Unsupervised Joint Learning for Headline Generation and Discourse Structure of Reviews

Masaru Isonuma*[1] Junichiro Mori*[1] Ichiro Sakata*[1]

*[1] The University of Tokyo

Recently, using a large amount of reference summary, supervised neural summarization models have achieved success. However, such datasets are rare, and trained models cannot be shared across domains. Although an unsupervised approach is a possible solution, models applicable for single-document summary or headline generation have not been established. Our work focuses on generating headlines for reviews, without supervision. We assume that reviews contain a discourse tree in which the headline is the root and the child sentences elaborate on the parent. By estimating the parent from their child recursively, our model learns such a structure and generates a headline that describes the entire review. Through the evaluation of the generated headline on actual reviews, our model achieved competitive performance with supervised models, especially on relatively long reviews. In induced structures, we confirmed that the child-sentences explain the parent in detail and generated headline abstracts for the entire review.

1. Introduction

The need for automatic document summarization is widely increasing, with the recently growing vast amounts of online textual data such as reviews on E-commerce websites. Under these circumstances, supervised neural network models have widely achieved success, using a large amount of reference summary. However, the model trained from them cannot be adopted in other domains as salient phrases are not common across domains. Few or no examples of summaries exist for most documents, and preparing such large volumes of reference summaries is very expensive.

An unsupervised approach is a possible solution for such a problem. Traditionally, the unsupervised approach has been widely applied to sentence extraction [Erkan 04]. The extractive approach can be effective for some types of documents, e.g. news articles, since the salient sentences should be the summary even though they describe only a part of the document. On the other hand, as for reviews, an appropriate summary is generally concise sentences that summarize the entire review. Therefore, the abstractive method is more effective for reviews because it condenses an entire review via paraphrasing and generalization [Gerani 14]. Our work focuses on headline generation of reviews; a kind of abstractive summarization tasks, without supervision.

Abstractive summarization techniques sometimes use discourse parsers [Hirao 13, Gerani 14]; however, [Ji 17] indicates the limitations of using external discourse parsers. In this context, [Liu 18] proposed a model that encodes a document while automatically inducing the discourse tree. Following [Liu 18], we aim to generate a headline based on the discourse tree, which is automatically constructed.

Figure 1 shows an example of a review about a jigsaw puzzle and its dependency-based discourse tree [Hirao 13] constructed manually. The headline describes its quality, and the child sentences explain it in terms of size and thickness. Each child sentence elaborate on the parent in detail.

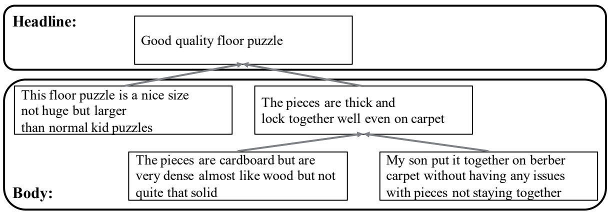

Headline:
Good quality floor puzzle
This floor puzzle is a nice size not huge but larger than normal kid puzzles
The pieces are thick and lock together well even on carpet
The pieces are cardboard but are very dense almost like wood but not quite that solid
My son put it together on berber carpet without having any issues with pieces not staying together
Body:

Figure 1: An example of discourse tree in a review

Thus, we assume that reviews can generally be described as a multi-root non-projective discourse tree in which the headline is the root, and the sentences construct each node. In such trees, child sentences elaborate on or provide background to the parent sentence. By estimating the parent from their children recursively, our model learns such a discourse tree and generates a headline.

In this work, we propose a model that generates the headline of a single review without reference through learning the discourse tree. Although there has been previous work without supervision using the Abstract Meaning Representation (AMR) parser [Dohare 18], our work is, to our knowledge, the first headline-generation model that requires no external parser. Through the evaluation of the headlines generated for actual online reviews and the induced discourse tree, we validate our assumption; child sentences elaborate on the parent sentences, and as a result, the generated headline summarizes the entire review.

