Multi-Granularity Interaction Network for Extractive and Abstractive Multi-Document Summarization

Hanqi Jin, Tianming Wang, Xiaojun Wan
Wangxuan Institute of Computer Technology, Peking University
Center for Data Science, Peking University
The MOE Key Laboratory of Computational Linguistics, Peking University
{jinhanqi,wangtm,wanxiaojun}@pku.edu.cn

Abstract

In this paper, we propose a multi-granularity interaction network for extractive and abstractive multi-document summarization, which jointly learn semantic representations for words, sentences, and documents. The word representations are used to generate an abstractive summary while the sentence representations are used to produce an extractive summary. We employ attention mechanisms to interact between different granularity of semantic representations, which helps to capture multi-granularity key information and improves the performance of both abstractive and extractive summarization. Experiment results show that our proposed model substantially outperforms all strong baseline methods and achieves the best results on the Multi-News dataset.

1 Introduction

Document summarization aims at producing a fluent, condensed summary for given documents. Single document summarization has shown promising results with sequence-to-sequence models that encode a source document and then decode it into a summary (See et al., 2017; Paulus et al., 2018; Gehrmann et al., 2018; Çelikyilmaz et al., 2018). Multi-document summarization requires producing a summary from a cluster of thematically related documents, where the given documents complement and overlap each other. Multi-document summarization involves identifying important information and filtering out redundant information from multiple input sources.

There are two primary methodologies for multi-document summarization: *extractive* and *abstractive*. Extractive methods directly select important sentences from the original, which are relatively simple. Cao et al. (2015) rank sentences with a recursive neural network. Yasunaga et al. (2017) employ a Graph Convolutional Network (GCN) to incorporate sentence relation graphs to improve the performance for the extractive summarization. Abstractive methods can generate new words and new sentences, but it is technically more difficult than extractive methods. Some works on multi-document summarization simply concatenate multiple source documents into a long flat sequence and model multi-document summarization as a long sequence-to-sequence task (Liu et al., 2018; Fabbri et al., 2019). However, these approaches don't take the hierarchical structure of document clusters into account, while the too-long input often leads to the degradation in document summarization (Cohan et al., 2018; Liu and Lapata, 2019). Recently, hierarchical frameworks have shown their effectiveness on multi-document summarization (Zhang et al., 2018; Liu and Lapata, 2019). These approaches usually use multiple encoders to model hierarchical relationships in the discourse structure, but other methods to incorporate the structural semantic knowledge have not been explored. The combination of extractive and abstractive has been explored in single document summarization. Chen and Bansal (2018) use the extracted sentences as the input of the abstractive summarization. Subramanian et al. (2019) concatenate the extracted summary to the original document as the input of the abstractive summarization.

In this work, we treat documents, sentences, and words as the different granularity of semantic units, and connect these semantic units within a three-granularity hierarchical relation graph. With the multi-granularity hierarchical structure, we can unify extractive and abstractive summarization into one architecture simultaneously. Extractive summarization operates on sentence-granularity and directly supervises the sentence representations while abstractive summarization operates on word-granularity and directly supervises the word repre-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6244–6254

sentations. We propose a novel multi-granularity interaction network to enable the supervisions to promote the learning of all granularity representations. We employ the attention mechanism to encode the relationships between the same semantic granularity and hierarchical relationships between the different semantic granularity, respectively. And we use a fusion gate to integrate the various relationships for updating the semantic representations. The decoding part consists of a sentence extractor and a summary generator. The sentence extractor utilizes the sentence representations to select sentences, while the summary generator utilizes the word representations to generate a summary. The two tasks are trained in a unified architecture to promote the recognition of important information simultaneously.

We evaluate our model on the recently released Multi-News dataset and our proposed architecture brings substantial improvements over several strong baselines. We explore the influence of semantic units with different granularity, and the ablation study shows that joint learning of extractive and abstractive summarization in a unified architecture improves the performance.

In summary, we make the following contributions in this paper:

- We establish multi-granularity semantic representations for documents, sentences, and words, and propose a novel multi-granularity interaction network to encode multiple input documents.
- Our approach can unify the extractive and abstractive summarization into one architecture with interactive semantic units and promote the recognition of important information in different granularities.
- Experimental results on the Multi-News dataset show that our approach substantially outperforms several strong baselines and achieves state-of-the-art performance. Our code is publicly available at https://github.com/zhongxia96/MGSum.

2 Related Work

The methods for multi-document summarization can generally be categorized to extractive and abstractive. The extractive methods produce a summary by extracting and merging sentences from the input documents, while the abstractive methods generate a summary using arbitrary words and expressions based on the understanding of the documents. Due to the lack of available training data, most previous multi-document summarization methods were extractive (Erkan and Radev, 2004; Christensen et al., 2013; Yasunaga et al., 2017).

