Journal of Data Science 19 (3), 450–464
July 2021
DOI: 10.6339/21-JDS1012
Data Science in Action

Augmented Abstractive Summarization with Document-Level Semantic Graph

Qiwei Bi[1], Haoyuan Li[2], Kun Lu[3], and Hanfang Yang[1,4,*]

[1]*School of Statistics, Renmin University of China, Beijing, China*
[2]*T.H. Chan School of Public Health, Harvard University, Boston, MA, USA*
[3]*Department of ORFE, Princeton University, Princeton, New Jersey, USA*
[4]*Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, China*

Abstract

Previous abstractive methods apply sequence-to-sequence structures to generate summary without a module to assist the system to detect vital mentions and relationships within a document. To address this problem, we utilize semantic graph to boost the generation performance. Firstly, we extract important entities from each document and then establish a graph inspired by the idea of distant supervision (Mintz et al., 2009). Then, we combine a Bi-LSTM with a graph encoder to obtain the representation of each graph node. A novel neural decoder is presented to leverage the information of such entity graphs. Automatic and human evaluations show the effectiveness of our technique.

Keywords *distant supervise; entity extraction; graph attention neural network; information extraction*

1 Introduction

Text summarization systems aim to condense a piece of text into a shorter summary, while preserving the key information and the meaning of the content. After showing superior performance in machine translation (Bahdanau et al., 2014), the sequence-to-sequence models have also made encouraging progress in abstractive summarization (Rush et al., 2015; Nallapati et al., 2016).

However, simply adopting this structure in the summarization task is frequently found to bring undesirable behavior such as inaccurate factual details, repetitions, and lack of coherence (See et al., 2017; Paulus et al., 2018). This is mainly because the length of the input text for summarization is in general much longer than that of the machine translation task, while the generated summary is much shorter than the original document. Without good guidance, the model will have difficulty in choosing which aspects of content to focus on during generation. Fan et al. (2018) suggests that existing methods have difficulty in identifying salient entities and related events. In order to alleviate this problem, Sharma et al. (2019); Gehrmann et al. (2018) propose to follow a two-step schema by first selecting important phrases and then paraphrasing them. However, the corpus does not provide the related 'importance label' which leads to the loss of information in the first stage.

Structured data is helpful to tackle this issue. Abstract Meaning Representation (AMR), proposed by Banarescu et al. (2013), is a semantic graph representation of a sentence that can grasp the vital entities and relations. Damonte and Cohen (2019) show the benefits of

*Corresponding author. Email: hyang@ruc.edu.cn.

Received December 7, 2020; Accepted April 10, 2021

graph encoders in AMR-to-text generation, which aims to generate sentences from AMR graphs. Semantic graphs also prove to be very useful in generating longer texts expressing complex ideas. Logan et al. (2019) introduce a language model conditioned on an external knowledge graph relevant to the context. Koncel-Kedziorski et al. (2019) propose an end-to-end trainable system for graph-to-text generation. However, due to the lack of golden annotations, introducing graph-related methods to summarization is difficult.

Intuitively, provided with a cheat sheet about the essential entities and their relationships, it would be much easier to generate an accurate and coherent summary. Inspired by this, in this paper, we use information extraction methods to extract entities, coreferences, and relations for each document. To extract entities and coreferences, we simply adopt the industrial-strength tool Spacy (Honnibal et al., 2020). There are many state-of-the-art relation extraction models (Luan et al., 2018; Zhang et al., 2018; Trisedya et al., 2019), but these methods are all limited to a specific domain. If we want to apply these models in summarization, we have to collect related annotations and retrain them on new data, e.g. CNN/Daily mail (CNNDM) and New York Times (NYT). Therefore, we adopt a semi-supervised method to address this problem. Inspired by distant-supervision (Mintz et al., 2009), two types of relationships are considered in this paper: (1) two mentions from different coreference clusters are deemed to have relations if they appear in the same sentence; (2) two entities that participate in a known knowledge-base relation are regarded to express that relationship if they belong to the same sentence. To avoid redundant information, we also set a series of rules to only keep salient nodes in the graph. We believe a graph built this way can capture the global structure information of the original text. We then combine a Bi-LSTM and a Graph Attention Neural Network to encode the graph. Having this vital information in mind, we design a novel decoder to take advantage of information from both the original document and the graph.

Sharma et al. (2019) tries to use the information of entities when extracting salient sentences. However, this method is based on a two-step procedure and does not take the relation between entities into consideration. In order to mitigate the long-distance relationship problem, Fernandes et al. (2019) uses graphs to model highly-structured data. Coreferences of entities are connected with a special REF edge, but the relationships between different coreference clusters and entities are not considered. The model still has difficulty identifying relationships between entities that belong to different coreferences.

Our Contributions We propose a new architecture, which we call DSGSUM, to integrate graph information in the standard encoder-decoder framework for abstractive summarization. The contribution of our work is in three-folds: (i) To the best of our knowledge, we are the first to leverage the relationship between different coreferences or entities in summarization tasks, which improve the ability of our model to comprehend complex relationships between entities. (ii) We develop a new method to build graphs from text for the summarization task, which is proven to capture the global structural information better than previous methods according to our experiments. (iii) We design a novel decoder to incorporate the graphical information, achieving new state-of-the-art results in abstractive summarization, which can serve as a stepping stone for future research.

