Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

What shall we read: the article or the citations? - A case study on scientific language understanding

Aman Sinha[1, 2] Sam Bigeard[1] Marianne Clausel[1] Mathieu Constant[2]
(1) IECL, Université de Lorraine, Nancy, France
(2) ATILF, Université de Lorraine, Nancy, France
{firstname.lastname}@univ-lorraine.fr

RÉSUMÉ

The number of scientific articles is increasing tremendously across all domains to such an extent that it has become hard for researchers to remain up-to-date. Evidently, scientific language understanding systems and Information Extraction (IE) systems, with the advancement of Natural Language Processing (NLP) techniques, are benefiting the needs of users. Although the majority of the practices for building such systems are data-driven, advocating the idea of "*The more, the better*". In this work, we revisit the paradigm - questioning what type of data : text (title, abstract) or citations, can have more impact on the performance of scientific language understanding systems.

ABSTRACT

Lire l'article ou les citations ? - Une étude de cas sur la compréhension du langage scientifique

Le nombre d'articles scientifiques explose dans tous les domaines à tel point qu'il est devenu difficile pour les chercheurs de suivre l'évolution de la littérature scientifique. De toute évidence, intégrer les avancées dans le domaine du traitement du langage naturel (TAL), ne peut être que bénéfique aux utilisateurs cherchant à faire de la veille scientifique. La majorité des systèmes actuels est basée sur l'exploitation du caractère massif des données, partant de l'idée que "*Plus de données implique de meilleures performances*". Dans ce travail de recherche, nous revisitons le paradigme : se demander quel type de données : texte (titre, résumé) ou de citations, peut avoir plus d'impact sur la performance des systèmes scientifiques de compréhension du langage.

MOTS-CLÉS : Littérature scientifique, Recherche d'information, Compréhension du langage scientifique.

KEYWORDS: Scientific text, Information Extraction, Scientific Language Understanding.

1 Introduction

Scientific databases [1] are accumulating a large amount of literature to such an extent that it is getting overwhelming (Johnson *et al.*, 2018) and practically impossible to be up-to-date for researchers. Several recent works have looked into building intelligent systems for tasks such as ad-hoc based retrieval, conversational-agents, recommendation, summarization, document search, and re-ranking, as shown by the advances in the area of Neural Information Retrieval(IR) and Biomedical text mining (Zhang *et al.*, 2016; Gu *et al.*, 2020; Thakur *et al.*, 2021). Given the unstructured nature, the lengthiness and other metadata (such as citations information), in the scientific documents, the key question that arises when dealing with scientific information extraction is what type of data can be helpful to build high-performance models : text and/or citation information ?

1. Examples : Pubmed (https ://pubmed.ncbi.nlm.nih.gov/), ASCO(https ://www.asco.org/) ; ArXiv (https ://arxiv.org/), etc.

Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 80–85.

We investigate the informativeness of different type of features : *text-only* features, *graph-only* features and *text-graph* features via the document classification task. Additionally, we further explore impact of increasing the amount of text features. The main contribution of our work is to benchmark the effectiveness of the three type of features via various machine learning and deep learning based models for document classification task.

2 Related Work

This work investigates the impact of different text and graph-based features for document classification task. We briefly discuss the works which are related to our study by grouping them into the following categories :

Neural IE & Transfer Learning. The effectiveness of more text for text-ranking and text-retrieval tasks has been studied previously(Lin, 2009). Several works (Guo *et al.*, 2011; Ermakova *et al.*, 2018; Yeganova *et al.*, 2021) have also investigated at the importance of features from different sections in scientific articles using methods such as traditional BM25 scoring, deep-NN (Huang *et al.*, 2013) and weighted word-count (Yu *et al.*, 2014). In parallel, transfer learning (Peng *et al.*, 2019; Gu *et al.*, 2020; Kanakarajan *et al.*, 2021) has received a lot of attention with domain specific pre-trained language models (PLMs) for various downstream tasks such NER, relation extraction, question answering, document classification for scientific and biomedical text mining.

Citation graphs. Viewing data as a *graph* has been positively explored for various NLP tasks (Wu *et al.*, 2023) involving knowledge graphs (Mondal *et al.*, 2021; Jin *et al.*, 2019) and ontology-integrated models (Sinha *et al.*, 2022), showcasing the existing informativeness in graphs. We explore graph representation learning (Perozzi *et al.*, 2014; Hamilton *et al.*, 2017; Kipf & Welling, 2017) combining the structural information of graph with text-features in order to enhance the learning of information extraction models.