2. Related research

2.1 Un-/Semi-supervised Summarization

Most of unsupervised summarization techniques have been focused on sentence extraction. [Erkan 04] constructed a graph that consists of sentences as the nodes, with their similarities as the edge. They extracted sentences with a higher eigenvector centrality so that the selected sentences are heavily connected to each sentence in the input document. On the other hand, with the re-

Contact: Masaru Isonuma, isonuma@ipr-ctr.t.u-tokyo.ac.jp

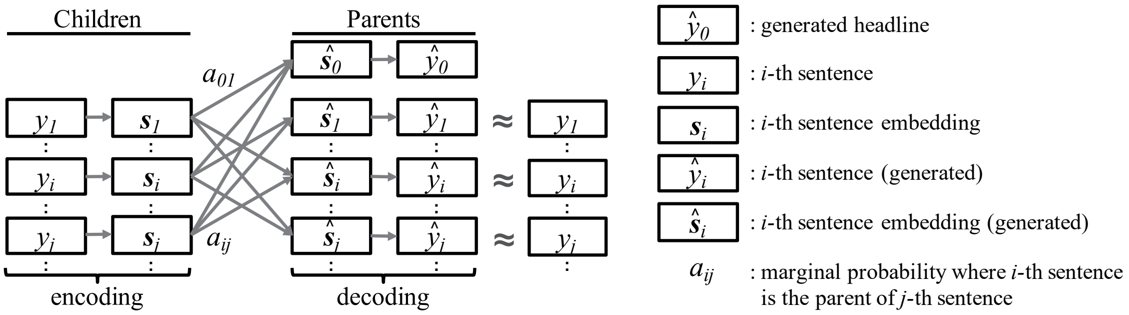


Figure 2: The outline of the proposed model

cently growing neural summarization models, unsupervised or semi-supervised summary generation is being attempted. [Miao 16] introduced the idea that compressed sentences that do not lose meaningful phrases can decode the input sentences. They applied a variational auto-encoder for the sentence compression task. [Chu 18] proposed a model that consists of an auto-encoder trained so that the mean of the representations of the input reviews is decoded to the summary. However, their model does not aim to generate the summary or headline of a single document. Against such a background, our model generates the headline of a single review without reference.

2.2 Discourse Parsing and its Application

Discourse parsing is broadly researched and used for various applications that utilize knowledge on conjunctions or corpora in which rhetorical structure is annotated. [Hirao 13] transformed a rhetorical structure theory-based discourse tree (RST-DT) into a dependency-based discourse tree to take a tree-trimming approach to summarization. [Ji 17] also constructed a dependency-based discourse tree, and applied a recursive neural network model for document classification. They indicated the limitations of using external parsers, showing that the performance depends on the amount of the RST-DT and the domain of documents. Against such a background, [Liu 18] proposed a model that can encode a document while automatically inducing a latent document structure. They reported that the child presents additional information regarding the parent in the induced document structure. Inspired by [Liu 18], we obtain a discourse tree without external parsers by estimating parent sentences from the children.

3. Proposed Model

In this section, we present our headline generation model by inducing a discourse tree without external parsers. Figure 2 shows the outline of our proposed model. In the following, we explain the training method and computation of the marginal probability of the dependency edges.