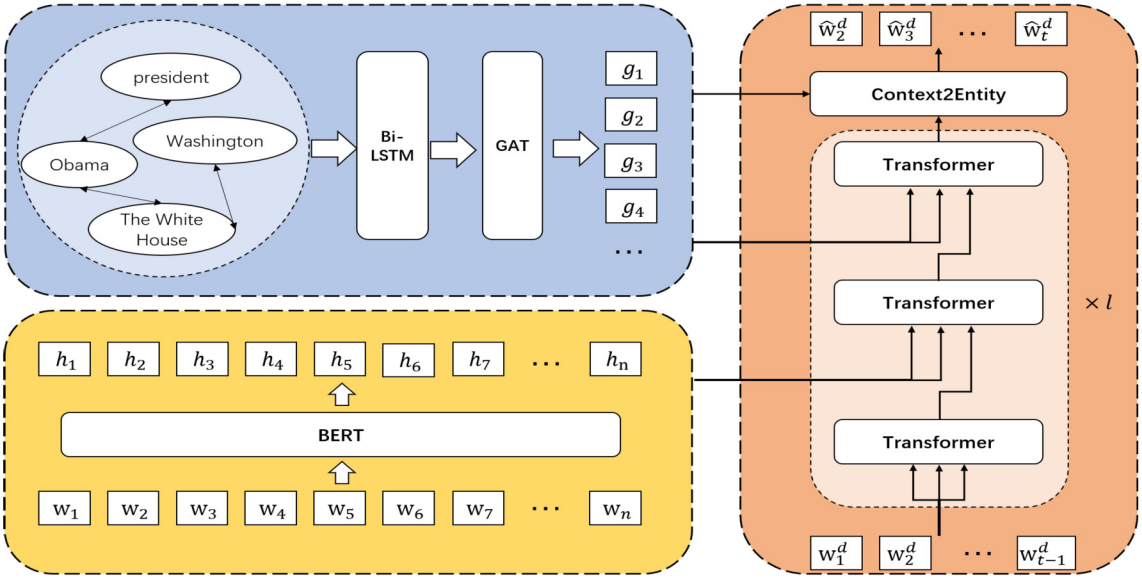

Figure 1: The framework of DSGSUM: (1) a graph encoder to obtain the contextualized representation of each graph node, (2) a BERT-based word representation encoder, and (3) a novel decoder structure which combines Transformer with Contxt2Entity attention to generate the final output $\widehat{\mathbf{w}}^d$, utilizing the structural information.

2 Model

In this section, we introduce our DSGSUM model that leverages graphic information in the abstractive summarization framework. Figure 1 shows three modules of our architecture: (1) a graph encoder to obtain the contextualized representation of each graph node, (2) a BERT-based word representation encoder similar to Liu and Lapata (2019), and (3) a novel decoder structure which incorporates the graph information.

Suppose we have a document D with n words. The document can be represented as $D = [x_1, x_2, \ldots, x_n]$, where x_i represents the i-th word in the document. Each word x_i can be embedded into a p-dimension vector $\mathbf{w}_i$, and we denote the embedding of input tokens at the decoding process as $\mathbf{w}^d$. Note that we insert [CLS] and [SEP] at the beginning and end of each sentence respectively. A chunk of consecutive words makes up an entity, which is a node in the graph structure. For example, the j-th entity consists of words from x_a to x_b can be denoted as $\tilde{e}_j = [x_a, \ldots, x_b]$. To obtain the representation $\mathbf{g}$ of each graph node, we firstly use a Bi-LSTM to encode each entity as $\mathbf{e}$. Then we use a graph encoder to take the relationship into consideration, which derives the final representation $\mathbf{g}$ of each node. To obtain the encoded representation $\mathbf{h}$ of each word in the document, we adopt a pre-trained BERT model. In the prediction process, we apply the normal Transformer decoding process first to obtain the hidden state $\mathbf{s}$. Then we adopt a context-to-entity attention inspired by Bi-Directional Attention Flow (BIDAF) (Seo et al., 2017) to leverage the structured information.

2.1 Graph Extraction

Before we introduce the model structure, it's necessary to clarify the procedure to generate the graph for each document. We use existing state-of-the-art tools to infer entities and extract the relation inspired by distant supervision.

Entity Extraction We use Spacy (Honnibal et al., 2020), a natural language processing tool with industrial-strength to process each document to extract entities $\tilde{e}_i$. As for coreference extraction, we use neuralcoref[1] developed by huggingface. The result is in the form of clusters $\tilde{c}_j$, consisting of two or more entities in a document referring to the same object (e.g. a person). For example, a cluster $\tilde{c}_j$ consists of two mentions $\tilde{e}_1$, $\tilde{e}_2$ can be represented as $\tilde{c}_j = [\tilde{e}_{j,1}, \tilde{e}_{j,2}]$, and the first entity is typically called the main mention of this cluster. Although these entities can provide our system with extra-linguistic information, trivial entities can discourage our model from concentrating on salient entities which are vital to detect crucial content. In order to avoid introducing redundant information, we only keep entities that are not stopwords. Moreover, a coreference cluster will be ignored if the total number of mentions is less than 3.

