Neural-DINF: A Neural Network based Framework for Measuring Document Influence

Jie Tan, Changlin Yang, Ying Li, Siliang Tang*, Chen Huang, Yueting Zhuang
Zhejiang University
{tanjie95, yclzju, ying.li}@zju.edu.cn,
{siliang, beside.huang, yzhuang}@zju.edu.cn

Abstract

Measuring the scholarly impact of a document without citations is an important and challenging problem. Existing approaches such as Document Influence Model (DIM) are based on dynamic topic models, which only consider the word frequency change. In this paper, we use both frequency changes and word semantic shifts to measure document influence by developing a neural network based framework. Our model has three steps. Firstly, we train word embeddings for different time periods. Subsequently, we propose an unsupervised method to align vectors for different time periods. Finally, we compute the influence value of documents. Our experimental results show that our model outperforms DIM.

1 Introduction

Identifying the most influential articles is of great importance in many areas of research. It is often the case that we are increasingly exposed to numerous papers published every day. Research on influence evaluation can be applied to measure the scholarly impact of universities and research facilities. Besides, it helps researchers to distinguish valuable research work from a large number of scientific papers. The common approach of assessing an article's research impact is to count the number of explicit references to it. However, citations are often not available. For example, collections including blog posts and government documents adopt ideas proposed in the documents without explicit references (Stringer et al., 2008; Macroberts and Macroberts, 2010).

To identify influential articles without citations, Gerrish and Blei (2010) and Gerow et al. (2018) proposed probabilistic methods, which are based on dynamic topic models (Blei and Lafferty, 2006). They aimed to identify influential articles by examining the word frequency change over time. In this paper, we aim to use both word frequency changes and word semantic shifts on measuring document influence without citations. For our purpose, we propose a neural network based method called Neural-DINF, which stands for a Neural Network based Framework for measuring Document Influence. Our idea is that words that have semantic shifts across time contribute significantly to the influence of a document. Recent studies show that words whose word embeddings across different time periods diverge significantly are suspected to have semantic shifts (Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016).

Neural-DINF first generates static word embeddings in each time period by using Word2Vec (Mikolov et al., 2013b,a) independently, then aligns embeddings to the same vector space with an unsupervised method, subsequently calculates differences of the embeddings of many words across time to identify words that experience semantic shifts, finally measures the influence of a document by counting these crucial words. In summary, this paper makes the following main contributions:

- We consider both word frequency changes and word semantic shifts on measuring document influence without citations by developing a novel neural network framework.
- In the semantic change detection step, we propose an unsupervised method to align word embeddings across time.
- Neural-DINF outperforms dynamic topic based models such as DIM, which only considers the word frequency change.

This paper is organized as follows: Section 2 states related work; Section 3 formulates our approach;

*Corresponding Author.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6004–6009

Section 4 presents our experiments; Section 5 concludes our work.

2 Related work

There are two lines of literature that are closely related to our work: document influence evaluation and semantic shift detection.

2.1 Document Influence Evaluation

Assessing document influence only based on texts is a challenging task. Garfield et al. (2002) considered that the impact of a journal is based on aggregate citation counts. To identify influential articles without citations, Gerrish and Blei (2010) proposed the document influence model (DIM), which is a probabilistic model based on the dynamic topic model (Blei and Lafferty, 2006). In DIM, they considered the word frequency change and a document whose words can help the way the word frequencies change will have a high influence score. Gerow et al. (2018) improved DIM by incorporating features, such as authorship, affiliation, and publication venue and they aimed to explain how influence arises. In practice, this additional information is not often available.

In this paper, we measure document influence from a more fine-grained level by considering word semantic shifts. Our work differs from the above studies by considering both word frequency changes and word semantic shifts. Specially, we aim to find words that present significant changes in their meanings and we think these words contribute significantly to document influence. Neural-DINF assigns influence scores to documents based on how many of these important words are included in these documents.

2.2 Semantic Shift Detection

There has been a lot of research on detecting semantic changes across time (Kay, 1979; Traugott, 1989; Blank, 1999; Zhang et al., 2016; Liao and Cheng, 2016; Bamler and Mandt, 2017). In general, most approaches learn individual embeddings for different time slices and recognize the changes by comparing these embeddings. These vectors have to be aligned into the same vector space for comparison. To achieve alignment, Kim et al. (2014) trained word vectors for different years and then initialized the word vectors in subsequent years with the word vectors obtained from the previous years. Kulkarni et al. (2015) and Hamilton et al. (2016) addressed the embedding alignment problem by learning a linear transformation of words between any two time periods. Most of the alignment methods require anchor words whose meaning does not change between the two time slices. However, it is difficult for us to acquire this kind of prior knowledge, which involves additional expert supervision.