3.1 Model Training

In Figure 2, y_i and $\boldsymbol{s}_i \in \mathcal{R}^d$ indicate i-th sentence and its embedding in a document $D = \{y_1, y_2, \dots, y_n\}$ respectively. w_i^t is t-th word in a sentence $y_i = \{w_i^1, w_i^2, \dots, w_i^l\}$. $\boldsymbol{s}_i$ is computed via max-pooling operation across hidden states $\boldsymbol{h}_i^t \in \mathcal{R}^d$ of Bi-directional Gated Recurrent Units (Bi-GRU):

$$\begin{aligned} \overrightarrow{\boldsymbol{h}}_i^t &= \overrightarrow{\mathrm{GRU}}(\overrightarrow{\boldsymbol{h}}_i^{t-1}, w_i^t) & (1)\\ \overleftarrow{\boldsymbol{h}}_i^t &= \overleftarrow{\mathrm{GRU}}(\overleftarrow{\boldsymbol{h}}_i^{t+1}, w_i^t) & (2)\\ \boldsymbol{h}_i^t &= [\overrightarrow{\boldsymbol{h}}_i^t, \overleftarrow{\boldsymbol{h}}_i^t] & (3)\\ \forall m \in \{1, \dots, d\},\ \boldsymbol{s}_{i,m} &= \max_t \boldsymbol{h}_{i,m}^t & (4) \end{aligned}$$

Here, we assume that the document D involves a discourse tree in which the root is the headline, and all the sentences are the nodes. We denote a_{ij} and a_{0j} as the marginal probability, where sentence i and the root are the parent node of sentence j under the constraint $\sum_{i=0}^{n} a_{ij} = 1$ (see Figure 2). From the sentence embeddings and a_{ij}, we compute the embedding of the parent sentence $\hat{s}_i$ and that of the headline $\hat{s}_0$. $\hat{s}_i$ ($i \in \{0, \dots, n\}$) are defined with parameters $\boldsymbol{W}_s \in \mathcal{R}^{d*d}$ and $\boldsymbol{b}_s \in \mathcal{R}^d$ as shown below:

$$\hat{\boldsymbol{s}}_i = \tanh\Big\{\boldsymbol{W}_s(\sum_{j=1}^{n} a_{ij}\boldsymbol{s}_j) + \boldsymbol{b}_s\Big\} \quad (5)$$

The higher the marginal probability a_{ij} is, the more the information of $\boldsymbol{s}_j$ are input into $\hat{\boldsymbol{s}}_i$. From $\hat{\boldsymbol{s}}_i$, the GRU-decoder learns to reconstruct the sentence i, i.e., search the parameters $\boldsymbol{\theta}$ that maximize the following log likelihood:

$$\sum_{i=1}^{n}\sum_{t=1}^{l} \log P(w_i^t | w_i^{<t}, \hat{\boldsymbol{s}}_i, \boldsymbol{\theta}) \quad (6)$$

Here, we explain how training the model contributes to headline generation. By reconstructing the sentences in documents, the decoder learns a language model to generate grammatical sentences. Therefore, the model can decode the headline embedding $\hat{\boldsymbol{s}}_0$ as a fluent sentence.

Besides, the more the j-th sentence contributes to generating the i-th sentence, the higher a_{ij} can be. This mechanism models our assumption; child sentences can generate their parent, but not vice versa, because the children have more information than the parents. Based on this assumption, the most abstractive k-th sentences in the body make less contribution to reconstruction of any other sentences. From the constraint $\sum_{i=0}^{n} a_{ik} = 1$, a_{0k} is expected to be larger and contribute toward generating the headline.

3.2 Marginal Probability of Dependency

We explain how to calculate the marginal probability a_{ij}. We first define the unnormalized weight f_{ij} of the edge between a parent node i and the child node j via a bilinear function and a linear function. For convenience, we assume the weighted adjacency matrix $\boldsymbol{F} = (f_{ij}) \in \mathcal{R}^{(n+1)*(n+1)}$. The index of the first column and row are 0, which denotes the root node. We assume that the discourse structure can be described as a multi-root non-projective tree. Therefore, based on [Liu 18], f_{ij} is defined as :

$$f_{ij} = \begin{cases} \exp(\boldsymbol{w}_r^\top \boldsymbol{s}_j) & (i = 0 \land j \geq 1) \\ \exp(\boldsymbol{p}_i^\top \boldsymbol{W}_f \boldsymbol{c}_j) & (i \geq 1 \land j \geq 1 \land i \neq j) \\ 0 & (j = 0 \lor i = j) \end{cases} \quad (7)$$