Since the neural abstractive models have achieved promising results on single-document summarization (See et al., 2017; Paulus et al., 2018; Gehrmann et al., 2018; Çelikyilmaz et al., 2018), some works trained abstractive summarization models on a single document dataset and adjusted the model to adapt the multi-document summarization task. Zhang et al. (2018) added a document set encoder into the single document summarization framework and tuned the pre-trained model on the multi-document summarization dataset. Lebanoff et al. (2018) combined an extractive summarization algorithm (MMR) for sentence extraction to reweight the original sentence importance distribution learned in the single document abstractive summarization model. Recently, two large scale multi-document summarization datasets have been proposed, one for very long input, aimed at generating Wikipedia (Liu et al., 2018) and another dedicated to generating a comprehensive summarization of multiple real-time news (Fabbri et al., 2019). Liu et al. (2018) concatenated multiple source documents into a long flat text and introduced a decoder-only architecture that can scalably attend to very long sequences, much longer than typical encoder-decoder architectures. Liu and Lapata (2019) introduced intermediate document representations and simply add the document representations to word representations for modeling the cross-document relationships. Compared with our proposed multi-granularity method, Liu and Lapata (2019) inclined to the traditional bottom-up hierarchical method and don't effectively utilize the hierarchical representations while ignoring the hierarchical relationships of sentences. Fabbri et al. (2019) incorporated MMR into a hierarchical pointer-generator network to address the information redundancy in multi-document summarization.

3 Our Approach

Our model consists of a multi-granularity encoder, a sentence extractor, and a summary generator. Firstly, the multi-granularity encoder reads multiple input documents and learns the multi-

granularity representations for words, sentences, and documents. Self-attention mechanisms are employed for capturing semantic relationships of the representations with same granularity, while cross-attention mechanisms are employed for the information interaction between representations with different granularity. Fusion gates are used for integrating the information from different attention mechanisms. Then the sentence extractor scores sentences according to the learned sentence representations. Meanwhile, the summary generator produces the abstractive summary by attending to the word representations. In the following sections, we will describe the multi-granularity encoder, the sentence extractor, and the summary generator, respectively.

3.1 Multi-Granularity Encoder

Given a cluster of documents, we establish explicit representations for documents, sentences, and words, and connect them within a hierarchical semantic relation graph. The multi-granularity encoder is a stack of L_1 identical layers. Each layer has two sub-layers: the first is the multi-granularity attention layer, and the second is multiple fully connected feed-forward networks. The multi-granularity attention sub-layer transfers semantic information between the different granularity and the same granularity, while the feed-forward network further aggregates the multi-granularity information. We employ multi-head attention to encode multi-granularity information and use a fusion gate to propagate semantic information to each other. Figure 1 shows the overview of the multi-granularity encoder layer, and Figure 2 illustrates how the semantic representations are updated, which takes the sentence representation as an example.

Let $w_{i,j,k}$ be the k-th word of the sentence $s_{i,j}$ in the document d_i. At the bottom of the encoder stack, each input word $w_{i,j,k}$ is converted into the vector representation $e_{i,j,k}$ by learned embeddings. We assign positional encoding to indicate the position of the word $w_{i,j,k}$ and three positions need to be considered, namely i (the rank of the document), j (the position of the sentence within the document), k (the position of the word within the sentence). We concatenate the three position embedding PE_i, PE_j,and PE_k to get the final position embedding $p_{i,j,k}$. The input word representation can be obtained by simply adding the word

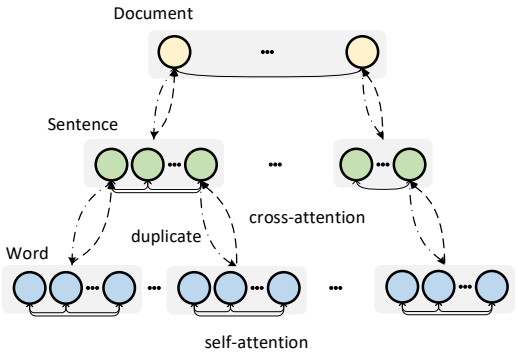

Figure 1: The overview of the multi-granularity encoder layer.

embedding $e_{i,j,k}$ and the position embedding $p_{i,j,k}$:

$$\begin{aligned} p_{i,j,k} &= [PE_i; PE_j; PE_k] \\ h^0_{w_{i,j,k}} &= e_{i,j,k} + p_{i,j,k} \end{aligned} \tag{1}$$

where the definition of positional encoding PE is consistent with the Transfomer (Vaswani et al., 2017). For convenience, we denote the output of l-th multi-granularity encoder layer as h^l and the input for the first layer as h^0. Symbols with subscripts $w_{i,j,k}$, $s_{i,j}$ and d_i are used to denote word, sentence, and document granularities, respectively. Both sentence representations $h^0_{s_{i,j}}$ and document representations $h^0_{d_i}$ are initialized to zeros.