In this paper, inspired by Conneau et al. (2017), we propose an adversarial network for unsupervised cross-time alignment. Different from existing approaches, our method is unsupervised and does not require expert information.

3 Method

Our Neural-DINF contains the following three steps. First, we generate static word embeddings in each time slice separately. Then, we implement an unsupervised approach with adversarial training and a refinement procedure to align these embeddings to the same vector space. Finally, we present a new metric to evaluate the influence of a document without citations.

3.1 Word Embedding Generation

Our method first learns individual word embeddings for different time periods and any reasonable word embedding generation approach can be used for this purpose.

We consider a text corpus collected across time and use the texts of the documents to train word embeddings. We define our text corpus as $\mathcal{D}$= $(\mathcal{D}_1, \ldots, \mathcal{D}_T)$, where each $\mathcal{D}_t (t = 1, \ldots, T)$ is the texts of all documents in the t-th time slice. The length of these time slices is years in our model. Given any time slice of the texts, our goal is to learn word embeddings through Word2Vec (Mikolov et al., 2013b,a).

3.2 Unsupervised Cross-time Alignment

As our word embeddings for different time periods are trained in different vector spaces, we need to align them to the unified vector space for comparison. We aim at learning a mapping between word vectors for two different time periods. Let $\mathcal{S}' = \{s'_1, s'_2, \ldots, s'_m\} \subseteq \mathbb{R}^d$ and $\mathcal{S} = \{s_1, s_2, \ldots, s_n\} \subseteq \mathbb{R}^d$ be two sets of m and n word embeddings from time slices t' and t respectively where $t' \in \{t+1, \ldots, T\}$. Ideally, we can use a known dictionary including words that do not experience semantic shifts. Then we can learn a linear mapping W between the two embedding

spaces such that:

$$W^* = \arg \min_{W \in \mathbb{R}^{d \times d}} \|WX - Y\|^2, \quad (1)$$

where d is the dimension of the embeddings, and X and Y are two aligned matrices of size $d \times k$ formed by k word embeddings selected from S' and S, respectively. During the inference time, the aligned embedding of any word w at time slices t' is defined as $\arg\max_{s_j \in \mathcal{T}} \cos(Ws'_w, s_j)$. In this paper, we aim to learn this mapping W without using anchor words, which does not change meaning between the two time slices. We first apply an adversarial network to learn an initial proxy of W, then refine the model by using a synthetic parallel dictionary.

Domain-Adversarial Training. We define a discriminator which aims at discriminating between elements randomly samples from $W\mathcal{S}' = Ws'_1, Ws'_2, \ldots, Ws'_m$ and $\mathcal{S}$. The mapping W can be regarded as a generator, which aims at preventing the discriminator from making accurate predictions. The discriminator is designed to maximize its ability to identify the origin of an embedding, and the generator makes $W\mathcal{S}'$ and $\mathcal{S}$ as similar as possible to prevent the discriminator from accurately predicting the embedding origins.

We denote the discriminator parameters as θ_D. Given the mapping W, the optimization objective of the discriminator can be defined as:

$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(\text{origin} = 1|Ws'_i) - \frac{1}{n}\sum_{j=1}^{n} \log P_{\theta_D}(\text{origin} = 0|s_j), \quad (2)$$

where $P_{\theta_D}(\text{origin} = 1|z)$ is the probability that z originates from the embedding space at time slice t' (as opposed to an embedding from the embedding space at time slice t).

The mapping W is trained to prevent the discriminator from accurately predicting embedding origins and the optimization objective can be defined as:

$$\mathcal{L}_W(W|\theta_D) = -\frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(\text{origin} = 0|Ws'_i) - \frac{1}{n}\sum_{j=1}^{n} \log P_{\theta_D}(\text{origin} = 1|s_j). \quad (3)$$

According to the standard training process of adversarial networks (Goodfellow et al., 2014), the discriminator θ_D and the mapping W are consecutively trained to respectively minimize $\mathcal{L}_D$ and $\mathcal{L}_W$.

Refinement Procedure. The refinement procedure is designed to improve the performance of alignment after the domain-adversarial training step. We obtain a linear transformation W that maps a word from time slices t' to t in the last step.

