**KGEA: A Knowledge Graph Enhanced Article Quality Identification Dataset**

Chunhui Ai¹, Derui Wang², Yang Xu², Wenrui Xie², Ziqiang Cao¹
¹ Institute of Artificial Intelligence, Soochow University, Suzhou, China
² Baidu Inc., Beijing, China

20215227120@stu.suda.edu.cn, {wangderui,xuyang24,xiewenrui01}@baidu.com, zqcao@suda.edu.cn

June 16, 2022

**ABSTRACT**

With so many articles of varying quality being produced at every moment, it is a very urgent task to screen this data for quality articles and commit them out to social media. It is worth noting that high quality articles have many characteristics, such as relevance, text quality, straightforward, multi-sided, background, novelty and sentiment. Thus, it would be inadequate to purely use the content of an article to identify its quality. Therefore, we plan to use the external knowledge interaction to refine the performance and propose a Knowledge Graph Enhanced Article quality identification dataset (KGEA) based on Baidu Encyclopedia. We quantified the articles through 7 dimensions and use co-occurrence of the entities between the articles and the Baidu encyclopedia to construct the knowledge graph for every article. We also compared some text classification baselines and found that external knowledge can guide the articles to a more competitive classification with the graph neural networks.

1 Introduction

In the information explosion, media articles are appearing all over the place, but the quality is uneven. Therefore, we want to filter media articles and screen them for excellent quality so that they can be read more meaningfully.

Excellent articles have many advantages. In addition to the quality of the article itself, an excellent article should also have depth of content, novelty, etc. In reading excellent articles, readers can understand things from the straightforward presentation of the author and can also deepen their reading impressions through the many aspects presented in the article. In addition, excellent articles are often topical and often highly current. In order to be able to quantify the excellence of the articles, we have identified 7 dimensions through our research. These include: relevance, text quality, straightforward, multi-sided, background, novelty and sentiment. By portraying the articles in various dimensions, we propose a Knowledge Graph Enhanced Article quality identification dataset (KGEA).

Traditional text classification models such as BERT [1], XLNet [2] and RoBERTa [3] focus on learning semantic information while ignoring the linking relationships between entities in the text. These models are pre-trained over open-domain corpora, but downstream tasks are fine-tuned in specific domains. Liu et al. [4] believes that while these models can work well on datasets such as GLEU [5], they often do not work well in domain-specific tasks due to lack of knowledge. One of the better ways of introducing knowledge at present is the Graph Neural Networks (GNN) [6]. Inspired by this, we want to adopt Graph Neural Networks as a tool for knowledge injection. The results show that graph neural networks are more effective than BERT, ERNIE, K-BERT.

In summary, the contributions of this work include:

- we propose an excellent article dataset KGEA.
- We compared a number of baselines to demonstrate the effectiveness of Graph Neural Networks.
2 Related Work

Text classification

Text classification is a very important sub-task of natural language understanding (NLU). The goal of text classification is to extract features from the original text data and predict the class of the text data based on these features. It includes sentiment analysis, news categorization, topic classification, question answering (QA), and natural language inference (NLI) [7]. With the continuous development of deep learning, the capabilities of text classification models and human evaluation are approaching. Currently, the main approaches to text classification are: bag of words [8][9], RNN [10][11], CNN [12][13], Graph Neural Networks [14], Attention [15] and Pre-train models such as BERT [1] and RoBERTa [3].

Graph Neural Networks

Graph Neural Networks (GNN) have shown great potential in tackling graph analytic problems, such as node classification [16][17], graph classification [18][19], link prediction [20][21], and recommendation [22][23]. GCN [20] is a further extension while researching on GNN. Recently, GCN was explored in some NLP tasks, such as Text Classification [24][25], Reading Comprehension [26][27], Machine Translation [28][29] and achieve state-of-the-art results in a number of datasets.

3 Dataset Construction

In this section we focus on describing the annotation criteria and construction of our dataset.

3.1 Data Collection

Topic based text classification has recently achieved tremendous success in advancing the performance. However, In-depth articles identification is a big challenge because of external knowledge interaction. We randomly extract 7835 Chinese articles as our training dataset, 1078 articles as dev dataset, and 1106 articles as test dataset. We then invited four data annotators to annotate the data we were given, and the data annotation criteria are described in section 3.2.

3.2 Data Annotation

In order to effectively distinguish in-depth articles, we scored them in 7 areas, in the range of \( \{0, 1, 2\} \). These can be summarized as follows:

- **Relevance** In-depth articles with a high degree of relevance between the content of the article and the article title. We ensure that in-depth articles have a high degree of consistency between content and title by measuring the relevance of the article content and title.

- **Text Quality** We propose this indicator to measure the standards of grammar and spelling of the sentences in the text. In-depth writing, sentences free of typos and grammatical errors, and overall fluidity and consistency.