Relation Extraction There are quite a few high-quality relation extraction models, but most of them can only be applied to a specific domain with annotated relation labels. However, the summarization dataset lacks golden relation annotations, therefore it's difficult to build up a model to automatically extract relations. In this paper, we mainly consider two types of relationships:

(1) Distant supervision (Mintz et al., 2009) is proposed to address the shortage of golden relationship annotations. Its idea comes from the fact that any sentence containing a pair of entities that exist in a known knowledge-base relation are likely to have that type of information. Large scale graphical knowledge databases such as ConceptNet (Speer and Havasi, 2012) use graphical structure to express intricate relations between concepts. In our work, we use ConceptNet to decide whether there is a relation between two entities. However, the relation of triples extracted by this method is often redundant or trivial, so we limit the relationship types to 'UsedFor', 'CapableOf', 'Causes', 'CausesDesire', 'Desires', and 'ObstructedBy'.

(2) Also inspired by distant supervision, we propose that two main mentions of different clusters have relations if they are in the same sentence. We notice that the object that a cluster refers to is of great importance because it appears in the original text more than twice. Additionally, the main mention of a cluster with more quotations in the source document is more likely to appear in the golden summary (Sharma et al., 2019). Most of the previous papers simply label different mentions in one cluster as coreference. These methods neglect the relationship between different clusters which reflects the interaction of salient mentions. We propose that if two entities come from different coreference clusters and appear in the same sentence, then the two clusters can be deemed to have a relationship. For example, 'Barack Obama' and 'the White House,' appear many times in the training documents, although not always in the same form. Then it is reasonable to infer that 'Barack Obama' and 'the white house' have a strong relationship if they are quoted in the same sentence.

Combining the entity extraction and the relation extraction, we are able to build a graph representation for each document.

[1] https://github.com/huggingface/neuralcoref

2.2 Graph Encoder

Next, we use an LSTM (Hochreiter and Schmidhuber, 1997) to encode entities and apply a Graph Attention Network to capture relationship information of the graph we get from last step.

LSTM Encoder Entity span $\tilde{e}_i$ is often multiple words as shown in the Entity Extraction section. The initial vector representation $\mathbf{e}$ should be obtained before applying the graph structure model. The representations of the start word and the end word of a span typically contain most of these mention's information (Lee et al., 2017). However, our experiment shows that simply using the last word works better here. We run a Bi-LSTM over the initial embedding of each word in entity span $\tilde{e}_i$, and the concatenation of backward and forward outputs of the last word is defined as $\mathbf{e}_i$.

Graph Attention Network The structure of our graph encoder is similar to the Graph Attention Network (GAT) which enables efficient and parallel computation. The hidden representation of each node in a graph is obtained by attending over its neighbors based on the Transformer structure (Vaswani et al., 2017). In the first iteration, nodes in the GAT can only interact with their nearest neighbor nodes. We run the GAT more than twice to obtain more global representation for each node.

In the l-th iteration, the representation $\mathbf{g}_i^l$ of $\tilde{e}_i$ is contextualized by attending over its neighbor nodes. Before the model goes on with the next iteration, the derived vector is processed through a specially designed attention mechanism:

$$\mathbf{g}_i^l = \texttt{FFN}(\texttt{LayerNorm}(\mathbf{g}_i^{l-1} + \sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{g}_j^{l-1})), \tag{1}$$

$$\alpha_{ij} = \frac{\texttt{exp}((W_1 \mathbf{g}_i)(W_2 \mathbf{g}_j)^T)}{\sum_{k \in \mathcal{N}_i} \texttt{exp}((W_1 \mathbf{g}_i)(W_2 \mathbf{g}_k)^T)}, \tag{2}$$

where $\mathcal{N}_i$ denotes the neighborhood of $\tilde{e}_i$, including $\tilde{e}_i$ itself; W_1 and W_2 denote trainable matrices parameters throughout the paper. FFN is a two-layer feedforward network and LayerNorm is a layer normalization operation.

In this paper, we adopt an N-headed self-attention setup, with N independent attentions calculated and concatenated before a residual connection is applied. The final vector representation of i-th graph node is denoted as $\mathbf{g}_i$. Tay et al. (2020) propose that random alignment attention matrices surprisingly perform quite competitively. We are interesting to apply this trick in our method, but the results are not improved. Therefore, we simply adopt the vanilla transformer in the paper.

Frequency Embedding We notice that the entities that appear frequently in the original text are more likely to appear in the golden summarization, which means that these entities are more important in the process of generating a summary. In order to better detect these salient entities, we additionally introduce a frequency embedding for these derived entity representations, which is similar to the idea of positional embeddings in BERT. We sort entities in descending order according to the number of occurrences in the original text. Before being fed into the graph encoder, the representation of each entity is defined as $\mathbf{g}_i^0 := \mathbf{e}_i + \text{FE}(i)$, where $\text{FE}(\cdot)$ denotes the pre-initialized frequency embedding, and i represents the ranking of the entities after sorting.