3 Experiments

Dataset We use the *PubMed-Diabetes* (Galileo Mark Namata & Huang, 2012) for our experiments. The dataset contains 19717 articles belonging to 3 classes of diabetes-mellitus, *Experimental*, *Type-1*, and *Type-2*. The dataset also provides citation information (eg. paper A $\xrightarrow{\text{cites}}$ paper B) and original PubMed-IDs. Using the graph terminology, PubMed-ID(s) denotes node(s) and the citation relation denotes edge(s) and therefore, the dataset can be also viewed as a citation graph with 19716 nodes [2] and 44338 edges.

Models We briefly describe the different models we use to perform our learning task. These models exploit standalone text features or graph features or combination of both the features.

— **Text-CNN** (Zhang & Wallace, 2015) : Besides the BERT (Devlin *et al.*, 2018) [CLS]-baseline, we used BERT-CNN architecture, which comprised of five 2d-CNN blocks applied on the token representations followed by max-pooling layer, to capture better quality text features.

— **DeepWALK** (Perozzi *et al.*, 2014) : Similar to Word2Vec algorithm (Mikolov *et al.*, 2013), this algorithm takes as input, a sequence of connected nodes and provides static embeddings for each node in the citation graph.

— **GraphSAGE** (GS) (Hamilton *et al.*, 2017) : This algorithm takes as input the citation graph and node-features (text features) to generate node embeddings, which contains both combina-

2. Note : One of the article-id ($pid.17874530$) was not anymore valid on PubMed, so we removed it from the nodes set ; no article was connected with the removed article-id and so the edge list was unaffected.

tion of graph-based and text-based information. The algorithm works under the concept of iterative message aggregation for every node from the neighborhood.

Features We briefly describe various features and PLMs that we used in our experiments.

— **random** : We use random noise feature in order to show the impact of addition of text/graph features for any model.

— **tfidf** : The dataset was provided with 500-length tf-idf features based on a curated list of keywords (Galileo Mark Namata & Huang, 2012) (tfidf-c). For our study on different text portions (title-only or abstract-only), we generated manual tf-idf features (tfidf-m) separately with each corpora.

— **BERT** (Devlin *et al.*, 2018) : We use the contextual embeddings ([CLS]-layer) from BERT model which is pretrained with English Wikipedia and BookCorpus. We use it as our baseline for domain-specific PLMs.

— **BioBERT** (Lee *et al.*, 2020) : This is a domain-specific language model obtained by training BERT on biomedical corpora including PubMed and PubMed Central (PMC).

— **PubMedBERT** (Gu *et al.*, 2020) : Contrary to continual pretraining as BioBERT, this model is pretrained from scratch using abstracts from PubMed. Its main advantage over BERT and BioBERT is its in-domain vocabulary, which helps it to identify medical terms (eg. *naloxone*, *acetyltransferase*) avoiding subword fragmentation [3].

— **BioELECTRA** (Kanakarajan *et al.*, 2021) : It is a biomedical domain-specific language encoder model that adapts ELECTRA (Clark *et al.*, 2020) by pretraining it from scratch with biomedical domain text from PubMed+PMC and uses domain specific vocabulary from PubMedBERT. It outperforms the previous models and achieves state-of-the-art (SOTA) in BLURB benchmark (Gu *et al.*, 2020).

Implementation Details In our experiments, we follow the random data splitting from the (Yang *et al.*, 2016) work by which we keep 20 random instances of each class for training, 500 random instances for development set, and 1000 constant instances for test. We report the average results for 5 seeded model runs. We keep the hyperparameter setting common across all the models. We use learning-rate (1e-06), epochs(500), maxlen (title=64, abst=512), and batch-size(title=64, abst=8).

4 Results & Discussion

Table 1 shows the results divided into 4 subsections. The first subsection shows *random* features to compare and show the impact of feature-based models. In the next subsection, we have *text-only* based features including *tfidf* and PLM-based models with two variants : [CLS]-finetuning and Text-CNN model. In the last two subsections, we report *graph-only* features and *text+graph*, which is a combination of text and graph features.

Random VS Features Comparing with *random* features, we notice the impact of text-only features (*tfidf-c/m*, PLM), graph-only features (*DW*) and combination of text+graph features (*tfidf+DW*, *GS*).

Text features We notice that *tfidf-c* performs better than *tfidf-m*, which can be attributed to the curated keyword list compared as automatic generated keyword list. We continue with *tfidf-m* to study **Title**- and **Abst**- text separately. We observed that *PLM*-based features perform predominately better than *tfidf* features with the exception of `BERT-base-case` (because of its general vocabulary) and further, all the **Abst**- models outperform the **Title**- based models showing the relevance of the Abstract section. We notice that surprisingly BioElectra performs lower than PubMedBERT for all cases (except for [CLS] baseline with title corpora) and its performance degrades with Text-CNN model.