To refine our mapping W, we utilize the learned W to build a syntactic parallel dictionary that specifies which $s'_i \in \mathcal{S}'$ refer to which $s_j \in \mathcal{S}$. Since the most frequent words are suspected to have better embeddings, we consider the most frequent words and keep only their mutual nearest neighbors. In the process of deciding mutual nearest neighbors, we use the Cross-Domain Similarity Local Scaling proposed in (Conneau et al., 2017) to alleviate the hubness problem (Dinu et al., 2014). Consequently, we use Eq. (1) on this obtained dictionary to refine W.

To compare vectors from different time periods, we propose an unsupervised approach. An adversarial network is first used to learn an initial proxy of W. To optimize the mapping W, we use a synthetic parallel dictionary in which words' semantics match the best.

3.3 Influence Evaluation

In this section, Neural-DINF evaluates document influence without citations. Our model makes use of both word frequency changes and word semantic shifts to compute an influence score for each document. We quantify the semantic change of the words by calculating the cosine similarity of the embedding vectors for the same words in different years. We represent aligned vectors of the word w in t and t' as $\mathbf{w}$ and $\mathbf{w}'$ respectively. We compute the word meaning shift of w as follows:

$$V_w = 1 - \cos\langle \mathbf{w}, \mathbf{w}' \rangle. \quad (4)$$

Given a document d of time slice t, the influence score of this document on the corpus $\mathcal{D}_{t'}$ can be defined as:

$$I_d^{t'} = \sum_{w \in D_{t,t'} \cap \mathbf{D}} V_w \cdot \frac{C_{d,w}^t}{C_w^t}, \quad (5)$$

where $D_{t,t'}$ is the vocabulary consisting of co-occurence words of corpus $\mathcal{D}_t$ and $\mathcal{D}_{t'}$, $\mathbf{D}$ is the

vocabulary of document d, $C^t_{d,w}$ represents the frequency of word w in the document d, C^t_w represents the frequency of word w in the corpus $\mathcal{D}_t$. The document published at time slice t can only affect documents published after that time slice, so the influence score of document d on the corpus $\mathcal{D}$ can be defined as:

$$\mathcal{I}_d = \sum_{t'=t+1}^{t'=T} I_d^{t'}. \quad (6)$$

4 Experiments

Similar to previous studies (Gerrish and Blei, 2010; Gerow et al., 2018) on measuring documents' scholarly impact, we evaluate the performance of Neural-DINF by Pearson correlation and Spearman rank correlation of influence scores and citation counts. We reproduce the DIM (Gerrish and Blei, 2010) as our baseline and its experimental setup is as follows: topics' Markov chain variance $\sigma^2 = 0.005$, topic number $K = 5$, LDA (Blei et al., 2003) hyperparameter $\alpha = 0.001$.

In Neural-DINF, word embeddings are generated by training on the corpus of each year and word embedding size is 300. We only select the first 10k most frequent words in each year in our experiments. This threshold is determined by the size of the smallest vocabulary in the years (2002-2013). In the unsupervised alignment, we use the default setting specified in (Conneau et al., 2017) to build a discriminator and the dimension of W is 300×300. Stochastic gradient descent(SGD) is used to train the discriminator and W with the learning rate of 0.1. We only feed the discriminator with 3000 most frequent words. This is because the embeddings of rare words are of low quality (Luong et al., 2013), which makes them harder to align. It is observed that feeding the discriminator with rare words had a small negative impact which cannot be ignored. In the refinement procedure, we retain the same setting presented in (Conneau et al., 2017).

4.1 Data

For evaluation, we analyze a sequential corpus *The Association for Computational Linguistics Anthology* (ACL Anthology), which is a collection of documents on the study of computational linguistics and natural language processing (Bird et al., 2008). Following the experimental setup in DIM, we only use the texts and dates of this corpus. We analyze a subsample from ACL Anthology, spanning from 2002 to 2013, which contains 11106 articles and 18960 unique tokens after preprocessing. We remove short documents and words that have low frequency and low TF-IDF value. Citation counts of articles are obtained from *ACL Anthology Network* (Joseph and Radev, 2007; Leskovec et al., 2009; Radev et al., 2013).

4.2 Result

We compare the correlation coefficient scores on DIM and Neural-DINF in Table 1. The Pearson correlation computed by Neural-DINF and DIM is 0.186 and 0.118 respectively. The Spearman rank correlation computed by Neural-DINF and DIM is 0.249 and 0.102 respectively. The results show that our model outperforms the DIM.

Method	Pearson correlation	Spearman rank correlation
DIM	0.118	0.102
Neural-DINF	**0.186**	**0.249**

Table 1: Pearson correlation and Spearman rank correlation between citation counts and the influence score.