2.3 Word Representation Encoder

Motivated by Liu and Lapata (2019), we use BERT to encode the document. Each token w_i is assigned three kinds of embeddings: a token embedding indicating the meaning of each token, a segmentation embedding discriminating between two sentences, and a position embedding denoting the position of each token. The encoder structure in this paper is similar to BERTSUMABS (Liu and Lapata, 2019). The encoded embedding of each word is denoted as $\mathbf{h} = \text{Transformer}(\mathbf{w}, \mathbf{w}, \mathbf{w})$. For simplicity, the first place denotes query vector, and the second and third place denote the key and value vector respectively.

2.4 Decoder

In a traditional seq-to-seq Transformer structure, the decoder obtains its contextualized representation $\mathbf{s}^d$ in a way similar to the encoder, with a modification to prevent the model from attending to subsequent positions. There is also a sub-layer to help the decoder focus on appropriate places in the input sequence. More details can be found in (Vaswani et al., 2017). Here we simply denote the output of this decoder as:

$$\mathbf{s}^d = \text{Transformer}(\mathbf{w}^d, \mathbf{w}^d, \mathbf{w}^d), \tag{3}$$

$$\mathbf{s} = \text{Transformer}(\mathbf{s}^d, \mathbf{h}, \mathbf{h}). \tag{4}$$

Context-to-Entity Attention The decoder state $\mathbf{s}_t$ is supposed to be entity-aware. In machine comprehension, Seo et al. (2017) introduce the Bi-Directional Attention Flow (BIDAF) network, a multi-stage hierarchical process that represents the context at different levels of granularity and uses a bidirectional attention flow mechanism to obtain a query-aware context representation. Inspired by this work, we propose a **context-to-entity** attention in this paper. In order to avoid redundant information our system, we try to introduce entity filter module before context2entity. However, the performance is below expectation, so we left our attention mechanism to decide which entity should be paid more attention.

In this module, we compute attentions from context to entities, which is derived from a similarity matrix, $\mathbf{M} := (M_{ti})_{1\leqslant t\leqslant T, 1\leqslant i\leqslant I} \in \mathbf{R}^{T\times I}$, between the contextual embeddings of the decoder word $\mathbf{s}$ and the entity $\mathbf{g}$, where T and I are size of $\mathbf{s}$ and $\mathbf{g}$, respectively. Each element M_{ti} indicates the similarity between the t-th decoder hidden vector and the i-th graph node, which is calculated by:

$$M_{ti} = \alpha\left(\mathbf{s}_t, \mathbf{g}_i\right), \tag{5}$$

where $\mathbf{s}_t$ is the t-th vector in decoder state, $\mathbf{g}_i$ is the embedding of the j-th entity. α is a function defined as $\alpha(\mathbf{s}, \mathbf{g}) = W_s\mathbf{s} + W_g\mathbf{g} + W_{sg}(\mathbf{s} \circ \mathbf{g})$, and $\circ$ denotes element-wise multiplication. Context-to-entity attention signifies which entities are most relevant to decode state $\mathbf{s}$ respectively. The attention weight is computed by, and the weighted entity vector is $\tilde{\mathbf{g}}_t = \sum_{i=1}^{I} a_{ti}\mathbf{g}_i$, which is a d-dimensional vector. The final graph-enhanced representation $\mathbf{s}^g$ can be defined as:

$$\mathbf{s}^g = W_G[\mathbf{s}; \tilde{\mathbf{g}}; \mathbf{s} \circ \tilde{\mathbf{g}}]. \tag{6}$$

With the decoder state embedding $\mathbf{s}^g$ at hand, we can simply predict the word distribution P_{vocab} as:

$$P_{vocab} = \text{softmax}(W_P\mathbf{s}^g), \tag{7}$$

where W_P projects the hidden state to the size of vocabulary. Here we want to mention that we also tried to add a entity-to-context attention flow to make a complete BIDAF, while the performance actually got worse. Experiments show that a single context-to-entity attention is good enough for our system. We also tried to introduce copy mechanism to copy salient entities to the generated summary. However, through ablation study, we found that it actually hurt the performance. This might be because the encoder based on powerful pre-trained model BERT has already learned part of the salient entities information, and copying redundant entities will add noise to the learning process. More details of this part can be found in the Appendix.

3 Experimental Setups

Datasets and Preprocessing We perform our experiments on the non-anonymous CNN/Dailymail (CNNDM) dataset[2] (Nallapati et al., 2016). We also adopt the New York Times dataset (Durrett et al., 2016), which licensed by LDC.[3] The source documents in the CNNDM training set on average have 766 words spanning 29.74 sentences while for summaries the numbers are 53 words and 3.72 sentences. We used the non-anonymized version of the CNNDM corpus and truncated source documents to 512 tokens. We followed the preprocessing steps in (Nallapati et al., 2016), obtaining 287,227 training pairs, 13,368 validation pairs and 11,490 test pairs.

For New York Times (NYT), we adopted the same split strategy as Durrett et al. (2016), and partitioned the whole dataset to 100,834 training pairs, 9,706 validation pairs, and 4000 test pairs. The test set contains all articles published since January 1, 2007. Then following Durrett et al. (2016), we created the NYT50 dataset by removing the pairs with summaries shorter than 50 words, and obtained a filtered test set with 3,452 examples. Input documents were truncated to 800 tokens, and all documents were tokenized with the Stanford CoreNLP toolkit (Manning et al., 2014).

