Neural News Recommendation with Topic-Aware News Representation

Chuhan Wu[1], **Fangzhao Wu**[2], **Mingxiao An**[3], **Yongfeng Huang**[1], **and Xing Xie**[2]
[1]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]Microsoft Research Asia, Beijing 100080, China
[3]University of Science and Technology of China, Hefei 230026, China
wuch15@mails.tsinghua.edu.cn, {fangzwu,xingx}@microsoft.com,
anmx@mail.ustc.edu.cn, yfhuang@tsinghua.edu.cn

Abstract

News recommendation can help users find interested news and alleviate information overload. The topic information of news is critical for learning accurate news and user representations for news recommendation. However, it is not considered in many existing news recommendation methods. In this paper, we propose a neural news recommendation approach with topic-aware news representations. The core of our approach is a topic-aware news encoder and a user encoder. In the news encoder we learn representations of news from their titles via CNN networks and apply attention networks to select important words. In addition, we propose to learn topic-aware news representations by jointly training the news encoder with an auxiliary topic classification task. In the user encoder we learn the representations of users from their browsed news and use attention networks to select informative news for user representation learning. Extensive experiments on a real-world dataset validate the effectiveness of our approach.

1 Introduction

Online news platforms such as Google News and MSN News have attracted hundreds of millions of users to read news online (Das et al., 2007; Lavie et al., 2010). Massive news are generated everyday, making it impossible for users to read all news to find their interested content (Phelan et al., 2011). Thus, personalized news recommendation is very important for online news platforms to help users find their interested news and alleviate information overload (IJntema et al., 2010).

Learning accurate representations of news and users is critical for news recommendation (Wu et al., 2019b,a). Several deep learning based methods have been proposed for this task (Okura et al., 2017; Wang et al., 2018; Kumar et al., 2017; Khattar et al., 2018; Zheng et al., 2018). For example, Okura et al. (2017) proposed to learn news representations from news bodies via denoising autoencoders, and learn user representations from the representations of their browsed news via a gated recurrent unit (GRU) network. Wang et al. (2018) proposed to learn news representations from news titles via a knowledge-aware convolutional neural network (CNN), and learn user representations from news representations using the similarity between candidate news and browsed news. However, these methods do not take the topic information of news into consideration.

Title	Topic
James Harden's incredible heroics lift Rockets over Warriors	Sports
These Are Some of The Safest Airlines in the World	Travel
Weekend snowstorm forecast from Midwest to East Coast	Unlabeled

Figure 1: Three example news articles.

Our work is motivated by the following observations. First, the topic information of news is useful for news recommendation. For example, if a user clicks many news with the topic "sport", we can infer she is probably interested in sports. Thus, exploiting the topic information of news has the potential to learn more accurate news and user representations. Second, not all news articles contain topic labels, since it is very expensive and time-consuming to manually annotate the massive news articles emerging everyday. Thus, it is not suitable to directly incorporate the topic labels of news as model input. Third, different words in the same news may have different informativeness in representing news. For example, in Fig. 1 the word "Airlines" is more informative than "Some". Besides, different news may also have different importance for user representation. For instance, the first news in Fig. 1 is more informative than the third one in inferring the interest of users.

In this paper, we propose a neural news recom-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1154–1159
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

mendation approach with topic-aware news representations (TANR) which exploit the useful topic information in news. The core of our approach is a topic-aware news encoder and a user encoder. In the news encoder, we learn the representations of news from their titles by capturing the local contexts via CNNs. Since different words may have different informativeness for news representation, we apply attention network to select important words for news representation learning. In addition, we propose to learn topic-aware news representations by jointly training the news encoder with an auxiliary topic classification task. In the user encoder, we learn representations of users from the representations of their browsed news. Since different news may have different informativeness for user representation, we apply attention network to select informative news for user representation learning. Extensive experiments are conducted on a real-world dataset. The results show our approach can effectively improve the performance of news recommendation.

2 Our Approach

In this section, we first introduce our basic neural news recommendation model. Then we introduce how to learn topic-aware news representations.

2.1 Neural News Recommendation Model

The architecture of our basic neural news recommendation model is shown in Fig. 2. It consists of three major modules, i.e., *news encoder*, *user encoder* and *click predictor*.

News Encoder. The *news encoder* module is used to learn representations of news from their titles. It contains three layers. The first one is word embedding, which converts a news title from a word sequence into a vector sequence. Denote a news title as $[w_1, w_2, ..., w_M]$, where M is title length. It is converted into word vector sequence $[\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_M]$ via a word embedding matrix.