- **Straightforward** Straightforward consists of two aspects:
Figure 2: Top is an example of Baidu encyclopedia’s Short Description and Infoboxes Structure. The diagram below is entity connection graph combination of articles and encyclopedia that we have constructed.

- The article uses a large number of accurate figures or data to portray the matter being written about. The article quantifies the whole picture of the matter through data, allowing the user to understand the matter more directly and the content of the article is more accurate.

- The article uses clear and concise language to summarise or define the characteristics of the matter being written about, so that the user receives the characteristics of the matter more directly and the ideas are expressed more concisely.

We propose this marking aspect in the hope that the resulting quality article will be universally readable and understandable.

**Multi-Sided**  This criterion is sufficient if the article meets any of the following characteristics:

- Cite sufficient external quotations (examples, evidence, arguments, etc.) to argue or analyse central ideas or personal opinions, to enhance the persuasiveness of the essay and make it stand up to scrutiny and convince readers.

- A rigorous logic of thought that uncovers the multifaceted nature of things (without resorting to external quotes or historical context) and provides a multidimensional presentation of objects and perspectives that enables users to understand the content from different dimensions, making the subject or point explicit and insightful.

However, if the essay is not in the style of an argumentative essay, we will award a score as 1.

**Background**  We want to have an in-depth article with a clear background to the subject matter, with points and lines of information, such as a timeline, a line of events, etc. We don’t think there is any depth to a straightforward narrative or a simple list of events.

**Novelty**  We define the novelty of an article as the following:
Figure 3: Data distribution for relevance, text quality, straightforward, multi-sided, background, novelty, sentiment and label.

| Model          | dev accuracy | dev precision | dev recall | dev f1     | test accuracy | test precision | test recall | test f1     |
|----------------|--------------|---------------|------------|------------|---------------|---------------|-------------|------------|
| BERT           | 0.7737       | 0.7087        | 0.7234     | 0.7169     | 0.7785        | 0.6743        | 0.6937      | 0.6839     |
| ERNIE 1.0      | 0.7514       | 0.7219        | 0.6033     | 0.6573     | 0.7649        | 0.6826        | 0.5969      | 0.6369     |
| K-BERT         | 0.7737       | 0.731         | 0.676      | 0.702      | 0.7821        | 0.696         | 0.654       | 0.675      |
| GCN+Random     | 0.7876       | 0.6966        | 0.8192     | 0.753      | 0.7694        | 0.6224        | 0.8445      | 0.717      |
| GCN+Word2Vec   | 0.7978       | 0.7055        | 0.838      | 0.7661     | 0.7929        | 0.6597        | 0.8272      | 0.734      |

Table 1: Comparison between BERT, graph-based methods, ERNIE and K-BERT.

- Content topics relate to popular events or extended developments of events and hottest topics of discussion (including but not limited to entertainment, film and television, etc.)
- Recent major events at home and abroad or by prominent figures, such as major domestic current affairs news, international conflicts, recent changes in cross-strait relations, etc.
- The author has an original and analytical perspective on common events, with novel and rare perspectives, and expresses content that is refreshing and expands the perception.

**Sentiment**  We believe that insightful articles spread enthusiasm and confidence. We define positive articles into the following three categories:

- The author’s perspective is inspiring and motivating.
- The atmosphere of the article is inviting, touching humanity, inspiring the people and arousing emotional resonance
- Use humorous language to promote the user’s sentiment, the content is expressed in a relaxing way to make the user feel comfortable.

The distribution of the data as a group is shown in the Fig 3. In the final data, we focus on text quality, multi-sided and multi-faceted, background information and novelty. We used the entropy method to determine the weighting of the indicators, and we ended up with 3,460 quality articles, representing a ratio of 37%.

### 3.3 Graph Construction

**Extracting Terms** As described above, terms can be extracted by NER tools. Among different NER tools, Baidu lexer can locate the characteristic or property of a particular word and find at most 38 types of name entities with 154 types of sub-categories. It helps tremendously with our feature extraction and provides external knowledge.

Regarding terms definition, A prior knowledge that can be gathered is that the better quality of terms being extracted, the more accurate content features are exposed to the model. These terms are all name entities instead of oral expression.
**Build Connection** We use the Baidu encyclopedia to build term connections. After extracting the terms, we searched them through Baidu Encyclopedia and constructed edges for co-occurring terms (both terms co-occur in the article and in Baidu Encyclopedia). Fig 2 is an illustration of our graph.