$$\boldsymbol{p}_i = \tanh(\boldsymbol{W}_p \boldsymbol{s}_i + \boldsymbol{b}_p) \quad (8)$$

$$\boldsymbol{c}_j = \tanh(\boldsymbol{W}_c \boldsymbol{s}_j + \boldsymbol{b}_c) \quad (9)$$

where $\boldsymbol{W}_f \in \mathcal{R}^{d*d}$ and $\boldsymbol{w}_r \in \mathcal{R}^d$ are parameters for the transformation. $\boldsymbol{W}_p \in \mathcal{R}^{d*d}$ and $\boldsymbol{b}_p \in \mathcal{R}^d$ are the weights and the bias for constructing the representation of the parent nodes. $\boldsymbol{W}_c \in \mathcal{R}^{d*d}$ and $\boldsymbol{b}_c \in \mathcal{R}^d$ are those of the child nodes.

We normalize f_{ij} into a_{ij}, following [Koo 07]. a_{ij} corresponds the proportion of the total weight of all the spanning trees containing the edge (i, j):

$$\begin{aligned} a_{ij}(\boldsymbol{F}) &= \frac{\sum_{g \in G:(i,j) \in g} v(g|\boldsymbol{F})}{\sum_{g \in G} v(g|\boldsymbol{F})} \\ &= \frac{\partial \log Z(\boldsymbol{F})}{\partial f_{ij}} \end{aligned} \quad (10)$$

$$v(g|\boldsymbol{F}) = \prod_{(i,j) \in g} f_{ij} \quad (11)$$

$$Z(\boldsymbol{F}) = \sum_{g \in G} v(g|\boldsymbol{F}) \quad (12)$$

Here, G denotes the set of all the spanning trees in a document D. $v(g|\boldsymbol{F})$ is the weight of a tree $g \in G$, and $Z(\boldsymbol{F})$ denotes the sum of the weights of all the trees in G. From the Matrix-Tree Theorem, $Z(\boldsymbol{F})$ can be rephrased as:

$$Z(\boldsymbol{F}) = L^{(0,0)}(\boldsymbol{F}) \quad (13)$$

where $L(\boldsymbol{F})$ and $L^{(0,0)}(\boldsymbol{F})$ be the Laplacian matrix of $\boldsymbol{F}$ and its minor, with respect to the row 0 and the column 0.

4. Experiments

In this section we present our experiments for evaluating the performance of headline generation. We compared the generated headlines on actual online reviews. The following explains the details of our experiments and the results.

Table 1: ROUGE-F1 score on the evaluation set （%）

Models	ROUGE-1	ROUGE-2	ROUGE-L
Seq-Seq	12.8	1.8	10.2
Seq-Seq-att	13.8	2.5	10.9
Our Model	11.4	1.6	9.1

4.1 Dataset

Our experiment uses Amazon product review data (Toys and Games) [He 16], which [Ma 18] used as evaluation for their supervised headline generation model. This dataset contains actual online reviews and their headlines.

Because our model generates a headline via learning the discourse tree, we assume that training will fail if the number of sentences in the review is too small. Therefore, we use reviews in which the number of sentence is in $[10, 20)$ for training and $[5, 20)$ for validation and evaluation. The number of reviews for training, validation, and evaluation are 21791, 416, and 464, respectively.

4.2 Experimental Details

For all the experiments, our model has 300-dimensional word embeddings and Bi-GRU with 256-dimensional hidden states. We initialize the word embeddings with pre-trained GloVe (840B tokens) [Pennington 14]. We train the model using Ada-grad with a learning rate of 10^{-1}, an initial accumulator value of 0.1, and a batch size of 16. At evaluation time, we use a beam search with a beam size of 10.

