Frame Semantic-Enhanced Sentence Modeling for Sentence-level Extractive Text Summarization

Yong Guan[1], **Shaoru Guo**[1], **Ru Li** [1,2*], **Xiaoli Li**[3], **Hongye Tan**[1,2],
1. School of Computer and Information Technology, Shanxi University, Taiyuan, China
2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, China
3. Institute for Infocomm Research, A*Star, Singapore
{guanyong0130,guoshaoru0928}@163.com
{liru,tanhongye}@sxu.edu.cn,xlli@i2r.a-star.edu.sg

Abstract

Sentence-level extractive text summarization aims to select important sentences from a given document. However, it is very challenging to model the importance of sentences. In this paper, we propose a novel **F**rame **S**emantic-Enhanced **S**entence Modeling for Extractive **S**ummarization (FS^3), which leverages Frame semantics to model sentences from both *intra-sentence level* and *inter-sentence level*, facilitating the text summarization task. In particular, *intra-sentence level semantics* leverage Frames and Frame Elements to model internal semantic structure within a sentence, while *inter-sentence level semantics* leverage Frame-to-Frame relations to model relationships among sentences. Extensive experiments on two benchmark corpus CNN/DM and NYT demonstrate that FS^3 model outperforms six state-of-the-art methods significantly.

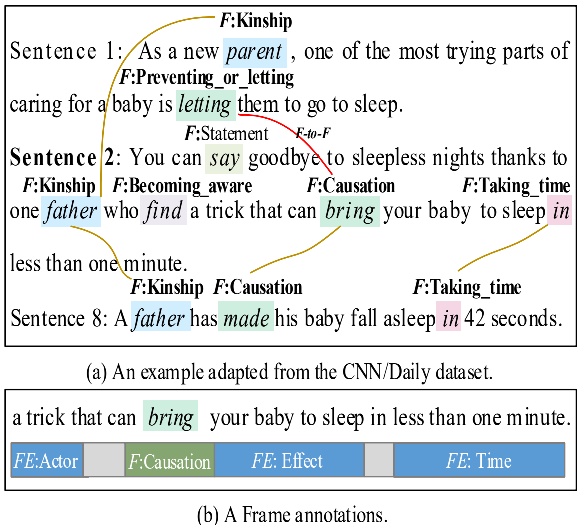

Figure 1: (a) Frame relations across different sentences. (b) Frame semantic structure within a sentence.

1 Introduction

Extractive text summarization selects words, phrases, or sentences from the original text to create a summary (Chan, 2018). In this paper, we focus on sentence-level extractive text summarization, which aims to select important and informative sentences from a given document. A key problem in this task is to model the salience of sentences, focusing on not only the semantic information within sentences (intra-sentence), but also the relationships among sentences (inter-sentences). For example, in Figure 1(a), the sentence 2 is an important sentence and a part of final output summary. At the intra-sentence level, it has key phrases, such as *father*, *bring your baby to sleep*, *less than one minute*. At the inter-sentence level, it has meaningful relations with other sentences 1 and 8.

Recently sequence-to-sequence models have produced promising results on the summarization task, which utilize encoder to obtain the representation of sentences and predict the final summary (Chen and Bansal, 2018; Zhong et al., 2020; Mendes et al., 2019; Zhou et al., 2018; Liu and Lapata, 2019). However, these methods mainly focus on modeling sentences word-by-word and ignore the internal semantic structure within a sentence. Besides, they usually focus on the similarity between sentences and ignore the semantic relationship between sentences.

On the other hand, some graph-based methods have been proposed to model relationships between sentences, by first treating sentences as nodes in a graph and then constructing relationships between nodes (Jin et al., 2020). However, most of these methods mainly construct relations based on surface features, e.g., trigram overlapping (Jia et al., 2020), coreference (Xu et al., 2020), without considering the semantic relations between sentences.

We notice FrameNet (Fillmore et al., 1976; Baker et al., 1998), a semantic database, can be lever-

*Corresponding author: Ru Li.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4045–4052
November 7–11, 2021.

aged to distil semantic structures of sentences by defining some semantic units, such as **Frame** (F), **Frame Element** (FE) and Target (T) (Zhao et al., 2020; Guo et al., 2020). As shown in Figure 1(b), the T word *bring* evokes the Frame *Causation*, which contains three FEs, i.e., *Actor*, *Effect*, *Time*, where the FE *Actor* is filled by phrase *a trick*. It is worth mentioning that FrameNet connects different relevant Frames by defining **Frame-to-Frame (F-to-F) relations** (e.g between Frame *Causation* and *Preventing_or_letting*), which facilitate providing natural and effective ways to model semantic relations among sentences (Guan et al., 2021).