In each multi-granularity attention sub-layers, the word representation is updated by the information of word granularity and sentence granularity. We perform multi-head self-attention across the word representations in the same sentence $h^{l-1}_{w_{i,j,*}} = \{h^{l-1}_{w_{i,j,k}} | w_{i,j,k} \in s_{i,j}\}$ to get the context representation $\tilde{h}^l_{w_{i,j,k}}$. In order to propagate semantic information from sentence granularity to the word granularity, we duplicate sentence-aware representation $\overleftarrow{h}^l_{w_{i,j,k}}$ from corresponding sentence $s_{i,j}$ and employ a fusion gate to integrate $\tilde{h}^l_{w_{i,j,k}}$ and $\overleftarrow{h}^l_{w_{i,j,k}}$ to get the updated word representation $f^l_{w_{i,j,k}}$.

$$\begin{aligned} f^l_{w_{i,j,k}} &= \text{Fusion}\left(\tilde{h}^l_{w_{i,j,k}}, \overleftarrow{h}^l_{w_{i,j,k}}\right) \\ \tilde{h}^l_{w_{i,j,k}} &= \text{MHAtt}\left(h^{l-1}_{w_{i,j,k}}, h^{l-1}_{w_{i,j,*}}\right) \\ \overleftarrow{h}^l_{w_{i,j,k}} &= h^{l-1}_{s_{i,j}} \end{aligned} \tag{2}$$

where MHAtt denotes the multi-head attention proposed in Vaswani et al. (2017) and Fusion denotes the fusion gate. $h^{l-1}_{w_{i,j,k}}$ is the query and $h^{l-1}_{w_{i,j,*}}$ are

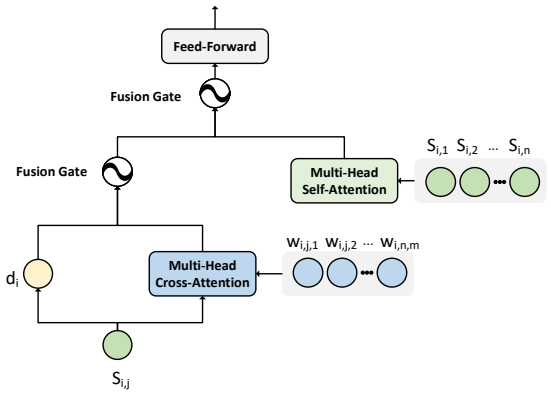

Figure 2: The multi-granularity encoder layer for updating sentence representation. The sentence representation is updated by using two fusion gates to integrate the information from different granularities.

the keys and values for attention. The fusion gate works as

$$\begin{aligned} \mathrm{z} &= \sigma\left([x;y]W_f + b_f\right) \\ \mathrm{Fusion}(x,y) &= \mathrm{z}\,x + (1-\mathrm{z})\,y \end{aligned} \tag{3}$$

where σ is the sigmoid function, parameters $W_f \in \mathbb{R}^{2*d_{model}\times 1}$ and $b_f \in \mathbb{R}$.

The sentence representation is updated from three sources: (1) We take the sentence representation $h^{l-1}_{s_{i,j}}$ as the query, the word representations $h^{l-1}_{w_{i,j,*}} = \{h^{l-1}_{w_{i,j,k}} | w_{i,j,k} \in s_{i,j}\}$ as the keys and values, to perform multi-head cross-attention to get the intermediate word-aware representation $\overrightarrow{h}^{l-1}_{s_{i,j}}$; (2) Multi-head self-attention across sentence representations $h^{l-1}_{s_{i,*}} = \{h^{l-1}_{s_{i,j}} | s_{i,j} \in d_i\}$ is performed to get the context representation $\tilde{h}^{l}_{s_{i,j}}$; (3) In order to propagate document granularity semantic information to the sentence, we duplicate the document-aware representation $\overleftarrow{h}^{l}_{s_{i,j}}$ from corresponding document d_i.

$$\begin{aligned} \overrightarrow{h}^{l}_{s_{i,j}} &= \mathrm{MHAtt}\left(h^{l-1}_{s_{i,j}}, h^{l-1}_{w_{i,j,*}}\right) \\ \tilde{h}^{l}_{s_{i,j}} &= \mathrm{MHAtt}\left(h^{l-1}_{s_{i,j}}, h^{l-1}_{s_{i,*}}\right) \\ \overleftarrow{h}^{l}_{s_{i,j}} &= h^{l-1}_{d_i} \end{aligned} \tag{4}$$

Semantic representations from the three sources are fused by two fusion gate to get the updated sentence representation $f^{l}_{s_{i,j}}$.