Similar to [Ma 18], our evaluation metric is the ROUGE-F1 score. We use ROUGE-1, ROUGE-2, and ROUGE-L.

4.3 Baseline

Following previous work [Ma 18], our baseline models are the supervised sequence-to-sequence models for headline generation. We denote the sequence-to-sequence model as Seq-Seq and that with the attention mechanism as Seq-Seq-att. Implementation details are the same as above.

4.4 Results

Table 1 shows the ROUGE score of our model and the baseline models on the evaluation sets. Our model achieves a slightly low performance, compared to Seq-Seq.

In Figure 3, we report the performance on the evaluation sets in which the number of the sentences are in $[5, 10)$, $[10, 15)$ and $[15, 20)$, respectively. We compared to the supervised baseline model (Seq-Seq-Att). In the case of the dataset with under 10 sentences, the performance of our model is inferior to that of the baseline. On the other hand, on the one with 10 or more sentences, our model achieves a competitive performance as for ROUGE-1 and ROUGE-L. Because our model generates headlines via learning discourse tree, our model sometimes appears to fail at constructing a tree as for short documents. It results in a decline in the performance.

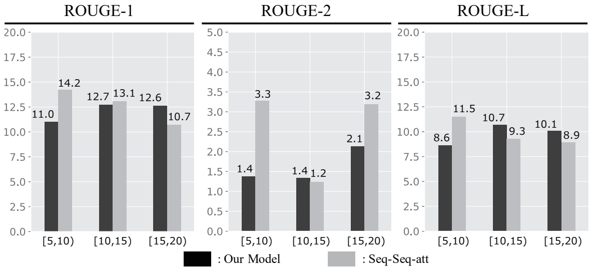
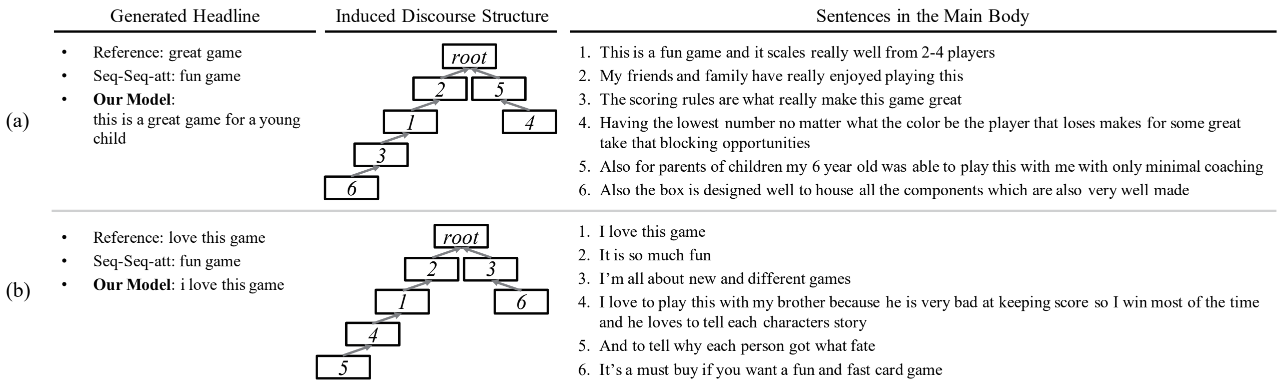


Figure 3: ROUGE-F1 score on evaluation set with various number of sentences



(a)

- Reference: great game
- Seq-Seq-att: fun game
- **Our Model**: this is a great game for a young child

1. This is a fun game and it scales really well from 2-4 players
2. My friends and family have really enjoyed playing this
3. The scoring rules are what really make this game great
4. Having the lowest number no matter what the color be the player that loses makes for some great take that blocking opportunities
5. Also for parents of children my 6 year old was able to play this with me with only minimal coaching
6. Also the box is designed well to house all the components which are also very well made

(b)

- Reference: love this game
- Seq-Seq-att: fun game
- **Our Model**: i love this game

