Knowledge-Enhanced Document Summarization with Pointer Generator

Anonymous ACL submission

Abstract

Since knowledge-aware methods have boosted a number of natural language processing applications over the last decades, we focus on researching leveraging semantic knowledge to improve a document summarizer for the multi-document summarization task. In this paper, we study to leverage knowledge from semantic knowledge graphs for abstractive multi-document summarization adopting the pointer-generator model architecture. Specifically, we investigate a manner to integrate two kinds of semantic knowledge into a pointer generator model, to analyze the benefit of each type of semantic knowledge insertion for a multi-document summarization task. We conduct experiments on two datasets, which are benchmarks in the document summarization field: Multi-News and CNN/DailyMail, with four evaluation metrics: ROUGE, BERTScore, duplication rate and novelty rate. Extensive experiments show that our models outperform the baseline pointer generator summarizer in higher ROUGE and BERTScore results, less redundant information, and more novel terms generated, implying an improved method for summarization that can produce competitively concise and informative summaries.

1 Introduction

As the development of information technology has incrementally sparked a large quantity of text data and corpora, there is an increasingly high demand for summarizing text data to assist users in gathering the most important and relevant information quickly and easily. Document summarization is a natural language processing task of generating an abridged version of the given single or multiple documents pursuing as concise and coherent as possible while preserving salient and consistent information from the source text. In contrast to the single-document summarization task, the multi-document summarization task prospers a broader range of real-world applications. Relatively, it requires conquering more challenges, such as content complementary, information overlapping, and conflicting problems. The Transformer-based deep-learning models have recently been increasingly employed in document summarization tasks, especially for multi-document summarization tasks, due to their capabilities to understand the relations among a broader range of sequential elements and parallelise the training, which can achieve higher accuracy while reducing the time cost. Compared to the standard Transformer (Vaswani et al., 2017), the pointer generator (See et al., 2017), which is a hybrid network architecture adjusted with the Transformer attention mechanism and between a baseline sequence-to-sequence model, can excel in summarization efficiency and summary conciseness. Against Transformer, the pointer generator uses an advanced word generation mechanism, generating a summary composed of a combination of copied words via pointing and generated words from a fixed vocabulary with word weights. Rather than only focusing on solo source text distribution to generate the summary, the pointer generator can additionally consider word distributions of the vocabulary set and attention set to acquire final word distribution for predicting summary words. Except for summarizer architecture research, over the last decades, many advanced summarization methods have emerged with knowledge graphs leveraged (Koncel-Kedziorski et al., 2019; Tang et al., 2020; Ji et al., 2020; Huang et al., 2020; Zhu et al., 2021; Ji and Zhao, 2021; Wang et al., 2022; Lyu et al., 2022). However, the empirical investigation on the incorporation of semantic knowledge into the document summarization model is still a potential research direction in the document summarization field. There is no clear or strong evidence that clarifies the benefit of each kind of knowledge for document summarization. Amid the mentioned knowledge that contains variant forms of knowledge, we focus on semantic knowledge as the knowledge for experiments only, which can be categorized into open knowledge and closed knowledge, from unrestricted knowledge graphs and external-source knowledge graphs, respectively. Since semantic knowledge contains the highest-level connotation information, it can significantly profit document summarization tasks.

We conduct experiments on the multi-document summarization task based on an augmented pointer generator model with integrated open knowledge and closed knowledge. Experimental results show that our improved knowledge-enhanced pointer generator summarizer outperforms the baseline pointer generator model considerably on general summarization evaluation metrics, such as the ROUGE and BERTScore scores. Also, it achieves better rate results on the duplication and novelty metrics of its generated summaries. In particular, our knowledge-enhanced pointer generator summarizer achieves 5.88% BERTScore higher on the Multi-News dataset and 3.94% BERTScore higher on the CNN/DailyMail dataset than the baseline pointer generator summarizer. We summarize our main contributions as follows.

- We propose a simple manner to incorporate semantic knowledge into the pointer generator summarizer. Both open semantic knowledge and closed semantic knowledge are experimented with.

- We experiment with the effectiveness of leveraging semantic knowledge in the pointer generator summarizer. Evaluation results show that the knowledge-enhanced pointer generator summarizer achieves better summarization performance than the baseline pointer generator summarizer.