Training Details and Parameters Both source and target texts were tokenized with BERT's subword tokenizer. We rarely observe issues with out-of-vocabulary words in the output thanks to the subword tokenizer. We used Spacy (Honnibal et al., 2020) to extract entities using model version 'en-core-web-sm'. As for coreference extraction, we used neuralcoref developed by huggingface.

After getting the graph, we first used a 2 layer Bi-LSTM with 384 hidden nodes to encode the entities, with the dropout rate to be 0.2. For the graph encoder, we adopted a 4-headed transformer implementation, where the number of graph layers is 3. We applied dropout with probability 0.2 before all linear layers. We adopted the transformer model implementation by (Wolf et al., 2019), and BERT-based models were used throughout the experiments. Our Transformer decoder has 768 hidden units and the hidden size for all feed-forward layers is 2,048. During decoding, we used beam search with size 5 and decoded until an end-of-sequence token is emitted. In order to reduce repetitions, repeated trigrams were blocked (Paulus et al., 2018). We used two Adam optimizers for the encoder and the decoder, respectively, each with the same warmup-steps and learning rate settings as Liu and Lapata (2019). The pre-trained encoder was fine-tuned with a smaller learning rate and smoother decay.

[2] https://cs.nyu.edu/~kcho/DMQA/
[3] https://catalog.ldc.upenn.edu/LDC2008T19

Table 1: ROUGE F1 results of various models on CNNDM test set using full-length F1 Rouge-1 (R1), Rouge-2 (R2), and Rouge-L (RL). For NYT50, we used limited-length ROUGE Recall (Liu and Lapata, 2019), where predicted summaries are truncated to the length of the gold summaries. We applied bootstrap resampling technique (Berg-Kirkpatrick et al., 2012), and our model performs significantly better than other models with $p < 0.05$.

Dataset	Model	R1	R2	RL
CNNDM	LEAD-3	40.42	17.62	36.67
	POINTGEN+COV (See et al., 2017)	39.53	17.28	36.38
	REWRITE (Chen and Bansal, 2018)	40.88	17.80	38.54
	DEEPREINFORCE (Paulus et al., 2018)	41.16	15.75	39.08
	BOTTOMUP (Gehrmann et al., 2018)	41.22	18.68	38.34
	SENECA (Sharma et al., 2019)	41.52	18.36	38.09
	DCA (Celikyilmaz et al., 2018)	41.69	**19.47**	37.92
	BERTSUMABS (Liu and Lapata, 2019)	41.72	19.39	38.76
	DSGSUM	**41.96**	19.29	**38.98**
NYT50	LEAD-3	39.58	20.11	35.78
	POINTGEN (See et al., 2017)	42.47	25.61	–
	POINTGEN+COV (See et al., 2017)	43.71	26.40	–
	DEEPREINFORCE (Paulus et al., 2018)	42.94	26.02	–
	BERTSUMABS (Liu and Lapata, 2019)	48.92	30.84	45.41
	DSGSUM	**49.86**	**31.04**	**46.64**

We chose the top-3 checkpoints based on their validation loss and reported the averaged results of these checkpoints on the test set. All of our models were trained for 150,000 steps on 2 GPUs (Tesla V100) with gradient accumulation every two steps, which roughly takes about a day and a half. We reported Rouge-1, Rouge-2, and Rouge-L as evaluation metrics of our models. Rouge-1 and Rouge-2 measure the unigram and bigram overlap respectively. Rouge-L is the longest common subsequence for assessing fluency (Durrett et al., 2016).

Baselines and Comparisons Besides baseline LEAD-3, we compared our model with existing popular state-of-the-art abstractive summarization models on CNNDM datasets: (1) pointer-generator model with coverage (POINTGEN+COV) (See et al., 2017); (2) sentence rewriting model (REWRITE) (Chen and Bansal, 2018); (3) RL-based abstractive summarization (DEEPREINFORCE) (Paulus et al., 2018); (4) bottom-up abstraction (BOTTOMUP) (Gehrmann et al., 2018); deep communicating agents-based summarization (DCA) (Celikyilmaz et al., 2018); a system for Entity-driven coherent abstractive summarization (SENECA) (Sharma et al., 2019), and an extended version of BERT for summarization (BERTSUMABS) (Liu and Lapata, 2019). For NYT50, we replicate the socore of POINTGEN and DEEPREINFORCE from (Sharma et al., 2019).