In this paper, we propose FS^3, a novel **F**rame **S**emantic-Enhanced **S**entence Modeling for Extractive **S**ummarization, which incorporates the hierarchical attention mechanism into the Graph Convolutional Network (GCN) (Kipf and Welling, 2017) to model sentences from intra-sentence level and inter-sentence level based on Frame semantics. Specifically, intra-sentence and inter-sentence modelings are applied to capture the semantic structure information within a sentence and the semantic relations among sentences respectively. The contribution of this paper can be summarized as follows.

1. To the best of our knowledge, we are the very first to leverage Frame semantics for extractive summarization. We propose a novel FS^3 method that leverages Frame semantics to model sentences from both *intra-sentence level* and *inter-sentence level*.

2. We incorporate the hierarchical attention mechanism into a graph neural network, which dynamically models the interactions within and among sentences. In particular, *intra-sentence level* leverages Frame and FE to model the structure within a sentence, while *inter-sentence level* leverages F-to-F relations to model relationships among sentences.

3. Extensive experimental results demonstrate FS^3 outperforms six state-of-the-art models on two benchmark data CNN/DM and NYT.

2 Methods

Figure 2 provides an overview of the proposed model, mainly consisting of three key modules:

(1) **Semantic Graph Construction** builds a Frame semantic-based semantic graph $\mathcal{G}$ for a given document $\mathcal{D}$.

(2) **Frame Semantic-Enriched Sentence Modeling** employs a novel attention-enhanced graph neural network to learn the semantic graph representations $\mathcal{C}^g$, which utilizes intra-sentence and inter-sentence modeling to capture the semantic structure information within a sentence and the semantic relations among sentences respectively.

(3) **Prediction** employs the document representation $\mathcal{C}^d$ and graph representation $\mathcal{C}^g$ to predict its summary, i.e. whether we should select a sentence.

2.1 Semantic Graph Construction

Formally, our semantic graph can be formalized as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the node set and $\mathcal{E}$ is the edge set. We utilize the frame semantic parser SEMAFOR (Das et al., 2014) to annotate documents. As shown in Figure 2, there are three types of nodes: Frame nodes $\mathcal{V}^f$, FE nodes $\mathcal{V}^{fe}$ and sentence nodes $\mathcal{V}^{sen}$, i.e., $\mathcal{V}^f \subset \mathcal{V}, \mathcal{V}^{fe} \subset \mathcal{V}, \mathcal{V}^{sen} \subset \mathcal{V}$. In particular, $\mathcal{V}^f$ represents *semantic scenario information* of sentences. $\mathcal{V}^{fe}$ describes *semantic unit* of sentences, which consists of the filled words corresponding to its FE. Different from HAHSum (Jia et al., 2020), which keeps each *individual* word as one node, we design an attention mechanism to *aggregate the filled words into one node* to keep the completeness of information (Pan et al., 2020). For example, in Figure 1(b), the filled words *your baby to sleep* of FE *Effect* are considered as a whole. Finally, each sentence has a node $\mathcal{V}^{sen}$, denoting overall information of the sentence.

To capture rich semantic relationships within and across sentences, we connect the different types of nodes to form three different types of edges in $\mathcal{E}$: First, as a Frame typically consists of several associated FEs, we connect multiple FEs to the corresponding Frame to obtain a better Frame representation. Moreover, as a sentence usually contains multiple Frames that represent several semantic scenario, we connect multiple Frames to their corresponding sentence to capture the enriched semantic representation of the sentence. Finally, the importance of a sentence is reflected in its connections to other sentences. The more connections a sentence has, the more important it is (Page et al., 1998; Mihalcea and Tarau, 2004). Therefore, we connect different sentences according to F-to-F relations.