- Empirical results demonstrate that generated summaries contain less redundant information with the open knowledge incorporated and include more novel information with the closed knowledge leveraged on Multi-News and CNN/DailyMail datasets.

2 Method

In this section, we provide details of the method adopted for this research on leveraging semantic knowledge into a multi-document summarization model based on the pointer generator architecture.

2.1 Knowledge Extraction

As for the open knowledge extraction, we utilize the information extractor OpenIE to produce a knowledgeable set of semantic three-element triples. Each extracted triple is in the format such that $\langle subject, relation, object \rangle$. As for the closed knowledge extraction, we crawl closed knowledge from an open-source knowledge graph, ConceptNet, which can be downloaded from the website Amazon Web Services (AWS). Since ConceptNet contains a large number of commonsense information but relatively also massive noisy knowledgeable contents for the summarization task, we follow the works (Ji et al., 2020) to filter out unnecessary concepts from text sequences. Specifically, we merge 42 origin relationships of the ConceptNet into 17 reasonable relationships with reversed links to the graph (Das et al., 2018; Lin et al., 2019; Ji et al., 2020), and apply fuzzy matching with the text pre-processing methods of lemmatization and stop word removal using the Spacy3. In addition, to capture high semantic relations, we follow (Guan et al., 2019; Ji et al., 2020) that only extract nouns and verbs to structure concepts, which are nodes with high occurrences and contributions to the information flow within the knowledge graph.

2.2 Knowledge Processing

In order to further leverage the knowledge information into the pointer generator model, we transform the extracted semantic knowledge into a vocabulary set with word weights. The weight of each word in the vocabulary set is the occurrence of this word in the extracted knowledgeable content. Also, we concatenate the weighted open knowledge vocabulary set to the source vocabulary set for obtaining a set of fused knowledge information in varied ways depending on the test cases. Similarly, we process the extracted concepts from the ConceptNet knowledge graph into a vocabulary set containing the closed knowledge information. This vocabulary set that is comprised of a number of closed knowledge words with their respective weights can be incorporated with the source vocabulary set in different ways depending on the test cases as well. Each vocabulary set constructed by knowledge information contains at least 120,000 words with their associated weights. After the concatenation with the source vocabulary set, each vocabulary set can retain more than 500,000 informative words. The first 5,000 words in the vocabulary set are empha-

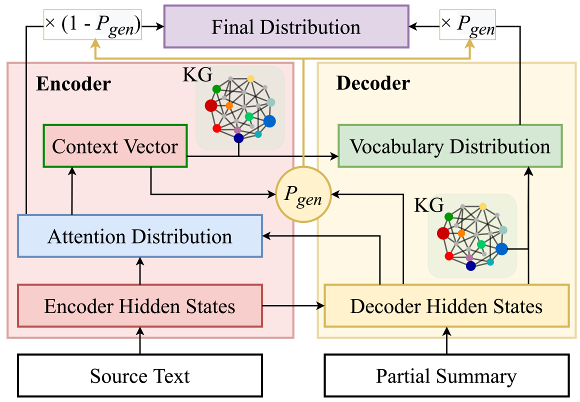

Figure 1: The architecture of an augmented pointer-generator model with the usage of semantic knowledge for multi-document summarization. P_{gen} is the probability for word generation.

sized and used as salient words for the training in the pointer generator model.

2.3 Knowledge Incorporation

Since the standard pointer-generator model has two modules: the encoder and decoder to receive and process the input information, we add the processed knowledge information from knowledge graphs into both of those two modules, in order to enhance the model performance on the multi-document summarization task. The main model architecture of our proposed knowledge-enhanced pointer generator summarizer is illustrated in Figure 1. Specifically, the knowledge information targets to contribute to the vocabulary distribution generation in the pointer generator model, advancing the final distribution for the summary word prediction. In detail, the vocabulary distribution P_{vocab} is a probability distribution covering all vocabulary words, produced as in Equation 1:

$$P_{vocab} = softmax(V'(V[s_t, \sum_i a_i^t h_i] + b) + b') \tag{1}$$

where a_i is the attention distribution, h_i is the encoder hidden state, s_t is the decoder state, V, V', b, and b' are learnable parameters.