The second layer is a CNN network (Kim, 2014). Local contexts are important for understanding news titles. For example, in the news title "90th Birthday of Mickey mouse", the local contexts of "mouse" such as "Mickey" is useful for inferring it is a comic character name. Thus, we use CNN to learn contextual word representations by capturing local contexts. The CNN layer takes the word vectors as input, and outputs the contextual word representations $[\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_M]$.

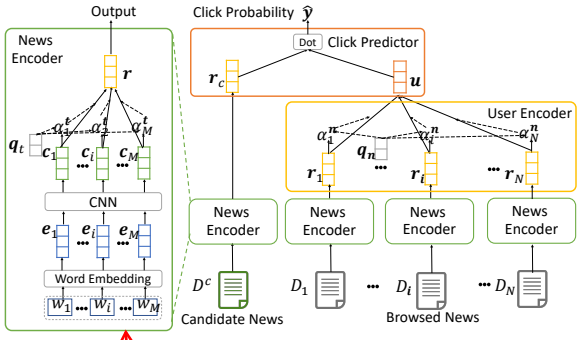

Figure 2: The framework of the basic model.

输入时新闻标题

The third layer is an attention network. Different words in the same news title may have different importance in representing news. For example, in the first news of Fig. 1, the word "Rockets" is more informative than "over" for news representation. Thus, we propose to use attention mechanism to select important words in news titles to learn informative news representations. Denote the attention weight of the i_{th} word in a news title as α_i^t:

$$a_i^t = \mathbf{q}_t^T \tanh(\mathbf{V}_t \times \mathbf{c}_i^t + \mathbf{v}_t), \quad (1)$$

$$\alpha_i^t = \frac{\exp(a_i^t)}{\sum_{j=1}^{M} \exp(a_j^t)}, \quad (2)$$

where $\mathbf{V}_t$ and $\mathbf{v}_t$ are parameters, $\mathbf{q}_t$ is the attention query vector. The final representation of a news title is the summation of the contextual representations of its words weighted by their attention weight, i.e., $\mathbf{r} = \sum_{i=1}^{M} \alpha_i^t \mathbf{c}_i$.

User Encoder. The *user encoder* module is used to learn the representations of users from the representations of their browsed news. Different news browsed by the same user may have different informativeness for representing this user. For example, the news "The best movies in 2018" is more informative than the news "Winter storms next week" in inferring user interests. Therefore, we apply a news attention network to select important news to learn more informative user representations. Denote the attention weight of the i_{th} browsed news as α_i^n:

$$a_i^n = \mathbf{q}_n^T \tanh(\mathbf{V}_n \times \mathbf{r}_i + \mathbf{v}_n), \quad (3)$$

$$\alpha_i^n = \frac{\exp(a_i^n)}{\sum_{j=1}^{N} \exp(a_j^n)}, \quad (4)$$

where $\mathbf{q}_n$, $\mathbf{V}_n$ and $\mathbf{v}_n$ are the parameters, and N is the number of browsed news. The final repre-

sentation of a user is the summation of the representations of her browsed news weighted by their attentions, i.e., $\mathbf{u} = \sum_{i=1}^{N} \alpha_i^n \mathbf{r}_i$.

Click Predictor. The *click predictor* module is used to predict the probability of a user clicking a candidate news based on their hidden representations. Denote the representation of a candidate news D^c as $\mathbf{r}_c$. Following (Okura et al., 2017), the click probability score $\hat{y}$ is calculated by the inner product of the representation vectors of the user and the candidate news, i.e., $\hat{y} = \mathbf{u}^T \mathbf{r}_c$.

Motivated by (Huang et al., 2013), we propose to use negative sampling techniques for model training. For each news browsed by a user (denoted as positive sample), we randomly sample K news displayed in the same impression but not click by this user as negative samples. We then jointly predict the click probability scores of the positive news $\hat{y}^+$ and the K negative news $[\hat{y}_1^-, \hat{y}_2^-, ..., \hat{y}_K^-]$. In this way, we formulate the news click prediction problem as a pseudo $K+1$-way classification task. The posterior click probability of a positive sample is calculated as follows:

$$p_i = \frac{\exp(\hat{y}_i^+)}{\exp(\hat{y}_i^+) + \sum_{j=1}^{K} \exp(\hat{y}_{i,j}^-)}. \quad (5)$$

The loss function for news recommendation is the negative log-likelihood of all positive samples:

$$\mathcal{L}_{NewsRec} = -\sum_{i \in \mathcal{S}} \log(p_i), \quad (6)$$

where $\mathcal{S}$ is the set of positive training samples.