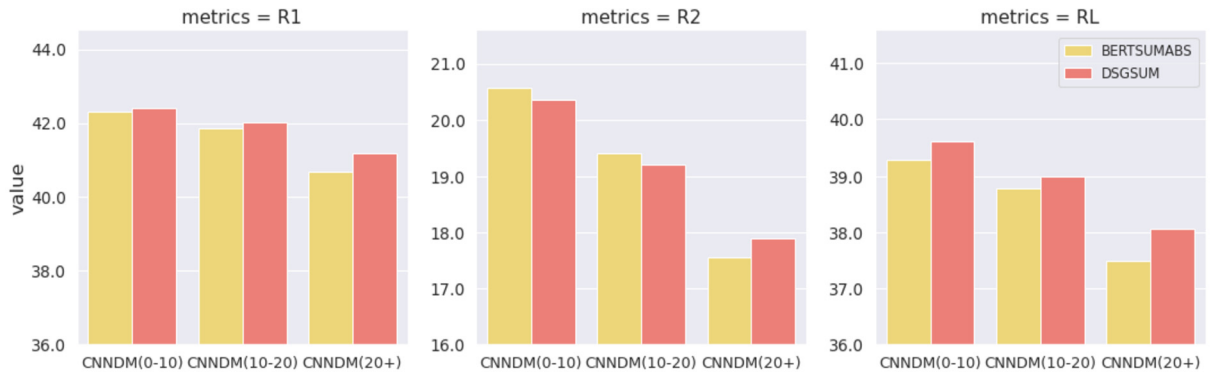

Figure 2: Results of our models and BERTSUMABS on CNNDM test set by the number of entities, using full-length F1 Rouge-1 (R-1), Rouge-2 (R-2), and Rouge-L (R-L). The split dataset has 3622, 5338, 2530 samples respectively. Our model is better than BERTSUMABS on all split datasets, and has the most significant improvement in the divided dataset containing more entities (e.g. CNNDM(20+)).

4 Results

4.1 Automatic Summary Evaluation

Table 1 summarizes the performances of all abstractive models on both CNN and NYT50 dataset. Our DSGSUM outperforms other methods by a large margin. Specifically, we can observe that DSGSUM beats the previous state-of-the-art model BERTSUMABS on R-1 and R-L. To verify the performance gap between DSGSUM and BERTSUMABS, we run another 4 training runs. We observe that the variance of the different models is non-overlapping, and all runs of DSGSUM outperformed BERTSUMABS on metrics of R-1 and R-L. It has been shown that the pretrained language model could capture a surprising amount of world knowledge. BERT may already have the information about the entities and the relations between them. However, the knowledge is stored implicitly in the parameters of a neural network (Guu et al., 2020). The result shows that the design of our model that extracts the salient entities and models their relations explicitly does bring extraction gains for the generation system. If the information is trivial, incorporation of the information will not bring significant improvement, and even spoil the model performance.

Under R-2 metric, though not significant, the score of our model is slightly lower than BERTSUMABS on CNNDM. This may partly due to the fact that our model focuses more on the logical relationships of the full text, causing the model to ignore some local grammars. As we will discuss in Section 4.4, BERTSUMABS also performs better on grammar on cnndm. Grammar often determines what kind of words should appear after the current word. There may be a positive correlation between R-2 score and grammatical performance. Consequently, BERTSUMABS performs better on R-2 score, which measures bi-gram overlap.

4.2 Ablation Study

The extracted mentions are of great importance, however the relationship between them can not be ignored. We believe that the writer would not mention two salient subjects in the same sentence if they do not have a strong correlation. For example, the coreference clusters ‘Barack Obama’ and ‘the White House’ are mentioned at least twice at different sentences and locations

Table 2: Results of ablation study on CNNDM. "- gat" denotes not to use graph attention network, in which the entities are only encoded by the lstm encoder; "- context2entity" denotes not to use the entity information while generating summary, and the results are almost the same as BERTSUMABS; "- FE" denotes not to use frequency embedding, in which the frequency embedding is not added into the representation of each entity.

Dataset	Model	R1	R2	RL
CNNDM	DSGSUM	**41.96**	**19.29**	**38.98**
	DSGSUM - gat	41.83	19.19	38.81
	DSGSUM - context2entity	41.74	19.15	38.79
	DSGSUM - FE	41.77	19.14	38.79

in the document, so the relationship between their main mentions can be deemed to have a long-range dependency, which is helpful to generate more coherent summaries (Sharma et al., 2019). To further highlight the value of using graphs rather than only salient entities, we propose an entity-version of our model by dropping the graph attention network. In this model, the embedding of each entity won't reflect the information of their relation. Both models make use of the same set of entities for the same test sample. The score on CNNDM is still better than BERTSUMABS. As shown in Table 2, due to the lack of information on relations, the performance is down to 41.83/19.19/38.81. We can conclude that related entities improve summary generation compared to an unrelated collection of entities. The result indicates that the introduction of the relationship between nodes is of great benefit for the model to detect salient information and transform this information into the surface form in the correct way. In addition, we conducted comparative experiments showing that context2entity and frequency embeddings are also effective.

4.3 The Value of Structured Information

We collect the entities in the summary written by humans. The generated summary of our model can cover about 62.01% of these entities on average, which is higher than that of BERTSUMABS (60.11%). Entities that appear in the golden summary are salient, often revealing the main topics of a sentence. The results show that our model can identify these important mentions and present them in the generated summary properly.

Why to Model Structured Data Explicitly We hold the idea that a system provided with structured information explicitly performs better. To verify this hypothesis, we split the test set according to the number of the extracted entities from the initial document and compare our model with BERTSUMABS. The entities we extract mainly include mentions from coreferences and entities recognized by Spacy. If a document contains many coreferences, it means that the logical relations in this article are more complicated. Most existing models are poor at comprehending such data sets and generating summaries. Provided explicitly with this information, it is expected that the benefit of our models will be more evident for those samples containing more entities. The result in Figure 2 shows that both our model and BERTSUMABS perform best on the divided dataset which has entities lower than 10. There are fewer entities in these kinds of documents, and models easily identify the main objects of these articles. When the entities extracted from the article exceed 20, both models perform the worst. But because our

Table 3: Sample golden summary, extracted entities and summaries generated by BERTSUMABS and DSGSUM. Different colored fonts indicate overlap with corresponding extracted entities. The results show that most extracted entities appear in the golden summary and also exist in the summary generated by DSGSUM. It verifies that entities are important for generating summaries and our model could properly use these entities while generating summaries.