$$f^{l}_{s_{i,j}} = \mathrm{Fusion}\left(\mathrm{Fusion}\left(\overrightarrow{h}^{l}_{s_{i,j}}, \overleftarrow{h}^{l}_{s_{i,j}}\right), \tilde{h}^{l}_{s_{i,j}}\right) \tag{5}$$

To update the document representation, multi-head self-attention across all document representations $h^{l-1}_{d_*} = \{h^{l-1}_{d_i}\}$ is performed to get the context representation $\tilde{h}^{l}_{d_i}$. Meanwhile, we take the document representation $h^{l-1}_{d_i}$ as the query, sentence representations $\{h^{l-1}_{s_{i,j}} | s_{i,j} \in d_i\}$ as the keys and values to perform multi-head cross-attention to get the intermediate sentence-aware representation $\overrightarrow{h}^{l}_{d_i}$. A fusion gate is used to aggregate the above outputs $\tilde{h}^{l}_{d_i}$ and $\overrightarrow{h}^{l}_{d_i}$.

$$\begin{aligned} f^{l}_{d_i} &= \mathrm{Fusion}\left(\tilde{h}^{l}_{d_i}, \overrightarrow{h}^{l}_{d_i}\right) \\ \tilde{h}^{l}_{d_i} &= \mathrm{MHAtt}\left(h^{l-1}_{d_i}, h^{l-1}_{d_*}\right) \\ \overrightarrow{h}^{l}_{d_i} &= \mathrm{MHAtt}\left(h^{l-1}_{d_i}, h^{l-1}_{s_{i,*}}\right) \end{aligned} \tag{6}$$

The feed-forward network FFN is used to transform multiple-granularity semantic information further. To construct deep network, we use the residual connection (He et al., 2016) and layer normalization (Ba et al., 2016) to connect adjacent layers.

$$\begin{aligned} \tilde{h} &= \mathrm{LayerNorm}\left(h^{l-1} + f^{l}\right) \\ h^{l} &= \mathrm{LayerNorm}(\tilde{h} + \mathrm{FFN}(\tilde{h})) \end{aligned} \tag{7}$$

where $l \in [1, L_1]$, FFN consists of two linear transformations with a ReLU activation in between. Note that we used different FFN and LayerNorm for the different granularity. The final representation $h^{L_1}_s$ is fed to the sentence extractor while $h^{L_1}_w$ is fed to the summary generator. For convenience, we denote $h^{L_1}_s$ as o_s, and $h^{L_1}_w$ as o_w.

3.2 Sentence Extractor

we build a classifier to select sentences based on the sentence representations o_s from the multi-granularity encoder. The classifier uses a linear transformation layer with the sigmoid activation function to get the prediction score for each sentence

$$\tilde{y}_s = \sigma\left(o_s W_o + b_o\right) \tag{8}$$

where σ is the sigmoid function, parameters $W_o \in \mathbb{R}^{d_{model}\times 1}$ and $b_o \in \mathbb{R}$.

These scores are used to sort the sentences of multiple documents and produce the extracted summary.

3.3 Summary Generator

The summary generator in our model is also a stack of L_2 identical layers. The layer consists of three parts: a masked multi-head self-attention mechanism, a multi-head cross-attention mechanism, and

a fully connected feed-forward network. As the input and output of multi-document summarization are generally long, the multi-head attention degenerates as the length increases (Liu and Lapata, 2019). Following Zhao et al. (2019) 's idea, we adopt a sparse attention mechanism where each query only attends to the top-k values according to their weights calculated by the keys rather than all values in the original attention (Vaswani et al., 2017). And k is a hyper-parameter. This ensures that the generator focuses on critical information in the input and ignores much irrelevant information. We denote the multi-head sparse attention as MSAttn.

Similar to the multi-granularity encoder, we add the positional encoding of words in the summary to the input embedding at the bottom of the decoder stack. We denote the output of the l-th layer as g^l and the input for the first layer as g^0. The self-attention sub-layer with masking mechanism is used to encode the decoded information. The masking mechanism ensures that the prediction of the position t depends only on the known output of the position before t.

$$\tilde{g} = \text{LayerNorm}(g^{l-1} + \text{MSAttn}\left(g^{l-1}, g^{l-1}\right)) \tag{9}$$

The cross-attention sub-layer take the self-attention output $\tilde{g}$ as the queries and the multi-granularity encoder output o_w as keys and values to performs multi-head sparse attention. The feed-forward network is used to further transform the outputs.

$$\begin{aligned} c &= \text{LayerNorm}\ \left(\tilde{g} + \text{MSAtt}\left(\tilde{g}, o_w\right)\right) \\ g^l &= \text{LayerNorm}\ \left(c + \text{FFN}(c)\right) \end{aligned} \tag{10}$$

The generation distribution p_t^g over the target vocabulary is calculated by feeding the output $g_t^{L_2}$ to a softmax layer.

$$p_t^g = \text{softmax}\left(g_t^{L_2} W_g + b_g\right) \tag{11}$$

where $W_g \in \mathbb{R}^{d_{model} \times d_{vocab}}, b_g \in \mathbb{R}^{d_{vocab}}$ and d_{vocab} is the size of target vocabulary.