2.2 Topic-Aware News Encoder

The topic information of news is useful for news recommendation. For example, if a user browses many news with the topic "sport", then she may be interested in sports. Thus, exploiting the news topics has the potential to improve the representations of news and users. However, not all news in online news platforms contain topic labels, since it is very costly and time-consuming to annotate the massive news emerging everyday. Thus, instead of incorporating news topics as model input, we propose to learn topic-aware news encoder which can extract topic information from news titles by jointly training it with an auxiliary news topic classification task, as shown in Fig. 3. We propose a news topic classification model for this task, which consists of a *news encoder* module and a *topic predictor* module. The *news encoder* module is shared with the news recommendation model. The *topic predictor* is used to predict the topic probability distribution from news representation as follows:

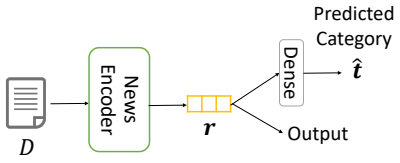

Figure 3: The framework of topic-aware news encoder.

$$\hat{\mathbf{t}} = softmax(\mathbf{W}_t \times \mathbf{r} + \mathbf{b}_t), \quad (7)$$

where $\mathbf{W}_t$ and $\mathbf{b}_t$ are parameters, and $\hat{\mathbf{t}}$ is the predicted topic distribution. The loss function of the topic classification task is formulated as follows:

$$\mathcal{L}_{Topic} = -\frac{1}{N_t} \sum_{i=1}^{N_t} \sum_{k=1}^{K_c} t_{i,k} \log(\hat{t}_{i,k}), \quad (8)$$

where N_t is the number of news with topic labels, K_c is the number of topic categories, and $t_{i,k}$ and $\hat{t}_{i,k}$ are the gold and predicted probability of the ith news in the k-th topic category.

We jointly train the news recommendation and topic classification tasks. The overall loss function is a weighted summation of the news recommendation and topic classification losses:

$$\mathcal{L} = \mathcal{L}_{NewsRec} + \lambda \mathcal{L}_{Topic}, \quad (9)$$

where λ is a positive coefficient. Since the news recommendation and the topic classification tasks share the same news encoder, via joint training, the news recommendation model can capture the topic information to learn topic-aware news and user representations for news recommendation.

3 Experiments

3.1 Datasets and Experimental Settings

We conducted experiments on a real-world dataset which is collected from MSN News[1] logs in one month (from 12/13/2018 to 01/12/2019). The basic statistics of this dataset are summarized in Table 1. In addition, the topic distributions in our dataset are illustrated in Fig. 4. We used the logs in the last week for test, and the rest for training. Besides, we randomly sampled 10% of training data as the validation set.

[1]https://www.msn.com/en-us/news

# users	10,000	avg. # words per title	11.29
# news	42,255	# topic categories	14
# impressions	445,230	# positive samples	489,644
# samples	7,141,584	# negative samples	6,651,940

Table 1: Statistics of our dataset.

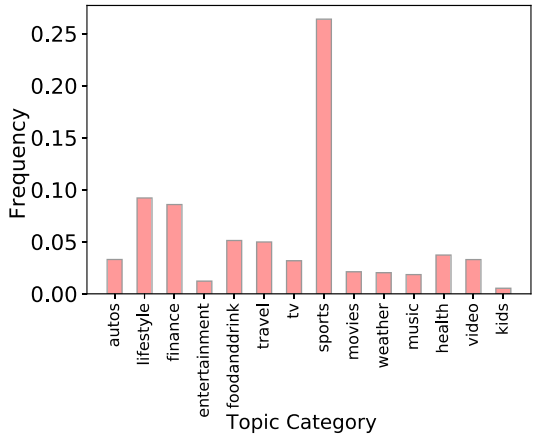

Figure 4: Topic distributions in our dataset.

In our experiments, word embeddings are 300-dimensional and were initialized by the pre-trained Glove embedding (Pennington et al., 2014). The CNN network has 400 filters, and their window sizes are 3. The negative sampling ratio K is 4 and the coefficient λ is 0.2. Adam (Kingma and Ba, 2014) is used as the optimization algorithm, and the batch size is 64. These hyperparameters were selected according to the validation set. The metrics used for result evaluation in our experiments include AUC, MRR, nDCG@5 and nDCG@10. We repeated each experiment 10 times and reported the average results.