2.2 Frame Semantic-Enriched Sentence Modeling

We first encode the document $\mathcal{D}$ and Frame F via BERT (Devlin et al., 2019) individually. Then, we

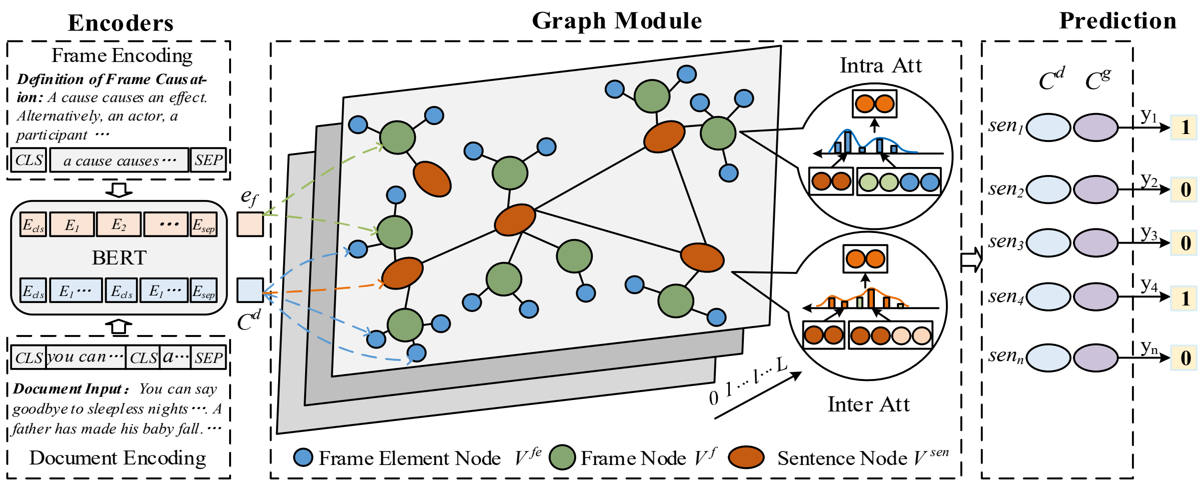

Figure 2: The overview of our proposed FS^3 model.

integrate them into GCN (Kipf and Welling, 2017) to obtain the Frame semantic-enriched sentence representation from both intra-sentence and inter-sentence level for extractive summarization.

2.2.1 Document Encoding

Given a document $\mathcal{D} = \{s_1, s_2, ..., s_n\}$, which contains n sentences, where s_i denotes the i-th sentence of document $\mathcal{D}$. Different from the original BERT, which is trained to encode a single sentence or sentence pair, we insert [CLS] and [SEP] tokens at the beginning and the end of each sentence (refer to bottom left of Figure 2). Then we obtain document representation $\mathcal{C}^d$ by feeding $\mathcal{D}$ into BERT.

$$\mathcal{C}^d = BERT(\mathcal{D}) \tag{1}$$

The token [CLS] representations in $\mathcal{C}^d$ are used as the sentence representations $\mathcal{C}^s$, to initialize the sentence nodes in our semantic graph $\mathcal{G}$.

2.2.2 Frame Encoding

Since each Frame has a unique definition F_{def}, we process the F_{def} into the input format of BERT as: [CLS] F_{def} [SEP] (refer to top left part of Figure 2). Then, we feed the f_{def} into BERT (Chen et al., 2020), and regard the token [CLS] representation as the Frame vector e_f, which initializes the corresponding Frame node in semantic graph $\mathcal{G}$.

2.2.3 Semantic Graph Encoding

As shown in middle part of Figure 2, we stack L graph encoder layers, which incorporate a hierarchical attention mechanism into the GCN to encode the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. At each encoder layer, we sequentially conduct intra- and inter-sentence modeling to update the node states. In this way, the final sentence node states encode the semantic information within and across sentences simultaneously.

Specifically, assume that the representation for the l-th encoder layer through graph convolution operation is represented as $h^l = \{h^l_f, h^l_{fe}, h^l_{sen}\}$, where h^l_f, h^l_{fe}, and h^l_{sen} are the l-th layer node states of $\mathcal{V}^f$, $\mathcal{V}^{fe}$ and $\mathcal{V}^{sen}$, respectively. The learned Frame representation e_f is used as the h^0_f and sentence representations $\mathcal{C}^s$ is used as the h^0_{sen}. Besides, we obtain the initial representation h^0_{fe} by computing the representation of the filled words from $\mathcal{C}^d$ with an attention mechanism. Then the updates of node states $h^l_f = \{h^l_{f_i}\}$, $h^l_{fe} = \{h^l_{fe_j}\}$, and $h^l_{sen} = \{h^l_{sen_k}\}$, consist of the following steps:

Intra-Sentence Modeling. It aims to enrich sentence representation by considering the semantic structure (Frame, FE) within a sentence. Specifically, we design an attention mechanism *Inner Att* to obtain Frame node representation $H^l_{f_i}$ by integrating the information from its associated FE nodes based on their importance (Bahdanau et al., 2015).

$$\alpha_{i,j} = softmax(\mathcal{A}^T_1 \sigma(W^l_1 h^l_{f_i} + W^l_2 h^l_{fe_j})) \tag{2}$$

$$H^l_{f_i} = \sum_{j \in \mathcal{N}_{f_i}} \alpha_{i,j} h^l_{fe_j} \tag{3}$$

Where σ is an activation function, and $\mathcal{N}_{f_i}$ is the associated FE nodes of $\mathcal{V}^f_i$. $\mathcal{A}_1$, W^l_1 and W^l_2 are parameter matrices.