Relying on the vocabulary distribution, the final distribution for predicting a summary word w to generate is as in Equation 2:

$$P(w) = P_{vocab}(w) \tag{2}$$

3 Experiments

In this section, we present details of our experiments on the knowledge-enhanced pointer generator model for multi-document summarization.

3.1 Experimental Settings

Datasets. We consider two representative document summarization benchmark datasets in our experiments for summarizer evaluation: Multi-News (Fabbri et al., 2019) and CNN/DailyMail (Hermann et al., 2015). The CNN/DailyMail dataset in our experiments is utilized as a small dataset and operated for testing purposes, which is a sub-dataset with the first 500 test data files of the original large dataset. A statistic summary of applied datasets is shown in Table 1.

Since the Multi-News dataset contains 56,216 pairs of news articles and professionally human-written summaries from over 1,500 sites and has been split into training (80%), validation (10%), and test (10%), which is notably the first large-scale dataset for multi-document summarization on news articles, we experiment with training mainly on the Multi-News dataset. Its average document length is 2,103.49 words, and its average summary length is 263.66 words.

The CNN/DailyMail dataset is a typical single-document summarization dataset but can also be used in multi-document summarization tasks. This dataset is collected from the CNN and Daily-Mail websites. Notably, golden summaries of the CNN/DailyMail dataset comprise a number of bullet points, concluding information aspects contained in those articles. The average document length is 810.57 words, and the average summary length is 56.20 words. In order to evaluate the generalization ability of the summarizer model, we consider and use the CNN/DailyMail dataset as a small-size dataset for experiments.

Evaluation Metrics. In terms of measuring the effectiveness of a document summarizer, there are two benchmarks for the summarizer evaluation: ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020), which are considered. In addition, in order to straightforwardly examine the text quality of the generated summaries, following the previous work (See et al., 2017), we measure and calculate two additional evaluation scores, which are the duplication rate and novelty rate, for judging our improved knowledge-enhanced summarizer.

	Multi-News	**CNN/DailyMail**
# Pairs	44,972 / 5,622 / 5,622	500
Doc. Length		
# Words	2,103.49	810.57
# Sent.	82.73	39.78
Sum. Length		
# Words	263.66	56.20
# Sent.	9.97	3.68

Table 1: Dataset statistics. # denotes the average number in indicating the average length of documents (**Doc.**) and summaries (**Sum.**).

# w	**R-1**	**R-2**	**R-L**	**R-SU**	**BS-F1**
100	27.56	7.27	16.87	7.09	82.71
200	32.62	9.35	18.96	10.68	**83.28**
300	33.41	10.01	19.40	11.49	83.22
400	**34.25**	**11.05**	**20.44**	**11.82**	82.76

Table 2: ROUGE and BERTScore results on the test set of Multi-News dataset with 70,000 training iterations, using the baseline pointer generator (**PG**) model with different settings of the number of output words (**# w**).

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a collection of evaluation indicators and the most prevalent metric for document summarization tasks. In the ROUGE category, the commonly used evaluation metrics are ROUGE-N (i.e., ROUGE-1 and ROUGE-2) and ROUGE-L, standing for ROUGE with N-gram Co-Occurrence Statistics and the longest common sub-sequence algorithm, respectively (Ma et al., 2020). In our experiments, we use ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU values for the evaluation. However, ROUGE is not adequate for the evaluation in some cases. This is because ROUGE scores can continually be high while summaries lack factual consistency, as the ROUGE calculation is generally based on token-level matching rather than semantic-level analysis (Zhu et al., 2021). Hence, we use multiple evaluation metrics for experiments to ensure the evaluation quality.

The BERTScore is an automatic evaluation metric, that computes similarity scores for tokens in candidate summary sentences with tokens in golden summary sentences. In contrast to common metrics, the BERTScore uses pre-trained BERT contextual embeddings to compute the similarity of to (Devlin et al., 2019). Additionally, it is reported that the BERTScore can correlate better with human judgments than existing metrics (Zhang et al., 2020). The precision, recall, and F1 scores of BERTScore results are considered for the summarizer evaluation in our experiments.

