(b1) Bigrams have recently been shown to be very successful features in supervised word sense disambiguation (Pedersen, 2001).	This paper shows that the combination of a simple bigrams and a standard decision tree learning algorithm results in accurate word sense disambiguation.	*+Prov*	*+Prov*	*+Prov*
(b2) A number of automatically acquired inference/rule paraphrase collections are available, such as (Szpektor et al., 2004).	In this paper, we will propose an unsupervised method to discover paraphrases from a large untagged corpus.	*+Prov*	*−Prov*	*+Prov*

Table 2 shows some examples of citation function and provenance classifications. Among them, (b2) significantly demonstrates the advantages of a deep learning model. We observe that there are almost no word overlap between the two text inputs; this is likely why the baseline which relies on word overlap fails to classify it correctly. However, phrases such as "automatically acquired", "inference" share a common meaning as "unsupervised". Our deep learning model classifies the instance correctly, as GloVe embeddings can successfully capture semantic relationships.

5 CONCLUSION

We have examined two related tasks in citation analysis: citation function and provenance. We leverage our key insight of the relationship between the tasks and employ multi-task learning on top of neural models, resulting in further performance improvement.

REFERENCES

[1] Steven Bird, Robert Dale, Bonnie J Dorr, Bryan R Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proc. of LREC*. 1755–1759.

[2] Rich Caruana. 1998. Multitask Learning. *Learning to Learn* (1998), 95–133.

[3] Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi, and Min-Yen Kan. 2018. Insights from CL-SciSumm 2016: the Faceted Scientific Document Summarization Shared Task. *International Journal on Digital Libraries* 19, 2-3 (2018), 163–171.

[4] Heng Wee Low. 2011. Citation Provenance. Bachelor's thesis. School of Computing, National University of Singapore.

[5] Animesh Prasad. 2017. WING-NUS at CL-SciSumm 2017: Learning from Syntactic and Semantic Similarity for Citation Contextualization. In *Proc. of BIRNDL*. 26–32.

[6] Eric Yulianto. 2012. Citation Typing. Bachelor's thesis. School of Computing, National University of Singapore.