Entity:
boston, the plane, delta, new jersey, newark, flight 271 from paris, logan international in boston, new york, king kong, another female passenger, an emergency landing in boston, one passenger, the turbulence

Gold:
delta said 'a small number of customers' on flight 271 from paris to newark complained of nausea and possible minor injuries. two people taken to massachusetts general hospital after unscheduled landing at logan airport in boston. extreme turbulence prevented the plane from landing at newark and then jfk before pilot headed to logan

BERTSUMABS:
delta flight from paris, france to newark, new jersey was diverted to boston. two people were taken to massachusetts general hospital with what are believed to be minor injuries. one passenger said it felt like 'king kong picked up the plane and shook it like there was no tomorrow'

DSGSUM:
delta said 'a small number of customers' on flight 271 from paris to newark airport in new jersey complained of nausea and possible minor injuries. two people were taken to massachusetts general hospital with what are believed to be minor injuries after the unscheduled landing at logan airport in boston

model explicitly utilizes structured information, our model performs significantly better than BERTSUMABS in this type of article.

4.4 Human Evaluation

Firstly, human evaluation is conducted to analyze the informativeness and readability of the summaries generated by our models. We conducted a manual evaluation with 3 proficient English speakers. We primarily used the following criteria to assess the generated summaries: **relevance**-whether the generated summary has strong relevance with the golden summary; **grammar**, **coherence**-whether the summary presents contents and entities in a coherent manner. To conduct the evaluation, we randomly selected 30 examples from the CNNDM testing set and asked the volunteers to subjectively rank the summaries against each other on a scale of 1 (worst) to 5 (best). For simplicity, we only provide the extracted entities to assist volunteers to judge, not the whole set of triples. Each example consisted of an article and three summaries, i.e., the golden summary, BERTSUMABS, and our proposed DSGSUM. Obviously, the summaries were randomly shuffled, and the model used to produce each summary was unknown to prevent bias. The results are presented in Table 4. Our model performs well on relevance and coherence. Surprisingly, BERTSUMABS ranks higher on GRAMMAR when compared to DSGSUM. Through manual inspection, we find that the most common cause of grammatical errors is the lack of adjacent components of words in a sentence, which can also lead to the lower bi-gram overlap compared with BERTSUMABS. A sample of case study is shown in Table 3.

Table 4: Results of human evaluation on CNNDM. Best results are in bold.

	RELEVANCE	GRAMMAR	COHERENCE
HUMAN	4.83	4.90	4.83
BERTSUMABS	4.33	**4.86**	4.07
DSGSUM	**4.57**	4.63	**4.43**

5 Conclusion

In this paper, we propose a new architecture, which we call DSGSUM, to integrate structured information for abstractive summarization. We develop a new method to build graphs for the summarization task and design a novel decoder to leverage the graphical information. Our model performs better than previous state-of-the-art models on the CNNDM and NYT50 datasets. We have shown that structured information is advantageous for abstractive summarization models. In this paper, our methods to extract triples also introduce noises and we have formulated some rules to filter triples. With more advanced methods to filter those noises, our model may perform better. And copy mechanism may also be helpful for our model, which directly copies important entities from the article into the abstract. We leave these parts to future work. More details can be found in the appendix.

Supplementary Material

Supplementary material online include: data link, python code and an instruction file needed to reproduce the results; an appendix containing additional structures and experiments we have tried. The web link is https://github.com/martin6336/DSGSum.

References

Bahdanau D, Cho K, Bengio Y (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint: https://arxiv.org/abs/1409.0473.

Banarescu L, Bonial C, Cai S, Georgescu M, Griffitt K, Hermjakob U, et al. (2013). Abstract Meaning Representation for sembanking. In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (S Dipper, M Liakata, A Pareja-Lora, eds.), 178–186. Association for Computational Linguistics, Sofia, Bulgaria.

Berg-Kirkpatrick T, Burkett D, Klein D (2012). An empirical investigation of statistical significance in NLP. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (M Pasca, J Henderson, J Tsujii, eds.), 995–1005. Association for Computational Linguistics, Jeju Island, Korea.

Celikyilmaz A, Bosselut A, He X, Choi Y (2018). Deep communicating agents for abstractive summarization. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (A Stent, H Ji, MA Walker, eds.), 1662–1675. Association for Computational Linguistics, New Orleans, Louisiana.

Chen YC, Bansal M (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)* (Y Miyao, I Gurevych, eds.), 675–686. Association for Computational Linguistics, Melbourne, Australia.

Damonte M, Cohen SB (2019). Structural neural encoders for AMR-to-text generation. 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)* (T Solorio, C Doran, J Burstein, eds.), 3649–3658. Association for Computational Linguistics, Minneapolis, Minnesota.