3. Word shattering for out-of-vocab words eg. *acetyltransferase* → [ace, ##ty, ##lt, ##ran, ##sf, ##eras, ##e]

	Method	Title				Abstract			
		P	R	F1	Acc	P	R	F1	Acc
R	random	$0.36_{0.03}$	$0.33_{0.03}$	$0.33_{0.03}$	$0.32_{0.03}$	-	-	-	-
T	tfidf-c	$0.70_{0.02}$	$0.68_{0.03}$	$0.67_{0.03}$	$0.68_{0.03}$	-	-	-	-
	tfidf-m	$0.61_{0.03}$	$0.60_{0.02}$	$0.59_{0.03}$	$0.60_{0.02}$	$0.71_{0.02}$	$0.70_{0.03}$	$0.69_{0.03}$	$0.70_{0.03}$
	BERT	$0.41_{0.09}$	$0.43_{0.02}$	$0.36_{0.04}$	$0.43_{0.02}$	$0.45_{0.16}$	$0.50_{0.07}$	$0.43_{0.11}$	$0.50_{0.07}$
	BioBERT	$0.67_{0.04}$	$0.64_{0.06}$	$0.62_{0.09}$	$0.64_{0.06}$	$0.72_{0.08}$	$0.67_{0.14}$	$0.66_{0.17}$	$0.67_{0.14}$
	PubmedBERT	$0.71_{0.06}$	$0.68_{0.07}$	$0.67_{0.09}$	$0.68_{0.07}$	$0.86_{0.02}$	$0.86_{0.02}$	$0.86_{0.02}$	$0.85_{0.03}$
	BioElectra	$0.76_{0.03}$	$0.71_{0.04}$	$0.70_{0.05}$	$0.71_{0.04}$	$0.86_{0.02}$	$0.85_{0.02}$	$0.85_{0.02}$	$0.84_{0.03}$
	BERTCNN	$0.60_{0.07}$	$0.59_{0.08}$	$0.59_{0.08}$	$0.59_{0.08}$	$0.66_{0.05}$	$0.65_{0.05}$	$0.65_{0.05}$	$0.65_{0.05}$
	BioBERTCNN	$0.70_{0.06}$	$0.69_{0.06}$	$0.69_{0.06}$	$0.69_{0.06}$	$0.72_{0.03}$	$0.70_{0.03}$	$0.70_{0.04}$	$0.69_{0.05}$
	PubmedCNN	$0.83_{0.03}$	$0.82_{0.04}$	$0.82_{0.04}$	$0.82_{0.04}$	$0.87_{0.03}$	$0.87_{0.03}$	$0.87_{0.04}$	$0.87_{0.04}$
	BioElectraCNN	$0.65_{0.10}$	$0.62_{0.06}$	$0.58_{0.09}$	$0.62_{0.06}$	$0.85_{0.02}$	$0.84_{0.02}$	$0.84_{0.02}$	$0.84_{0.03}$
G	DeepWALK (DW)	$0.67_{0.03}$	$0.64_{0.03}$	$0.65_{0.03}$	$0.64_{0.03}$	-	-	-	-
	random+GraphSAGE(GS)	$0.17_{0.10}$	$0.36_{0.08}$	$0.21_{0.08}$	$0.36_{0.08}$	-	-	-	-
T+G	tfidf-c+DW	$0.68_{0.02}$	$0.65_{0.02}$	$0.65_{0.02}$	$0.65_{0.02}$	-	-	-	-
	tfidf-m+DW	$0.68_{0.02}$	$0.65_{0.02}$	$0.66_{0.02}$	$0.65_{0.02}$	$0.69_{0.02}$	$0.65_{0.02}$	$0.66_{0.02}$	$0.65_{0.02}$
	tfidf-c+GS	$0.78_{0.03}$	$0.76_{0.02}$	$0.76_{0.02}$	$0.76_{0.02}$	-	-	-	-
	tfidf-m+GS	$0.42_{0.02}$	$0.40_{0.04}$	$0.40_{0.04}$	$0.40_{0.04}$	$0.41_{0.01}$	$0.39_{0.02}$	$0.38_{0.01}$	$0.39_{0.02}$

R : *no-feature* ; T : *text-only* ; G : *graph-only* ; T+G : *text+graph*

TABLE 1 – Experiment results. Subscript denotes standard deviation across multiple runs

Graph Features We report *DeepWALK (DW)* model performance as we notice its informativeness is comparable to *tfidf* features. We also experiment with noise features as node-feature input to GraphSAGE(GS) model, and we notice drastic drop in performance compared to static graph features(DW). This indicates the impact of initial node-features for GS learning algorithm.