As mentioned in 2.1, a sentence typically contains several Frames and each Frame consists of multiple FEs. We thus leverage the *two level hierarchical semantic information* of Frame and FE to enhance the sentence representation $\bar{H}^l_{sen}$.

$$\beta_{k,i} = \sigma(W^l_3 h^l_{sen_k} + W^l_4 H^l_{f_i} + W^l_5 \sum_{u \in A_{f_i}} h^l_{fe_u}) \quad (4)$$

$$\bar{H}^l_{sen_k} = \sum_{i \in A_{sen_k}} \beta_{k,i} H^l_{f_i} \quad (5)$$

Where A_{sen_k} is Frame nodes of $\mathcal{V}^{sen}_k$, and A_{f_i} is the set of FEs associated with $\mathcal{V}^f_i$.

Inter-Sentence Modeling. It aims to model relationships among sentences. Specifically, we design an attention mechanism *Inter Att* with an element-wise operation to gather the semantic information of a sentence from its related sentence nodes to update its current state $H^l_{sen_k}$ (Kim et al., 2018).

$$a_{k,t} = \rho(W^l_6 \bar{H}^l_{sen_k} + W^l_7 \bar{H}^l_{sen_t}) \quad (6)$$

$$H^l_{sen_k} = \sum_{t \in \mathcal{N}_{sen_k}} a_{k,t} \bar{H}^l_{sen_t} \quad (7)$$

Where $\mathcal{N}_{sen_k}$ is the neighbors of sentence node $\mathcal{V}^{sen}_k$, and ρ is an activation function. After L layers of graph propagation, we finally obtain final graph representation $\mathcal{C}^g = \{H^l_f, h^l_{fe}, H^l_{sen}\}$.

2.3 Prediction

We integrate semantic sentence representation $\mathcal{C}^s$ and sentence node representation H^l_{sen} to predict the oracle labels (refer to the right part of Figure 2).

$$y_i = \delta(FFN(\mathcal{C}^s, H^l_{sen})) \quad (8)$$

Where δ represents the logistic function, FFN is a feed-forward network, and y_i denotes the prediction probability. During training, we minimize a loss function, which is binary classification entropy of prediction y_i against ground-truth label y^*_i.

$$\mathcal{L} = -\frac{1}{|\mathcal{D}|} \sum \log p(y|\mathcal{D}; \theta) \quad (9)$$

3 Experiments

3.1 Datasets & Evaluation Metrics

We evaluate our proposed FS^3 model on two popular benchmark datasets, i.e., CNN/DM (Rush et al., 2015)[1], NYT (Sandhaus, 2008)[2]. For CNN/DM, we employ the standard splits of training, validation, and test (109,962/13,368/11,490), following the previous work (Jia et al., 2020; Liu and Lapata, 2019). For NYT dataset, we follow the same splits as (Durrett et al., 2016), i.e., 100834, 4000 and 9706 samples for training, validation and test.

Following the existing work, the performance is evaluated using standard ROUGE (Lin, 2004), including ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L), which calculates the overlapping lexical units between those extracted sentences and ground-truth summary.

3.2 Parameter Settings

We employ BERT for document encoder, whose implementation is based on the PyTorch version [3]. The implementation of graph encoding is based on DGL with 2 layers. We train our model for 100,000 steps on 2 GPUs (Nvidia Tesla V100, 32G) with gradient accumulation every two steps. We select the top-3 checkpoints according to the evaluation loss on validation set and report the averaged results on the test set. We select Adam as the optimizer, with a learning rate and the dropout probability, setting as 5e-5 and 0.3 respectively.