The duplication rate of a summary is the percentage of duplicated terms contained in this summary. The textual unit term can be a 1-grams term, 2-grams term, 3-grams term, 4-grams term, or sentence. In addition, the mathematical formula for calculating the duplication rate DUP for a summary Sum_t, which is composed of a number of terms t, is as in Equation 3:

$$DUP = \frac{N_{dt}}{N_t} = \frac{\#\{S_t\} - \#\{s \mid s \in S_t\}}{\#\{S_t\}} \quad (3)$$

where N_{dt} is the number of duplicated terms appeared in one summary, and N_t is the number of terms in that summary.

Analogously to the duplication rate, the novelty rate of a summary is the percentage of new terms contained in this summary in contrast to the source text. The textual unit term can be a 1-grams term, 2-grams term, 3-grams term, 4-grams term, or sentence. Moreover, the mathematical formula for calculating the novelty rate NOV for a summary S_t of a source text $Text$, which is composed by a number of terms t, is as in Equation 4:

$$NOV = \frac{N_{nt}}{N_t} = \frac{\#\{t \mid t \notin Text\}}{\#\{S_t\}} \quad (4)$$

where N_{nt} is the number of novel terms that appeared in one summary, and N_t is the number of terms in that summary.

Implementation Details. We set the maximal number of words for generating summaries to be 300 words with training and testing on the Multi-News dataset. This setting refers to the average length of golden summaries and our experimental results as shown in Table 2. Also, all experiment results presented are obtained without activating the coverage functionality, which can significantly profit to eliminate redundant texts from generated summaries, of the baseline pointer generator model. The purpose of this is to better observe the performance of leveraging semantic knowledge into document summarization models.

3.2 Results

We test overall six cases of leveraging two types of semantic knowledge into a multi-document summarization model. Information on each test case is listed as follows:

1. *S+OK*: The summarizer utilizes the source information with an additional set of word information extracted from the source corpus, which implies the open knowledge.

2. *SwOK*: The summarizer utilizes the additionally weighted source information, which is enhanced by adding the weights of the open knowledge extracted from the source corpus.

3. *OK*: The summarizer utilizes the open knowledge sorely without the source information.

4. *S+CK*: The summarizer utilizes the source information with an additional set of word information extracted from the open-source knowledge graph ConceptNet, which implies the closed knowledge.

5. *SwCK*: The summarizer utilizes the additionally weighted source information, which is enhanced by adding the weights of the closed knowledge extracted from the open-source knowledge graph ConceptNet.

6. *CK*: The summarizer utilizes the closed knowledge sorely without any source information from the source text.

10,000 Training Iterations on Multi-News. The ROUGE results of each test case on the Multi-News dataset for the multi-document summarization task are shown in Table 3. It implies that with a number of training iterations, the performance of the knowledge-enhanced summarizer with the source information on the ROUGE evaluation metric is prominently better than the baseline pointer generator model. Also, Table 3 presents a similar outcome that knowledge-enhanced summarization models mostly can achieve a higher performance than the baseline pointer generator summarizer from the BERTScore evaluation perspective. As for the duplication and novelty evaluations, experimental results are provided in Table 4. From the results, we observe that the knowledge-enhanced summarizer with the open knowledge stands out in the duplication reduction, and the summarizer with the closed knowledge can surpass the summary novelty more than others. In addition, the knowledge-enhanced summarizer with the enhanced source information by the open knowledge achieved the highest performance, indicating the effectiveness of this combination manner between the source information and the open knowledge.

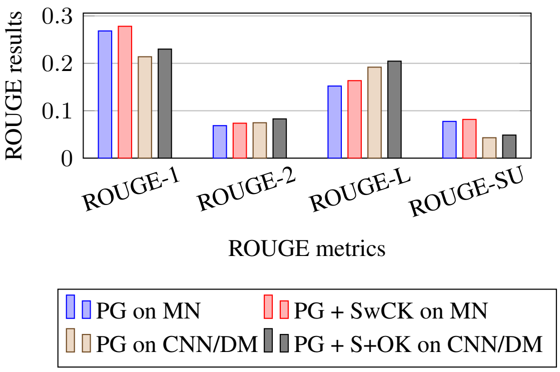

Figure 2: Comparisons of ROUGE results between the baseline pointer generator model and each the best implementation incorporated with semantic knowledge on Multi-News and CNN/DailyMail datasets. **CK** and **OK** denote the closed and open knowledge, respectively.