The copy mechanism (Gu et al., 2016) is employed to tackle the problem of out-of-vocabulary (OOV) words. We compute the copy attention ε_t with the decoder output g^{L_2} and the input representations o_w, and further obtain copy distribution p_t^c.

$$\begin{aligned} \varepsilon_t &= \text{softmax}(g_t^{L_2} o_w^\top + b_\varepsilon) \\ p_t^c &= \sum_{i,j,k} \varepsilon_t z_{i,j,k}^\top \end{aligned} \tag{12}$$

where $z_{i,j,k}$ is the one-hot indicator vector for $w_{i,j,k}$ and $b_\varepsilon \in \mathbb{R}^{d_{vocab}}$.

A gate is used over the the decoder output g^{L_2} to control generating words from the vocabulary or copying words directly from the source text. The final distribution p_t is the "mixture" of the two distributions p_t^g and p_t^c.

$$\begin{aligned} \eta_t &= \sigma\left(g_t^{L_2} W_\eta + b_\eta\right) \\ p_t &= \eta_t * p_t^g + (1 - \eta_t) * p_t^c \end{aligned} \tag{13}$$

where σ is the sigmoid function, $W_\eta \in \mathbb{R}^{d_{model} \times 1}, b_\eta \in \mathbb{R}$.

3.4 Objective Function

We train the sentence extractor and the summary generator in a unified architecture in an end-to-end manner. We use the cross entropy as both the extractor loss and the generator loss.

$$\begin{aligned} L_{ext} = -\frac{1}{N} \sum_{n=1}^{N} \Big(& y_s^{(n)} \log \tilde{y}_s^{(n)} + \\ & \left(1 - y_s^{(n)}\right) \log\left(1 - \tilde{y}_s^{(n)}\right)\Big) \\ L_{abs} = -\frac{1}{N} \sum_{n=1}^{N} & \log p(y_w^{(n)}) \end{aligned} \tag{14}$$

where y_s is the ground-truth extracted label, y_w is the ground-truth summary and N is the number of samples in the corpus.

The final loss is as below

$$L_{mix} = L_{abs} + \lambda L_{ext} \tag{15}$$

where λ is a hyper-parameter.

4 Experiment

4.1 Dataset

We experiment with the latest released Multi-News dataset (Fabbri et al., 2019), which is the first large scale multi-document news summarization dataset. It contains about 44972 pairs for training, 5622 pairs for development, and 5622 for the test. Each summary of the average of 264 words is paired with a documents cluster of average 2103 words discussing a topic. The number of source documents per summary presents as shown in Table 1 . While the dataset contains abstractive gold summaries, it is not readily suited to training extractive models. So we follow the work of Zhou et al. (2018) on extractive summary labeling, constructing gold-label sequences by greedily optimizing ROUGE-2 F1 on the gold-standard summary.

# of source	Frequency	# of source	Frequency
2	23,894	7	382
3	12,707	8	209
4	5,022	9	89
5	1,873	10	33
6	763		

Table 1: The distribution of number of source articles per instance in Multi-News dataset.

4.2 Implementation Details

We set our model parameters based on preliminary experiments on the development set. We prune the vocabulary to 50k and use the word in the source documents with maximum weight in copy attention to replace the unknown word of the generated summary. We set the dimension of word embeddings and hidden units d_{model} to 512, feed-forward units to 2048. We set 8 heads for multi-head self-attention, masked multi-head sparse self-attention, and multi-head sparse cross-attention. We set the number of multi-granularity encoder layer L_1 to 5 and summary decoder layer L_2 to 6. We set dropout (Srivastava et al., 2014) rate to 0.1 and use Adam optimizer with an initial learning rate $\alpha = 0.0001$, momentum $\beta_1 = 0.9$, $\beta_2 = 0.999$ and weight decay $\epsilon = 10^{-5}$. When the valid loss on the development set increases for two consecutive epochs, the learning rate is halved. We use a mini-batch size of 10, and set the hyper-parameter $k = 5$ and $\lambda = 2$. Given the salience score predicted by the sentence extractor, we apply a simple greedy procedure to select sentences. We select one sentence based on the descending order of the salience scores and append to the extracted summary until the summary reaches 300 words. We disallow repeating the same trigram (Paulus et al., 2018; Edunov et al., 2019) and use beam search with a beam size of 5 for summary generator.