1. I love this game
2. It is so much fun
3. I'm all about new and different games
4. I love to play this with my brother because he is very bad at keeping score so I win most of the time and he loves to tell each characters story
5. And to tell why each person got what fate
6. It's a must buy if you want a fun and fast card game

Figure 4: Examples of generated headline and induced discourse tree

5. Discussion

Figure 4 shows the generated headline and the discourse tree induced by our model. We obtain the maximum spanning tree from the probability distribution of dependency, using Chu–Liu–Edmonds algorithm. In Figure 4 (a), our model generates the headline, "this is a great game for a young child" while the actual headline is "great game." On the discourse tree, the child nodes of the root are the 2nd and 5th sentence. Both of them elaborate on the generated headline. The 3rd sentence explains the parent, the 1st sentence, by explaining the cause of fun.

Figure 4 (b) shows that the generated headline is "i love this game," while the reference is "love this game". In the induced tree, the 2nd sentence elaborates on the generated headline, while the 3rd sentence describes its background. The 4th and 5th sentences describe why the author loves the game, i.e., they explain the 1st sentence in detail.

As shown above, we confirmed that the child sentences elaborate on the parent in the induced discourse tree, while the headline abstracts for the child sentences.

6. Conclusion

In this work, we proposed a model to generate the headline of reviews by learning the latent discourse tree with neither a reference summary nor an external parser.

We evaluate our proposed model in comparison with supervised models on actual reviews. Our model achieves a competitive performance when the number of sentences is relatively large. On the reviews that contain few sentences, our model fails to construct the discourse tree and generate a reasonable headline.

Furthermore, our model induced a discourse tree in which the child sentences elaborate on the parent. We also confirmed that the headline abstracts for the entire review.

Acknowledgements

This work was supported by CREST, JST, the New Energy and Industrial Technology Development Organization (NEDO) and Deloitte Tohmatsu Financial Advisory LLC.

References

[Chu 18] Chu, E. and Liu, P. J.: Unsupervised Neural Multi-document Abstractive Summarization, *arXiv preprint arXiv:1810.05739* (2018)

[Dohare 18] Dohare, S., et al.: Unsupervised Semantic Abstractive Summarization, in *ACL Student Research Workshop*, pp. 74–83 (2018)

[Erkan 04] Erkan, G. and Radev, D. R.: LexPageRank: Prestige in Multi-Document Text Summarization, in *EMNLP*, Vol. 4, pp. 365–371 (2004)

[Gerani 14] Gerani, S., et al.: Abstractive summarization of product reviews using discourse structure, in *EMNLP*, pp. 1602–1613 (2014)

[He 16] He, R. and McAuley, J.: Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering, in *WWW*, pp. 507–517 (2016)

[Hirao 13] Hirao, T., et al.: Single-document summarization as a tree knapsack problem, in *EMNLP*, pp. 1515–1520 (2013)

[Ji 17] Ji, Y. and Smith, N. A.: Neural Discourse Structure for Text Categorization, in *ACL*, Vol. 1, pp. 996–1005 (2017)

[Koo 07] Koo, T., et al.: Structured prediction models via the matrix-tree theorem, in *EMNLP-CoNLL* (2007)

[Liu 18] Liu, Y. and Lapata, M.: Learning structured text representations, *TACL*, Vol. 6, pp. 63–75 (2018)

[Ma 18] Ma, S., et al.: A Hierarchical End-to-End Model for Jointly Improving Text Summarization and Sentiment Classification, in *IJCAI*, pp. 4251–4257 (2018)

[Miao 16] Miao, Y. and Blunsom, P.: Language as a Latent Variable: Discrete Generative Models for Sentence Compression, in *EMNLP*, pp. 319–328 (2016)

[Pennington 14] Pennington, J., et al.: Glove: Global vectors for word representation, in *EMNLP*, pp. 1532–1543 (2014)