3.3 Experimental Results

Table 1 lists the comparison results among all seven models on two widely used benchmark datasets, i.e., CNN/DM, NYT. From the table we can see that our model achieves significantly better results on the two datasets. We have the following observations: 1) the basic pre-trained model (BERT) and BERT-based models BertSumExt and MatchSum, perform better than other models, which provides good justifications on why we utilize BERT as our encoder. 2) Our proposed FS^3 achieves better results than six other models consistently across the two datasets in terms of three evaluation metrics (R-1, R-2 and R-L), indicating FS^3 can obtain a better sentence semantic representations by effectively integrating multiple semantic information, such as Frame and FE, using our constructed semantic graph based on GCN.

3.4 Ablation Studies

We perform ablation studies to investigate the influence of two different modules in our FS^3 model. As shown in Table 2, the deceasing results for removing Intra-Sentence Modeling (*-w/o Intra-att*) or Inter-Sentence Modeling (*-w/o Inter-att*) indicate that both *Intra-* and *Inter-* sentence modelings are useful for our model. Not surprisingly, the

[1]https://cs.nyu.edu/ kcho/DMQA/

[2]https://catalog.ldc.upenn.edu/LDC2008T19

[3]https://github.com/huggingface/pytorch-pretrained-BERT

Method	CNN/DM dataset			NYT dataset		
	R-1	R-2	R-L	R-1	R-2	R-L
Lead3 (Liu and Lapata, 2019)	40.42	17.62	36.67	39.58	20.11	35.78
SummaRuNNer (Nallapati et al., 2017)	39.60	16.20	35.30	-	-	-
Exconsumm (Mendes et al., 2019)	41.7	18.6	37.8	43.18	24.43	38.92
BERT (Devlin et al., 2019)	42.46	19.57	38.34	45.25	25.33	40.41
BertSumExt (Liu and Lapata, 2019)	43.85	20.34	39.90	46.66	26.35	42.6
MatchSum (Zhong et al., 2020)	44.41	20.86	40.55	-	-	-
FS^3	**44.72**	**21.38**	**40.87**	**47.32**	**26.87**	**43.25**

Table 1: The comparison results among six state-of-the-art models and our proposed FS^3 model on CNN/DM and NYT datasets.

-w/o both att model, removing the Intra- and Inter-Sentence Modeling, obtains significantly worse performance, signifying our proposed two attention modules are critical for improving the quality of summarization, especially when these two mechanisms as a whole make a greater effect together. Besides, we add case study for further analysis in Appendix A.

Method	R-1	R-2	R-L
FS^3	44.72	21.38	40.87
-w/o Intra-att	44.65	21.31	40.82
-w/o Inter-att	44.57	21.23	40.76
-w/o both att	43.74	20.41	40.13

Table 2: Ablation studies on CNN/DM data.

3.5 Human Evaluation

In addition to automatic evaluation by ROUGE, we also evaluated system output with human judgments (Owczarzak et al., 2012). Following the existing work (Cheng and Lapata, 2016; Narayan et al., 2018), we randomly select 40 examples from CNN/DM test data, and then the participants are presented with an article and summaries generated by five systems (Lead3, BERT, BertSumExt, FS^3 and Ground-Truth (GT)). Each document is annotated by three different participants separately, and we ask the participants to rank the summaries from best to worst in order of informativeness and fluency.

The results are shown in Table 3, unsurprisingly, the summaries of Ground-Truth are considered best and ranked 1st 63% of cases, however closely followed by our FS^3 model which is ranked 1st 42% of the time.

Models	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
Lead3	0.11	0.20	0.31	0.24	0.14
BERT	0.22	0.32	0.21	0.13	0.12
BertSumExt	0.26	0.39	0.15	0.12	0.08
FS^3	0.42	0.36	0.14	0.06	0.02
GT	0.63	0.20	0.14	0.03	0.00

Table 3: Human evaluation on CNN/DM test data.

4 Conclusion

In this paper, we focus on sentence-level extractive summarization which is an important yet challenging task. We propose a novel FS^3 method that leverages Frame semantics to model sentences from *intra-sentence level* and *inter-sentence level*, which can better model semantic information within a sentence and across difference sentences. Extensive experiments on two popular datasets demonstrate its effectiveness on benchmark data for summarization task.

In future work, there are two potential directions for research. Firstly, we will apply multidimensional information (Zhang et al., 2017; Gollapalli et al., 2017) to enhance the text representation. Secondly, to address the problem of coherence, we will apply semantic information to help produce summaries with few or no broken inter-relations.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and suggestions. This work was sponsored by the National Key Research and Development Program of China (No.2018YFB1005103) and National Natural Science Foundation of China (No.61772324, No.62076155).

References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Garrett G Chan. 2018. Extractive text summarization with deep learning.