Durrett G, Berg-Kirkpatrick T, Klein D (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)*, 1998–2008. Association for Computational Linguistics, Berlin, Germany.

Fan L, Yu D, Wang L (2018). Robust neural abstractive summarization systems and evaluation against adversarial information. arXiv preprint: https://arxiv.org/abs/1810.06065

Fernandes P, Allamanis M, Brockschmidt M (2019). Structured neural summarization.

Gehrmann S, Deng Y, Rush A (2018). Bottom-up abstractive summarization. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (J Tsujii, J Hockenmaier, D Chiang, E Riloff, eds.), 4098–4109. Association for Computational Linguistics, Brussels, Belgium.

Guu K, Lee K, Tung Z, Pasupat P, Chang MW (2020). Realm: Retrieval-augmented language model pre-training. arXiv preprint: https://arxiv.org/abs/2002.08909

Hochreiter S, Schmidhuber J (1997). Long short-term memory. *Neural Comput.*, 9(8): 1735–1780.

Honnibal M, Montani I, Van Landeghem S, Boyd A (2020). spaCy: Industrial-strength Natural Language Processing in Python. Zenodo, https://doi.org/10.5281/zenodo.1212303

Koncel-Kedziorski R, Bekal D, Luan Y, Lapata M, Hajishirzi H (2019). Text generation from knowledge graphs with graph transformers. arXiv preprint: https://arxiv.org/abs/1904.02342

Lee K, He L, Lewis M, Zettlemoyer L (2017). End-to-end neural coreference resolution. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (M Palmer, R Hwa, S Riedel, eds.), 188–197. Association for Computational Linguistics, Copenhagen, Denmark.

Liu Y, Lapata M (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)* (K Inui, J Jiang, V Ng, X Wan, eds.), 3730–3740. Association for Computational Linguistics, Hong Kong, China.

Logan R, Liu NF, Peters ME, Gardner M, Singh S (2019). Barack's wife Hillary: Using knowledge graphs for fact-aware language modeling. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A Korhonen, DR Traum, L Màrquez, eds.), 5962–5971. Association for Computational Linguistics, Florence, Italy.

Luan Y, He L, Ostendorf M, Hajishirzi H (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (E Riloff, D Chiang, J Hockenmaier, J Tsujii, eds.), 3219–3232. Association for Computational Linguistics, Brussels, Belgium.

Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014). The Stanford CoreNLP natural language processing toolkit. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. Association for Computational Linguistics, Baltimore, Maryland.

Mintz M, Bills S, Snow R, Jurafsky D (2009). Distant supervision for relation extraction without labeled data. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (J Wiebe, J Su, K-Y Su, eds.), 1003–1011. Association for Computational Linguistics, Suntec, Singapore.

Nallapati R, Zhou B, dos Santos C, Gulçehre Ç, Xiang B (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning* (S Riezler, Y Goldberg, eds.), 280–290. Association for Computational Linguistics, Berlin, Germany.

Paulus R, Xiong C, Socher R (2018). A deep reinforced model for abstractive summarization. In: *International Conference on Learning Representations*.

Rush AM, Chopra S, Weston J (2015). A neural attention model for abstractive sentence summarization. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (L Màrquez, C Callison-Burch, J Su, D Pighin, Y Marton, eds.), 379–389. Association for Computational Linguistics, Lisbon, Portugal.

See A, Liu PJ, Manning CD (2017). Get to the point: Summarization with pointer-generator networks. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (R Barzilay, M-Y Kan, eds.), 1073–1083. Association for Computational Linguistics, Vancouver, Canada.

Seo M, Kembhavi A, Farhadi A, Hajishirzi H (2017). Bidirectional attention flow for machine comprehension. arXiv preprint: https://arxiv.org/abs/1611.01603

Sharma E, Huang L, Hu Z, Wang L (2019). An entity-driven framework for abstractive summarization. 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)* (K Inui, J Jiang, V Ng, X Wan, eds.), 3280–3291. Association for Computational Linguistics, Hong Kong, China.

Speer R, Havasi C (2012). Representing general relational knowledge in ConceptNet 5. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (N Calzolari, K Choukri, Declerck T, MU Dogan, B Maegaard, J Mariani, J Odijk, S Piperidis,eds.), 3679–3686. European Language Resources Association (ELRA), Istanbul, Turkey.

Tay Y, Bahri D, Metzler D, Juan DC, Zhao Z, Zheng C (2020). Synthesizer: Rethinking self-attention in transformer models. arXiv preprint: https://arxiv.org/abs/2005.00743

Trisedya BD, Weikum G, Qi J, Zhang R (2019). Neural relation extraction for knowledge base enrichment. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A Korhonen, DR Traum, L Màrquez, eds.), 229–240. Association for Computational Linguistics, Florence, Italy.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. (2017). Attention is all you need. In: *Advances in Neural Information Processing Systems* (I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett, eds.), volume 30. Curran Associates, Inc.

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint: https://arxiv.org/abs/1910.03771

Zhang Y, Qi P, Manning CD (2018). Graph convolution over pruned dependency trees improves relation extraction. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (E Riloff, D Chiang, J Hockenmaier, J Tsujii, eds.), 2205–2215. Association for Computational Linguistics, Brussels, Belgium.