We also visualize the performances of DIM and our Neural-DINF to validate the effectiveness of our proposed model. As shown in Figure 1, for ACL documents with the highest 60% of influence scores. Neural-DINF covers 83% of citations, which outperforms DIM (68%) by a large marge.

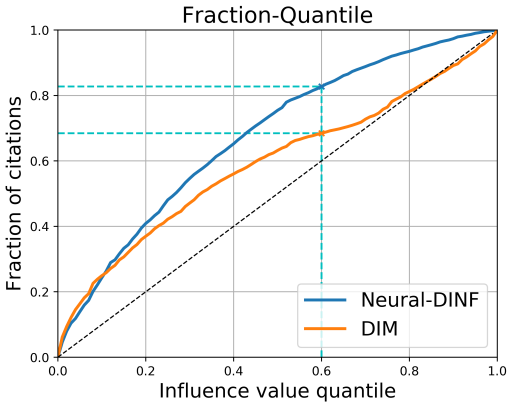

Figure 1: Fraction of citations explained by influence scores.

In fact, the qualitative analysis does present some evidence that in many cases the Neural-DINF is a better model to produce reasonable scores for the most-cited papers in the used datasets. For

example, *A Systematic Comparison of Various Statistical Alignment Models* (Och and Ney, 2003) is a top-cited article (citation ranking 3) in the dataset. This article receives a very high score both on the DIM and the Neural-DINF. However, the result of Neural-DINF ranking (31) is more close to its citation ranking than the DIM (236). Moreover, in some cases, only Neural-DINF can produce the correct score. For example, DIM assigns a relatively low influence score to (Collins, 2002) (citation ranking 9) in our dataset and ranks this article 11,106 out of 11,106 articles, while the Neural-DINF gives a relatively reasonable score to this article, ranking it 1,199 out of 11,106 articles.

5 Conclusion

In this paper, we aim to evaluate document influence from a fine-grained level by additionally considering word semantic shifts. For our purpose, we develop Neural-DINF which measures document influence from the texts of documents. Besides, we propose an unsupervised method to address the alignment problem. The document receives an influence score based on how it explains the word frequency change and the word semantic shift. Our experimental results show that our model performs better than the DIM on ACL Anthology.

Acknowledgments

This work has been supported in part by National Key Research and Development Program of China (2018AAA010010), NSFC (No.61751209, U1611461), University-Tongdun Technology Joint Laboratory of Artificial Intelligence, Zhejiang University iFLYTEK Joint Research Center, Chinese Knowledge Center of Engineering Science and Technology (CKCEST), China Engineering Expert Tank, Engineering Research Center of Digital Library, Ministry of Education, the Fundamental Research Funds for the Central Universities.

References

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 380–389. JMLR. org.

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan 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.

Andreas Blank. 1999. Why do new meanings occur? a cognitive typology of the motivations for lexical semantic change. *Historical semantics and cognition*, 13:6.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

Eugene Garfield, Alexander I Pudovkin, and VS Istomin. 2002. Algorithmic citation-linked historiography—mapping the literature of science. *Proceedings of the American Society for Information Science and Technology*, 39(1):14–24.

Aaron Gerow, Yuening Hu, Jordan Boyd-Graber, David M Blei, and James A Evans. 2018. Measuring discursive influence across scholarship. *Proceedings of the national academy of sciences*, 115(13):3308–3313.

Sean Gerrish and David M Blei. 2010. A language-based approach to measuring scholarly impact. In *ICML*, volume 10, pages 375–382. Citeseer.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Mark T Joseph and Dragomir R Radev. 2007. Citation analysis, centrality, and the acl anthology. Technical report, Citeseer.

Margarita Kay. 1979. Lexemic change and semantic shift in disease names. *Culture, medicine and psychiatry*, 3(1):73–94.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM.

Xuanyi Liao and Guang Cheng. 2016. Analysing the semantic change based on word embedding. In *Natural language understanding and intelligent applications*, pages 213–223. Springer.

Minh-Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

M. H. Macroberts and B. R. Macroberts. 2010. Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the Association for Information Science & Technology*, 61(1):1–12.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.

Michael J. Stringer, Sales Pardo Marta, Amaral Luís A. Nunes, and Scalas Enrico. 2008. Effectiveness of journal ranking schemes as a tool for locating information. *Plos One*, 3(2):e1683–.

Elizabeth Closs Traugott. 1989. On the rise of epistemic meanings in english: An example of subjectification in semantic change. *Language*, pages 31–55.

Yating Zhang, Adam Jatowt, Sourav S Bhowmick, and Katsumi Tanaka. 2016. The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807.