4.3 Metrics and Baselines

We use ROUGE (Lin, 2004) to evaluate the produced summary in our experiments. Following previous work, we report ROUGE F1[1] on Multi-News dataset. We compare our model with several typical baselines and several baselines proposed in the latest years.

Lead-3 is an extractive baseline which concatenates the first-3 sentences of each source document as a summary. LexRank (Erkan and Radev, 2004) is an unsupervised graph based method for computing relative importance in extractive summarization. TextRank (Mihalcea and Tarau, 2004) is also an unsupervised algorithm while sentence importance scores are computed based on eigenvector centrality within weighted-graphs for extractive sentence summarization. MMR (Carbonell and Goldstein, 1998) extracts sentences with a ranked list of the candidate sentences based on the relevance and redundancy. HIBERT (Zhang et al., 2019) first encodes each sentence using the sentence Transformer encoder, and then encode the whole document using the document Transformer encoder. It is a single document summarization model and cannot handle the hierarchical relationship of documents. We migrate it to multi-document summarization by concatenating multiple source documents into a long sequence. These extractive methods are set to give an output of 300 tokens. PGN (See et al., 2017) is an RNN based model with an attention mechanism and allows the system to copy words from the source text via pointing for abstractive summarization. CopyTransformer (Gehrmann et al., 2018) augments Transformer with one of the attention heads chosen randomly as the copy distribution. Hi-MAP (Fabbri et al., 2019) expands the pointer-generator network model into a hierarchical network and integrates an MMR module to calculate sentence-level scores, which is trained on the Multi-News corpus. The baseline above has been compared and reported in the Fabbri et al. (2019), which releases the Multi-News dataset, and we directly cite the results of the above methods from this paper. HT (Liu and Lapata, 2019) is a Transformer based model with an attention mechanism to share information cross-document for abstractive multi-document summarization. It is used

Model	R-1	R-2	R-SU4
Lead-3	39.41	11.77	14.51
LexRank (Erkan and Radev, 2004)	38.27	12.70	13.20
TextRank (Mihalcea and Tarau, 2004)	38.44	13.10	13.50
MMR(Carbonell and Goldstein, 1998)	38.77	11.98	12.91
HIBERT (Zhang et al., 2019)	43.86	14.62	18.34
PGN (See et al., 2017)	41.85	12.91	16.46
CopyTransformer(Gehrmann et al., 2018)	43.57	14.03	17.37
Hi-MAP(Fabbri et al., 2019)	43.47	14.89	17.41
HF(Liu and Lapata, 2019)	43.85	15.60	18.80
MGSum-*ext*	44.75	15.75	19.30
MGSum-*abs*	**46.00**	**16.81**	**20.09**
oracle ext	49.02	29.78	29.19

Table 2: ROUGE F1 evaluation results on the Multi-News test set.

[1]The ROUGE evaluation option: -c 95 -2 4 -U -r 1000 -n 4 -w 1.2 -a

initially to generate Wikipedia, and we reproduce their method for the multi-document news summarization.

4.4 Automatic Evaluation

Following previous work, we report ROUGE-1 (unigram), ROUGE-2 (bigram), and ROUGE-SU4 (skip bigrams with a maximum distance of 4 words) scores as the metrics for automatic evaluation (Lin and Hovy, 2003). In Table 2, we report the results on the Multi-News test set and our proposed multi-granularity model (denoted as MGSum) outperforms various previous models. Our abstractive method achieves scores of 46.00, 16.81, and 20.09 on the three ROUGE metrics while our extractive method achieves scores of 44.75, 15.75, and 19.30 on the three ROUGE metrics. We can also see that the abstractive methods perform better than the extractive methods. We attribute this result to the observation that the gold summary of this dataset tends to use new expressions to summarize the original input documents.

Owing to the characteristics of the news, lead-3 is superior to all unsupervised extractive methods. Our extractive method achieves about 1.13 points improvement on ROUGE-2 F1 compared with HIBERT. We attribute the improvement to two aspects: Firstly, the abstractive objective can promote the recognition of important sentences for the extractive model with the multi-granularity interaction network. Besides, while extractive gold-label sequences are obtained by greedily optimizing ROUGE-2 F1 on the gold-standard summary, gold labels may not be accurate. Joint learning of two objectives may correct some biases for the extractive model due to the inaccurate labels. We calculate the oracle result based on the gold-label extractive sequences, which achieves a score of 29.78 on ROUGE-2 F1 and is 14.03 points higher than the score of our extractive method. While there is a big gap between our model and the oracle, more efforts can be made to improve extractive performance.

Among the abstractive baselines, CopyTransformer performs much better than PGN and achieves 1.12 points improvement on the ROUGE-2 F1, which demonstrates the superiority of the Transformer architecture. Our abstractive model gains an improvement of 2.78 points compared with CopyTransformer, 1.92 points compared with Hi-MAP, and 1.21 points compared with HF on ROUGE-2 F1, which verifies the effectiveness of the proposed multi-granularity interaction network for the summary generation.