Xiao Chen, Changlong Sun, Jingjing Wang, Shoushan Li, Luo Si, Min Zhang, and Guodong Zhou. 2020. Aspect sentiment classification with document-level sentiment preference modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3677, Online. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

Dipanjan Das, Desai Chen, André FT Martins, Nathan Schneider, and Noah A Smith. 2014. Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008, Berlin, Germany. Association for Computational Linguistics.

Charles J Fillmore et al. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280, pages 20–32.

Sujatha Das Gollapalli, Xiao-Li Li, and Peng Yang. 2017. Incorporating expert knowledge into keyphrase extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Yong Guan, Shaoru Guo, Ru Li, Xiaoli Li, and Hu Zhang. 2021. Frame semantics guided network for abstractive sentence summarization. *Knowledge-Based Systems*, 221:106973.

Shaoru Guo, Yong Guan, Ru Li, Xiaoli Li, and Hongye Tan. 2020. Incorporating syntax and frame semantics in neural network for machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2635–2641, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ruipeng Jia, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao, and Shi Wang. 2020. Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3622–3631, Online. Association for Computational Linguistics.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Semsum: Semantic dependency guided neural abstractive summarization. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8026–8033. AAAI Press.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1571–1581, Red Hook, NY, USA. Curran Associates Inc.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Afonso Mendes, Shashi Narayan, Sebastião Miranda, Zita Marinho, André F. T. Martins, and Shay B. Cohen. 2019. Jointly extracting and compressing documents with summary state representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3955–3966, Minneapolis, Minnesota. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3075–3081. AAAI Press.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montréal, Canada. Association for Computational Linguistics.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475, Online. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times Annotated Corpus.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Yuxiang Zhang, Yaocheng Chang, Xiaoqing Liu, Sujatha Das Gollapalli, Xiaoli Li, and Chunjing Xiao. 2017. *MIKE: Keyphrase Extraction by Integrating Multidimensional Information*, page 1349–1358. Association for Computing Machinery, New York, NY, USA.

Hongyan Zhao, Ru Li, Xiaoli Li, and Hongye Tan. 2020. Cfsre: Context-aware based on framesemantics for distantly supervised relation extraction. *Knowledge-Based Systems*, 210:106480.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.

A Appendix of Case Study

We further conduct case studies to showcase summaries extracted by different systems, e.g., Lead3, BERT and FS^3. As shown in Table 4, Lead3 inherently captures the first three sentences of the article to form the document. Similarly, BERT selects sentences 1, 3 and 4 in this example, and thus lacks informativeness as it is focused on one part of the document. In contrast, FS^3 extracts summary mentioning *the details and meaning of the tennis competition*, which spans the whole document and captures more comprehensive information.

Document **sen1:** Rafa nadal got his clay court season off to a perfect, confidence-boosting start with a 6-2, 6-1 win over frenchman lucas pouille in the monte carlo masters. **sen2:** It was a businesslike display from the world no 5. **sen3:** He didn't unleash the full power of his forehand but played sensible, measured tennis and made only five unforced errors in the match. **sen4:** His talented 21-year-old opponent showed flashes of attacking flair. ... **sen13:** this victory was the first step in rebuilding his self-belief after disappointing losses in the quarter-finals of the Indian wells masters. ...
Lead3 **sen1:** Rafa nadal got his clay court season off to a perfect, confidence-boosting start with a 6-2, 6-1 win over frenchman lucas pouille in the monte carlo masters. **sen2:** It was a businesslike display from the world no 5. **sen3:** He didn't unleash the full power of his forehand but played sensible, measured tennis and made only five unforced errors in the match.
BERT **sen1:** Rafa nadal got his clay court season off to a perfect, confidence-boosting start with a 6-2, 6-1 win over frenchman lucas pouille in the monte carlo masters. **sen3:** He didn't unleash the full power of his forehand but played sensible, measured tennis and made only five unforced errors in the match. **sen4:** His talented 21-year-old opponent showed flashes of attacking flair.
$\boldsymbol{FS^3}$ **sen1:** Rafa nadal got his clay court season off to a perfect, confidence-boosting start with a 6-2, 6-1 win over frenchman lucas pouille in the monte carlo masters. **sen3:** He didn't unleash the full power of his forehand but played sensible, measured tennis and made only five unforced errors in the match. **sen13:** This victory was the first step in rebuilding his self-belief after disappointing losses in the quarter-finals of the Indian wells masters.

Table 4: An example of summaries extracted by different models. In this case, the results of our FS^3 are the same as the reference summary.