3.2 Performance Evaluation

We evaluate the performance of our *TANR* approach by comparing it with several baseline methods, including: (1) *LibFM* (Rendle, 2012), a feature based matrix factorization method for recommendation; (2) *CNN* (Kim, 2014), using Kim CNN to learn news representations from news titles, and building user representations via max pooling; (3) DSSM (Huang et al., 2013), using the deep structured semantic model by regarding the concatenation of browsed news titles as the query and candidate news as the documents; (4) *Wide&Deep* (Cheng et al., 2016), a combination of a wide linear channel and a deep neural network channel; (5) *DeepFM* (Guo et al., 2017), a combination of factorization machines and neural networks; (6) *DFM* (Lian et al., 2018), a deep fusion model by combining dense layers with different depths and using attention mechanism to select important features; (7) *GRU* (Okura et al., 2017), using autoencoders to learn news representations and using a GRU network to learn user representations; (8) *DKN* (Wang et al., 2018), a neural news recommendation method which can utilize entity information in knowledge graphs via a knowledge-aware CNN; (9) *TANR-basic*, our basic neural news recommendation model; (10) *TANR*, our approach with topic-aware news representations. The results of different methods are summarized in Table 2.

Methods	AUC	MRR	nDCG@5	nDCG@10
LibFM	0.5660	0.2924	0.3015	0.3932
CNN	0.5689	0.2956	0.3043	0.3955
DSSM	0.6009	0.3099	0.3261	0.4185
Wide&Deep	0.5735	0.2989	0.3094	0.3996
DeepFM	0.5774	0.3031	0.3122	0.4019
DFM	0.5860	0.3034	0.3175	0.4067
DKN	0.5869	0.3044	0.3184	0.4071
GRU	0.6102	0.2811	0.3035	0.3952
TANR-basic	0.6221	0.3246	0.3487	0.4329
TANR*	**0.6289**	**0.3315**	**0.3544**	**0.4392**

Table 2: The results of different methods. *The improvement is significant at $p < 0.01$.

From Table 2, We have several observations. First, the methods based on neural networks (e.g., *CNN*, *DSSM* and *TANR*) outperform *LibFM*. This is because neural networks can learn better news and user representations than traditional matrix factorization methods. Second, both *TANR-basic* and *TANR* can outperform many baseline methods. This is because our approaches can select important words and news for learning informative news and user representations via a hierarchical attention network, which is not considered in baseline methods. Third, *TANR* consistently outperforms *TANR-basic*. It validates the news topics are useful for news recommendation, and our approach can effectively exploit the topic information.

Then, we want to evaluate the performance of our approach in topic classification. The performance in Fscore over each topic category is shown in Fig. 5. From Fig. 5, we find the classification of most topic classes is satisfactory, except for the class "kids". This may be because the training data in this class is too scarce and difficult to be recognized. These results show that our approach can capture useful topic information by training the news encoder with an auxiliary topic classification task to learn topic-aware news representations.

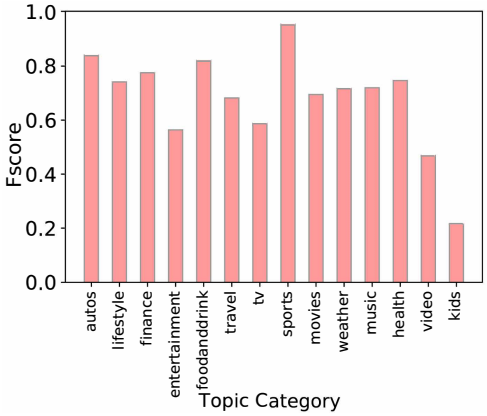

Figure 5: Performance of topic classification.

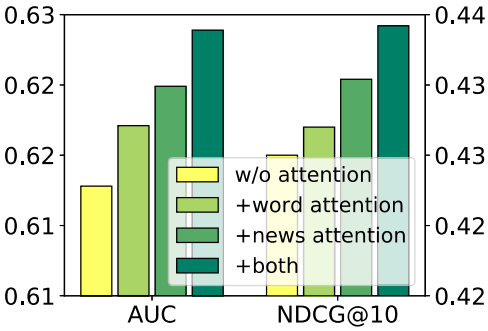

Figure 6: Effectiveness of different attention networks.

3.3 Effectiveness of Hierarchical Attention

We conducted experiments to explore the hierarchical attentions in our approach. The results are shown in Fig. 6. We find the news-level attention network can effectively improve the performance of our approach. This is because different news usually have different informativeness in representing users, and selecting important news can help learn more informative user representations. In addition, the word-level attention network is also useful. This is because different words usually have different importance for representing news, and our approach can select important words to learn informative news representations. Moreover, combining both attention networks can further improve the performance of our approach. These results validate the effectiveness of hierarchical attentions in our approach.