4 Experimental Evaluation

4.1 Baseline

We apply a number of baseline models to our experiments. They are listed below:

- **BERT**: This method fine-tune BERT with the training data and selects the character with the highest probability for classification.
- **GCN**: A graph based text classification model, which constructs graph on the vocabulary instead of the documents. NOTE: GCN+Random means that nodes set by random initialization. GCN+Word2Vec means graph node initialization with Tencent-Word2Vec.
- **ERNIE 1.0**: ERNIE 1.0 improves the content of BERT masks. BERT masks word pieces, but it is easy for the model to predict the content of a mask from information about the word alone, without paying attention to some syntactic and semantic information. ERNIE 1.0 uses both entity-level and phrase-level mask mechanisms, which forces the model to learn to focus on some syntactic and semantic information.
- **K-BERT**: K-BERT is a model that incorporates a knowledge graph in the input process. k-BERT first identifies the entity in the sentence and then obtains the triples from the knowledge graph with that entity as the head entity. Also, to avoid introducing knowledge noise, K-BERT applies the mask operation to self-attention

4.2 Models Settings

In our baselines, model parameters are trained by AdamW optimizer with batch size 16, learning rate $5e^{-5}$, dropout rate 0.5. The number of epochs is set to 10 with early stopping strategy.

4.3 Results Analysis

Table 1 shows the experiment results. We find that the BERT model shows lower performance than graph-based models GCN(with Word2Vec, node initialization) in test dataset with separately 5% on F1-score. However, ERNIE 1.0 and K-BERT did not perform as well as expected compared to BERT, possibly because the external knowledge involved in our dataset is more complex than the knowledge graph pre-trained by the pretrain model.

5 Conclusion & Future work

We present a knowledge graph enhanced article quality identification dataset that differs from a common text classification task. Ultimately, the introduction of external knowledge can be found in the experimental results to improve the effectiveness of text classification. In future, we will process more experiments comparing this method with other methods on common text classification or NLP tasks to prove its general purpose.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018.

[2] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. CoRR, abs/1906.08237, 2019.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692, 2019.

[1] https://ai.tencent.com/ailab/nlp/en/embedding.html
[4] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph, 2019.

[5] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[6] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. CoRR, abs/1609.02907, 2016.

[7] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlou, and Jianfeng Gao. Deep learning based text classification: A comprehensive review. CoRR, abs/2004.03705, 2020.

[8] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013.

[9] Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. CoRR, abs/1310.4546, 2013.

[10] Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. A deep architecture for semantic matching with multiple positional sentence representations. CoRR, abs/1511.08277, 2015.

[11] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi-perspective matching for natural language sentences. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pages 4144–4150, 2017.

[12] Hoa T. Le, Christophe Cerisara, and Alexandre Denis. Do convolutional networks need to be deep for text classification? CoRR, abs/1707.04108, 2017.

[13] Bao Guo, Chunxia Zhang, Junmin Liu, and Xiaoyi Ma. Improving text classification with weighted word embeddings via a multi-channel textcnn model. Neurocomputing, 363:366–374, 2019.

[14] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. CoRR, abs/1809.05679, 2018.

[15] Seonhoon Kim, Jin-Hyuk Hong, Inho Kang, and Nojun Kwak. Semantic sentence matching with densely-connected recurrent and co-attentive information. CoRR, abs/1805.11360, 2018.

[16] Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, Hrur Harutyunyan, Nazanin Alipourfard, Kristina Lerman, Greg Ver Steeg, and Aram Galstyan. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. CoRR, abs/1905.00067, 2019.

[17] Jun Wu, Jingrui He, and Jiejun Xu. Demo-net: Degree-specific graph neural networks for node and graph classification. CoRR, abs/1906.02319, 2019.

[18] Hongyang Gao and Shuiwang Ji. Graph u-nets. CoRR, abs/1905.05178, 2019.

[19] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), Apr. 2018.

[20] Thomas N Kipf and Max Welling. Variational graph auto-encoders. NIPS Workshop on Bayesian Deep Learning, 2016.

[21] Jiaxuan You, Rex Ying, and Jure Leskovec. Position-aware graph neural networks. CoRR, abs/1906.04817, 2019.

[22] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. CoRR, abs/1806.01973, 2018.

[23] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Yihong Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. CoRR, abs/1902.07243, 2019.

[24] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. CoRR, abs/1606.09375, 2016.

[25] Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In Proceedings of the 2018 World Wide Web Conference, WWW ’18, page 1063–1072, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[26] Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. Cognitive graph for multi-hop reading comprehension at scale. CoRR, abs/1905.05460, 2019.
[27] Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. Multi-hop reading comprehension through question decomposition and rescoring. *CoRR*, abs/1906.02916, 2019.

[28] Jasmijn Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima’an. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1957–1967, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[29] Daniel Beck, Gholamreza Haffari, and Trevor Cohn. Graph-to-sequence learning using gated graph neural networks. *CoRR*, abs/1806.09835, 2018.

[30] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019.