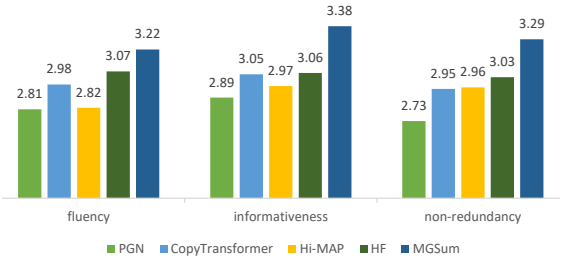

Figure 3: Human evaluation. The compared system summaries are rated on a Likert scale of 1(worst) to 5(best).

4.5 Human Evaluation

To evaluate the linguistic quality of generated summaries, we carry out a human evaluation. We focus on three aspects: **fluency**, **informativeness**, and **non-redundancy**. The fluency indicator focuses on whether the summary is well-formed and grammatical. The informativeness indicator can reflect whether the summary covers salient points from the input documents. The measures whether the summary contains repeated information. We sample 100 instances from the Multi-News test set and employ 5 graduate students to rate each summary. Each human judgment evaluates all outputs of different systems for the same sample. 3 human judgments are obtained for every sample, and the final scores are averaged across different judges.

Results are presented in Figure 3. We can see that our model performs much better than all baselines. In the fluency indicator, our model achieves a high score of 3.22, which is higher than 2.98 of CopyTransformer and 3.07 of HF, indicating that our model can reduce the grammatical errors and improve the readability of the summary. In the informativeness indicator, our model is 0.32 better than HF on ROUGE-2 F1. It indicates that our model can effectively capture the salient information. In the non-redundancy indicator, MGSum outperforms all baselines by a large margin, that indicates the multi-granularity semantic information and joint learning with extractive summarization does help to avoid the repeating information of the generated summary.

Model	R-1	R-2	R-SU4
MGSum-*ext*	45.04	15.98	19.53
only sentence extractor	44.65	15.67	19.27
without doc representation	44.67	15.58	19.15
MGSum-*abs*	46.08	16.92	20.15
only summary generator	45.57	16.32	19.56
without doc representation	45.71	16.62	19.80
without doc&sent representation	44.05	15.31	18.27

Table 3: Results of ablation study on the Multi-News development set.

4.6 Ablation Study

We perform an ablation study on the development set to investigate the influence of different modules in our proposed MGSum model. Modules are tested in four ways: (1) we remove the sentence extractor and only train the generator to verify the effectiveness of joint learning on the abstractive summarization; (2) we remove the summary generator part and only train the sentence extractor to verify the effectiveness of joint learning on the extractive summarization; (3) we remove the document representation and use only the sentence and word representations to verify the effectiveness of the document granularity semantic information; (4) We remove the document and sentence representation and use only the word representation to verify the importance of the sentence representation further. Since there are no interactions between the sentences of different documents without document representations, we establish connections between all sentences after the document representation is removed. Furthermore, we also establish connections between all the words after the sentence representation is removed, and the model degenerates into Transformer at this time.

Table 3 presents the results. We find that the ROUGE-2 F1 score of extractive summarization drops by 0.31 after the summary generator is removed. This indicates that the joint learning method helps extractive summarization to benefit from the abstractive summarization. ROUGE-2 F1 score of abstractive summarization drops by 0.6 after the sentence extractor is removed. This indicates that extractive summarization does help abstractive summarization identify important sentences during the interactive encoding phrase. ROUGE-2 F1 score of extractive summarization drops by 0.4, while the ROUGE-2 F1 score of abstractive summarization drops by 0.3 after the document representation is removed. It indicates establishing the document representation to simulate the relationships between documents is necessary to improve the performance of both extractive and abstractive summarization. ROUGE-2 F1 score drops by 1.61 compared with MGSum and 1.01 compared with the only summary generator after removing both the document representation and the sentence representation. And there is no extractive summarization to co-promote the recognition of important information for abstractive summarization after the sentence representation is removed. It indicates the semantic information of sentence granularity is of great importance to encode multi-