10,000 Training Iterations on CNN/DailyMail. The ROUGE results of each test case on the CNN/DailyMail dataset for the multi-document summarization task are shown in Table 5. The experiments are trained by the Multi-News training dataset and tested on the CNN/DailyMail dataset. It implies that with a number of training iterations, the performance and generalization ability of the knowledge-enhanced summarizer with the source information on the ROUGE evaluation metric is also better than the baseline pointer generator model. Besides, Table 5 presents a similar outcome that the knowledge-enhanced summarization model can achieve a higher performance than the baseline pointer generator summarizer from the BERTScore evaluation perspective. As for the duplication and novelty evaluations, experimental results are provided in Table 6. From the results, we can also see that the knowledge-enhanced summarizer with the open knowledge stands out in the duplication reduction, and the summarizer with the closed knowledge surpasses the summary novelty more than others. In addition, the knowledge-enhanced summarizer with the enhanced source information by the open knowledge achieved the highest performance, indicating the effectiveness of this combination manner between the source information and the open knowledge.

Discussion. According to the above empirical results, intuitive comparisons of ROUGE scores between the baseline pointer generator model and each of the best improvements incorporated with semantic knowledge on Multi-News and CNN/DailyMail datasets are shown in Figure 2.

	R-1	**R-2**	**R-L**	**R-SU**	**BERT-P**	**BERT-R**	**BERT-F1**
Source	26.83	6.86	15.21	7.75	79.91	78.55	79.20
S+OK	27.52	6.94	15.81	8.15	80.21	78.82	79.47
SwOK	27.15	7.08	15.90	7.70	80.11	**80.48**	**80.26**
OK	26.49	6.65	15.24	7.78	76.43	76.90	76.65
S+CK	23.59	**8.26**	**21.00**	5.00	**80.70**	78.89	79.76
SwCK	**27.81**	7.37	16.35	**8.17**	77.56	77.46	77.48
CK	14.16	3.45	8.32	3.74	78.23	72.08	74.94

Table 3: ROUGE and BERTScore results on the test set of Multi-News dataset with 10,000 training iterations.

	Duplication					**Novelty**				
	1-gs	**2-gs**	**3-gs**	**4-gs**	**sents**	**1-gs**	**2-gs**	**3-gs**	**4-gs**	**sents**
Source	77.29	69.12	66.94	65.57	34.09	55.10	93.89	99.61	99.95	3.74
S+OK	75.11	65.61	62.98	61.28	26.39	54.76	93.09	99.65	99.95	3.31
SwOK	70.90	60.51	56.74	54.15	21.67	52.11	91.44	99.27	99.90	3.69
OK	73.27	63.49	60.60	58.80	27.31	52.42	92.20	99.36	99.94	3.91
S+CK	79.50	72.27	71.05	70.35	34.91	55.91	94.11	99.73	99.96	0.43
SwCK	41.85	7.15	1.36	0.36	**0.00**	52.95	92.21	99.49	99.88	**4.12**
CK	89.07	85.02	84.20	83.72	17.34	**78.69**	**97.31**	**99.92**	**99.99**	1.90
Ref.	**41.70**	**7.09**	**1.27**	**0.33**	**0.00**	55.88	94.21	99.75	99.98	0.43

Table 4: Duplication and novelty rates on the test set of Multi-News dataset with 10,000 training iterations.

As for the Multi-News dataset, the implementation of the pointer generator model with the enhanced weights from the closed knowledge achieved high ROUGE results. Specifically, the results are about 3.65% higher on ROUGE-1, 7.43% higher on ROUGE-2, 7.50% higher on ROUGE-L, and 5.42% higher on ROUGE-SU than the baseline pointer generator model. In addition, the implementation with additional closed knowledge achieved high ROUGE results for the CNN/DailyMail dataset, which are about 18.65% higher on ROUGE-1 and 64.91% higher on ROUGE-SU. However, the improvements in BERTScore results are slight, as about 1.34% higher on BERTScore F1 result for the Multi-News dataset and 0.56% higher on BERTScore F1 result for the CNN/DailyMail dataset. Overall, the empirical results indicate that implementations of leveraging closed knowledge can profit from a high evaluation result on both ROUGE and BERTScore. This may be due to limitations of tools and mechanisms when extracting open semantic knowledge from the source text, and also the lack of efficient techniques to evaluate and determine the effectiveness value of extracted open semantic knowledge for summarization.