3.4 Influence of Hyperparameter

In this section, we explore the influence of the coefficient λ in Eq. (9) on our approach. It controls the relative importance of the topic classification task. The results are shown in Fig. 7. We find if λ is too small, the performance of our approach

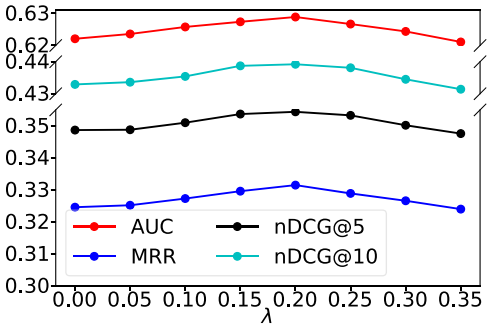

Figure 7: Influence of the hyperparameter λ.

is not optimal, since the useful topic information is not fully exploited. Thus, the performance improves when λ increases from 0. However, when λ becomes too large, the performance of our approach declines. This is because the topic classification task is over-emphasized and the news recommendation task is not fully respected. A moderate value of λ (e.g., 0.2) is the most appropriate.

4 Conclusion

In this paper, we propose a neural news recommendation approach with topic-aware news representations. In our approach we propose a new encoder to learn news representations from news titles, and use attention network to select important words. In addition, we propose to train a topic-aware news encoder via jointly training it with an auxiliary topic classification task to extract the topic information in news. In addition, we propose a user encoder to learn representations of users from their browsed news, and use attention network to select important news. Extensive experiments on a real-world dataset validate the effectiveness of our approach.

Acknowledgments

The authors would like to thank Microsoft News for providing technical support and data in the experiments, and Jiun-Hung Chen (Microsoft News) and Ying Qiao (Microsoft News) for their support and discussions. This work was supported by the National Key Research and Development Program of China under Grant number 2018YFC1604002, the National Natural Science Foundation of China under Grant numbers U1836204, U1705261, U1636113, U1536201, and U1536207, and the Tsinghua University Initiative Scientific Research Program under Grant number 20181080368.

References

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *DLRS*, pages 7–10. ACM.

Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *WWW*, pages 271–280. ACM.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: a factorization-machine based neural network for ctr prediction. In *AAAI*, pages 1725–1731. AAAI Press.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *CIKM*, pages 2333–2338.

Wouter IJntema, Frank Goossen, Flavius Frasincar, and Frederik Hogenboom. 2010. Ontology-based news recommendation. In *Proceedings of the 2010 EDBT/ICDT Workshops*, page 16. ACM.

Dhruv Khattar, Vaibhav Kumar, Vasudeva Varma, and Manish Gupta. 2018. Weave& rec: A word embedding based 3-d convolutional network for news recommendation. In *CIKM*, pages 1855–1858. ACM.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Vaibhav Kumar, Dhruv Khattar, Shashank Gupta, and Vasudeva Varma. 2017. Word semantics based 3-d convolutional neural networks for news recommendation. In *2017 IEEE International Conference on Data Mining Workshops*, pages 761–764.

Talia Lavie, Michal Sela, Ilit Oppenheim, Ohad Inbar, and Joachim Meyer. 2010. User attitudes towards news content personalization. *International journal of human-computer studies*, 68(8):483–495.

Jianxun Lian, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2018. Towards better representation learning for personalized news recommendation: a multi-channel deep fusion approach. In *IJCAI*, pages 3805–3811.

Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *KDD*, pages 1933–1942. ACM.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Owen Phelan, Kevin McCarthy, Mike Bennett, and Barry Smyth. 2011. Terms of a feather: Content-based news recommendation and discovery using twitter. In *ECIR*, pages 448–459. Springer.

Steffen Rendle. 2012. Factorization machines with libfm. *TIST*, 3(3):57.

Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Dkn: Deep knowledge-aware network for news recommendation. In *WWW*, pages 1835–1844.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019a. Neural news recommendation with attentive multi-view learning. In *IJCAI*.

Chuhan Wu, Fangzhao Wu, Junxin Liu, and Yongfeng Huang. 2019b. Npa: Neural news recommendation with personalized attention. In *KDD*. ACM.

Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. Drn: A deep reinforcement learning framework for news recommendation. In *WWW*, pages 167–176.