Human: – it ' s a race for the governor ' s mansion in 11 states today , and the gop could end the night at the helm of more than two-thirds of the 50 states . the gop currently controls 29 of the country ' s top state offices ; it ' s expected to keep the three republican ones that are up for grabs (utah , north dakota , and indiana) , and wrest north carolina from the dems . that brings its toll to 30 , with the potential to take three more , reports npr . races in montana , new hampshire , and washington are still too close to call , and in all three , democrat incumbents aren ' t seeking reelection . the results could have a big impact on health care , since a supreme court ruling grants states the ability to opt out of obamacare ' s medicaid expansion . " a romney victory would dramatically empower republican governors , " said one analyst . click for npr ' s state-by-state breakdown of what could happen .
HF: – delaware , new hampshire , and missouri are expected to notch safe wins in 11 states , reports npr . the state ' s top state of the state has seen its top state offices , and it ' s expected to be more than two-thirds of the nation ' s state , reports the washington post . the top 10 : montana , montana , and rhode island . indiana : missouri : the state is home to the top of the list of state offices . new hampshire : montana : incumbent john kasich : he ' s the first woman to hold a state seat in the state , notes the huffington post . north carolina : the only state to win gop-held seats in vermont and delaware . new jersey : the biggest state in the history of the year has seen a population of around 40 % of the population , reports ap . montana : new hampshire and missouri : a state department of emergency has been declared a state of emergency . click for the full list , or check out a list of the states that voted tonight .
MGSum-*ext*: gop eyes gains as voters in 11 states pick governors enlarge this image toggle caption jim cole/ap jim cole/ap voters in 11 states will pick their governors tonight , and republicans appear on track to increase their numbers by at least one , with the potential to extend their hold to more than two-thirds of the nation ' s top state offices . and that ' s health care , says political scientist thad kousser , co-author of the power of american governors . " republicans currently hold 29 governorships , democrats have 20 , and rhode island ' s gov . lincoln chafee is an independent . eight of the gubernatorial seats up for grabs are now held by democrats ; three are in republican hands . polls and race analysts suggest that only three of tonight ' s contests are considered competitive , all in states where incumbent democratic governors aren ' t running again : montana , new hampshire and washington .
MGSum-*abs*: – voters in 11 states will pick their governors tonight , and republicans appear on track to increase their numbers by at least one , with the potential to extend their hold to more than two-thirds of the nation ' s top state offices . republicans currently hold 29 governorships , democrats have 20 , and rhode island ' s gov . lincoln chafee is an independent . the seat is expected to be won by former charlotte mayor walter dalton , who won his last election with 65 % of the vote , reports the washington post . democrats are expected to hold on to their seats in west virginia and missouri , and democrats are likely to hold seats in vermont and delaware , reports npr . polls and race analysts say that only three of tonight ' s contests are considered competitive , and all in states where incumbent democratic governors aren ' t running again . " no matter who wins the presidency , national politics is going to be stalemated on the affordable care act , " says one political scientist .

Table 4: Sample summaries for a document cluster from the Multi-News test set. The underline shows the overlap parts between our abstractive summary and human summary. The extractive and abstractive summary generated by MGSum have the high overlap (different overlaps are marked in different colors).

ple documents.

4.7 Case Study

In Table 4, we present example summaries generated by strong baseline HF, and our extractive and abstractive methods. The output of our model has the highest overlap with the ground truth. Moreover, our extractive and abstractive summary show consistent behavior with the high overlap, which further indicates that the two methods can jointly promote the recognition of important information. Compared with the extracted summary, the generated summary is more concise and coherent.

5 Conclusion and Future Work

In this work, we propose a novel multi-granularity interaction network to encode semantic representations for documents, sentences, and words. It can unify the extractive and abstractive summarization by utilizing the word representations to generate the abstractive summary and the sentence representations to extract sentences. Experiment results show that the proposed method significantly outperforms all strong baseline methods and achieves the best result on the Multi-News dataset.

In the future, we will introduce more tasks like document ranking to supervise the learning of the multi-granularity representations for further improvement.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61772036), Tencent AI Lab Rhino-Bird Focused Research Program (No.JR201953) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We thank the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2153–2159. AAAI Press.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 335–336. ACM.

Asli Çelikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1662–1675. 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 675–686. Association for Computational Linguistics.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards coherent multi-document summarization. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 1163–1173. The Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 615–621. Association for Computational Linguistics.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4052–4059. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

Alexander Richard Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: A

large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1074–1084. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4098–4109. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4131–4141. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*. The Association for Computational Linguistics.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5070–5081. 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 , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 404–411. ACL.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Sandeep Subramanian, Raymond Li, Jonathan Pilault, and Christopher J. Pal. 2019. On extractive and abstractive neural document summarization with transformer language models. *CoRR*, abs/1909.03186.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir R. Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 452–462. Association for Computational Linguistics.

Jianmin Zhang, Jiwei Tan, and Xiaojun Wan. 2018. Towards a neural network approach to abstractive multi-document summarization. *CoRR*, abs/1804.09010.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5059–5069. Association for Computational Linguistics.

Guangxiang Zhao, Junyang Lin, Zhiyuan Zhang, Xuancheng Ren, Qi Su, and Xu Sun. 2019. Explicit sparse transformer: Concentrated attention through explicit selection. *CoRR*, abs/1912.11637.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 654–663. Association for Computational Linguistics.