Moreover, comparisons of duplication rates between the baseline pointer generator model and each of the best improvements incorporated with semantic knowledge on Multi-News and CNN/DailyMail datasets are shown in Figure 3. The implementation of an augmented pointer generator summarizer with enhanced source information weights by closed knowledge presents its effectiveness for the summarization task on both the Multi-News dataset and the CNN/DailyMail dataset, since it eliminates about an average of 81.82% duplicates in the form of N-grams on the Multi-News dataset, and about an average of 90.48% duplicates in the form of N-grams on the CNN/DailyMail dataset. The empirical results indicate that implementations of leveraging open knowledge can also profit from a high performance on reducing redundant information in the generated summaries, improving the summary conciseness. However, even for the best-improved implementation, its average duplication rate is still about 0.65%

	R-1	**R-2**	**R-L**	**R-SU**	**BERT-P**	**BERT-R**	**BERT-F1**
Source	21.39	7.47	19.19	4.31	85.53	81.57	83.48
S+OK	23.00	**8.28**	**20.46**	4.87	85.46	81.78	83.57
SwOK	20.48	6.65	18.29	3.99	84.80	81.28	82.99
OK	20.64	6.92	18.26	4.10	85.06	81.23	83.09
S+CK	**25.38**	5.80	14.02	**7.11**	**86.21**	**81.84**	**83.95**
SwCK	21.06	7.18	18.83	4.22	84.41	80.45	82.36
CK	21.44	7.38	19.22	4.42	85.56	80.55	82.94

Table 5: ROUGE and BERTScore results on the test set of CNN/DailyMail dataset with 10,000 training iterations.

	Duplication					**Novelty**				
	1-gs	**2-gs**	**3-gs**	**4-gs**	**sents**	**1-gs**	**2-gs**	**3-gs**	**4-gs**	**sents**
Source	54.95	42.61	39.79	37.73	9.61	58.46	95.21	99.83	99.99	4.16
S+OK	51.63	38.18	35.56	33.80	7.55	58.50	94.97	99.83	99.98	3.93
SwOK	56.47	43.95	40.07	37.19	6.16	57.16	93.67	99.60	99.97	3.77
OK	55.21	42.43	39.26	36.97	9.15	56.12	93.85	99.56	99.93	4.06
S+CK	47.66	32.11	29.95	28.75	6.23	**62.97**	**96.89**	**99.93**	99.98	**7.01**
SwCK	15.35	**1.10**	**0.18**	**0.04**	**0.00**	56.32	93.49	99.75	99.98	4.22
CK	53.37	39.21	37.08	35.82	6.20	62.13	95.94	99.88	**99.99**	4.05
Ref.	**15.32**	**1.10**	**0.18**	**0.04**	**0.00**	62.84	96.69	99.73	99.78	7.00

Table 6: Duplication and novelty rates on the test set of CNN/DailyMail dataset with 10,000 training iterations.

more than the reference on the Multi-News dataset and about 0.18% more than the reference on the CNN/DailyMail dataset, indicating the potential improvement space for leveraging closed knowledge in summarization.

Furthermore, the comparisons of novelty rates between the baseline pointer generator model and each of the best improvements incorporated with semantic knowledge on Multi-News and CNN/DailyMail datasets are shown in Figure 4. The implementation of an augmented pointer generator summarizer with closed knowledge presents its effectiveness for the summarization task on both the Multi-News dataset and the CNN/DailyMail dataset, since it generates about an average of 7.85% novel terms in the form of N-grams on the Multi-News dataset, and about an average of 1.78% novel terms in the form of N-grams on the CNN/DailyMail dataset. Notably, the average novelty rates from the best-improved implementation also about 7.46% exceed the reference on the Multi-News dataset and about 0.2% exceed the reference on the CNN/DailyMail dataset. In summary, the empirical results indicate that implementations leveraging closed semantic knowledge can also profit from a high performance on adding novel information into its generated summaries, improving the summary informativeness.