Text+Graph features Initially, we examine the concatenation of *DW* and *tfidf* and notice that *tfidf-m*(**Abst**)+*DW* performs comparable to *tfidf-m*(**title**)+*DW* but both the models perform lower than *tfidf-m*(**Abst**). Further, it is interesting to see that *DW* features perform better than *tfidf-m* trained on **Title**-text (when compared to *tfidf-m* only), but not for **Abstract**-text. Next, we experiment with *GS* with two settings, first with *tfidf-m* features we notice the performance decreases compared to *text-only* setting. This can be attributed to the iterative message passing during the process of node embedding generation can damage the original information. Lastly, with *tfidf-c* we notice an increased performance compared to *tfidf-m* model, which can be attributed to the curated list of keywords and spareness of *tfidf-m* features.

5 Conclusion

In this work, we provide a benchmark to study the informativeness of different features : text-based and citation information-based for document classification. We show that the models perform better with the Abstract text and we show that the information contained in the citation graph can be useful in addition to the text based features. In future, we would like to extend our study to investigate the relevance of other sections in scientific articles by combining graph models with PLMs. Additionally, the dynamic nature of the citation graphs can be interesting to understand the evolution and change points phenomena to study novelty and discovery in scientific citation graphs.

6 Acknowledgement

This work has been partly funded by the Project OLKI (Lorraine Université d'Excellence 2018-2021).

Références

CLARK K., LUONG M.-T., LE Q. V. & MANNING C. D. (2020). Electra : Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv :2003.10555*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

ERMAKOVA L., BORDIGNON F., TURENNE N. & NOEL M. (2018). Is the abstract a mere teaser ? evaluating generosity of article abstracts in the environmental sciences. *Frontiers Res. Metrics Anal.*, **3**, 16.

GALILEO MARK NAMATA, BEN LONDON L. G. & HUANG B. (2012). Query-driven active surveying for collective classification. In *International Workshop on Mining and Learning with Graphs*, Edinburgh, Scotland.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2020). Domain-specific language model pretraining for biomedical natural language processing.

GUO Y., KORHONEN A., LIAKATA M., SILINS I., HÖGBERG J. & STENIUS U. (2011). A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics*, **12**, 69 – 69.

HAMILTON W., YING Z. & LESKOVEC J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, **30**.

HUANG P.-S., HE X., GAO J., DENG L., ACERO A. & HECK L. (2013). Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, p. 2333–2338.

JIN W., ZHANG C., SZEKELY P. A. & REN X. (2019). Recurrent event network for reasoning over temporal knowledge graphs. *ArXiv*, **abs/1904.05530**.

JOHNSON R., WATKINSON A. & MABE M. (2018). The stm report : An overview of scientific and scholarly publishing.

KANAKARAJAN K. R., KUNDUMANI B. & SANKARASUBBU M. (2021). BioELECTRA :pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, p. 143–154, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.bionlp-1.16.

KIPF T. & WELLING M. (2017). Semi-supervised classification with graph convolutional networks. *ArXiv*, **abs/1609.02907**.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2020). Biobert : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240.

LIN J. J. (2009). Is searching full text more effective than searching abstracts ? *BMC Bioinformatics*, **10**, 46 – 46.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, **26**.

MONDAL I., HOU Y. & JOCHIM C. (2021). End-to-end construction of nlp knowledge graph. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 1885–1895.

PENG Y., YAN S. & LU Z. (2019). Transfer learning in biomedical natural language processing : an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv :1906.05474*.

PEROZZI B., AL-RFOU R. & SKIENA S. (2014). Deepwalk : Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 701–710.

SINHA A., OLLINGER S. & CONSTANT M. (2022). Word sense disambiguation of french lexicographical examples using lexical networks. In *TextGraphs-16 : Graph-based Methods for Natural Language Processing*, p. 70–76.

THAKUR N., REIMERS N., RÜCKLÉ A., SRIVASTAVA A. & GUREVYCH I. (2021). Beir : A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv :2104.08663*.

WU L., CHEN Y., SHEN K., GUO X., GAO H., LI S., PEI J., LONG B. *et al.* (2023). Graph neural networks for natural language processing : A survey. *Foundations and Trends® in Machine Learning*, **16**(2), 119–328.

YANG Z., COHEN W. & SALAKHUDINOV R. (2016). Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, p. 40–48 : PMLR.

YEGANOVA L., KIM W. G., COMEAU D., WILBUR W. J. & LU Z. (2021). Measuring the relative importance of full text sections for information retrieval from scientific literature. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, p. 247–256, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.bionlp-1.27.

YU L., HERMANN K. M., BLUNSOM P. & PULMAN S. (2014). Deep learning for answer sentence selection. *arXiv preprint arXiv :1412.1632*.

ZHANG Y., RAHMAN M. M., BRAYLAN A., DANG B., CHANG H.-L., KIM H., MCNAMARA Q., ANGERT A., BANNER E., KHETAN V. *et al.* (2016). Neural information retrieval : A literature review. *arXiv preprint arXiv :1611.06792*.

ZHANG Y. & WALLACE B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv :1510.03820*.