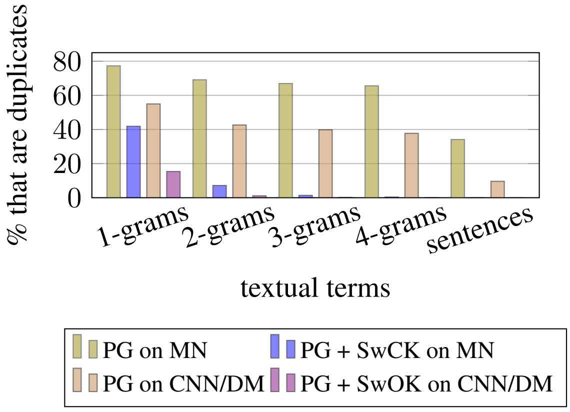

Figure 3: Comparisons of duplication rates between the baseline pointer generator (**PG**) model and each the best implementation leveraged knowledge on Multi-News (**MN**) and CNN/DailyMail (**CNN/DM**) datasets.

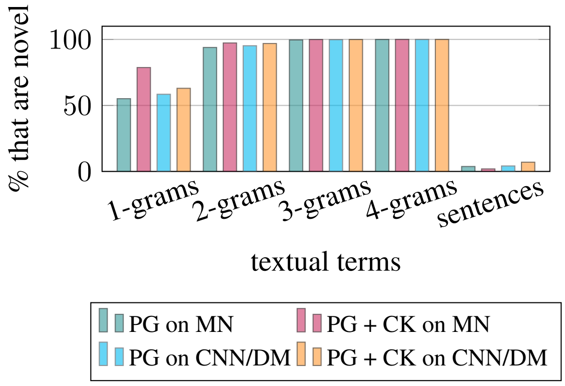

Figure 4: Comparisons of novelty rates between the baseline pointer generator (**PG**) model and each the best implementation leveraged knowledge on Multi-News (**MN**) and CNN/DailyMail (**CNN/DM**) datasets.

4 Conclusion

We demonstrate details of experiments on our improved pointer generator summarizer by incorporating kinds of semantic knowledge for the multi-document summarization task. Empirical results show that the pointer generator summarization model integrated knowledge achieved competitive results on ROUGE, BERTScore, duplication rate, and novelty rate evaluation metrics on both the Multi-News and CNN/DailyMail benchmark datasets compared to the baseline pointer generator model. In addition, a number of cases of leveraging kinds of semantic knowledge have experimented with corresponding detailed investigation results provided, which can provide a simple overview of the effectiveness of each type of semantic knowledge usage on the document summarization task. We hope that the experimented results and observations can assist in further investigations in leveraging knowledge and knowledge embeddings into document summarization tasks.

References

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alexander J. Smola, and Andrew McCallum. 2018. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

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 (NAACL-HLT)*, pages 4171–4186.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1074–1084.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence and 31st Innovative Applications of Artificial Intelligence Conference and 9th AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI-IAAI-EAAI)*, pages 6473–6480.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, pages 1693–1701.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5094–5107.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–736.

Xin Ji and Wen Zhao. 2021. SKGSUM: Abstractive document summarization with semantic knowledge graphs. In *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2284–2293.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. 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 2829–2839.

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

Yuanjie Lyu, Chen Zhu, Tong Xu, Zikai Yin, and Enhong Chen. 2022. Faithful abstractive summarization via fact-aware consistency-constrained transformer. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*, pages 1410–1419.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2020. Multi-document summarization via deep learning techniques: A survey. *arXiv*, abs/2011.04843.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1073–1083.

Tiancheng Tang, Tianyi Yuan, Xinhuai Tang, and Delai Chen. 2020. Incorporating external knowledge into unsupervised graph model for document summarization. *Electronics*, 9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pages 1–11.

Pancheng Wang, Shasha Li, Kunyuan Pang, Liangliang He, Dong Li, Jintao Tang, and Ting Wang. 2022. Multi-document scientific summarization from a knowledge graph-centric view. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING))*, pages 6222–6233.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *Proceedings of 8th International Conference on Learning Representations (ICLR)*.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 718–733.