Conversation Summarization of Japanese Text by a Recursive Attention Model

Takashi Kuremoto*, Daisuke Shitamoto, Shingo Mabu
Yamaguchi University, Yamaguchi, Japan
*Corresponding author: wu@yamaguchi-u.ac.jp

Abstract: Text summarization study has been developed several years inspiring by the machine translation studies. Meanwhile, conversations of our lives are difficult to be interrupted by the machine because there are a lot of utterance words independent from grammars and vague expressions of spoken-language. In this paper, a challenge of conversation summarization is proposed by a recursive attention model (RAM), which is a variant of neural networks based language model (NNLM). Experiment results of Japanese conversation summarization showed the priority of the proposed RAM comparing to the conventional method, attention-based summarization system (ABS), given by Rush, Chopra, and Weston in 2015.

Key-Words: *abstractive summarization, recursive attention model, recurrent neural network, language model*

1. Introduction

It is interest and important to automatically summarize texts, documents, reports, articles, books, or textual contents in websites in nowadays big data era. As a field of natural language processing (NLP) study, abstractive summarization methods, i.e., different from stochastic words extraction, have been developed since 2000s [1-8]. The basic approach of summarization comes from the development of machine translation methods which used recurrent neural networks (RNN) generally [2] [3]. However, conversations of our lives are difficult to be interrupted by the machine because there are a lot of utterance words independent from grammars and vague expressions of spoken-language. So the abstractive summarization of spoken conversation is more difficult than the normal text summarization [4]. In this paper, a challenge of conversation summarization is proposed by a recursive attention model (RAM), which is a variant of neural networks based language model (NNLM). The proposed model RAM is based on the previous models in [3] [5], by adding a recursive layer to the part of decoder, to address the important words for the abstractive summarization. Comparing to the conventional method "attention-based summarization system (ABS)" [3], the results of experiments with Japanese conversation texts showed the priority of the proposed RAM.

Received: 2020/06/03, Accepted: 2020/08/09
*Corresponding author: Takashi Kuremoto
E-mail address:wu@yamaguchi-u.ac.jp

2. Recursive Attention Model

In this section, the conventional abstractive summarization method RAS is introduced and then an improved method RAM is proposed.

2.1 ABS

Attention-based summarization system (ABS) [3] proposed by Rush, Chopra, and Weston in 2015 is composed by neural network language model (NNLM) [2] which serves as a decoder and an attention-based contextual encoder as shown in Fig. 1.

Let the input consists of a sequence of M words $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M$ coming from a fixed vocabulary V of

size $|V|$,where $\mathbf{x}_i \in \{0,1\}^V, i \in \{1,...,M\}$ is an indicator vector, and X as the set of possible inputs. A summarizer takes **x** as input and outputs a shortened sentence **y** ($\mathbf{y}_1$, $\mathbf{y}_2$, …, $\mathbf{y}_j$, …, $\mathbf{y}_N$) of length $N < M$, where $\mathbf{y}_j \in \{0,1\}^V, j \in \{1,...,N\}$ is an indicator vector,. Let the possible output sentences $\mathbf{y} \in Y, Y \subset (\{0,1\}^V,...\{0,1\}^V)$, we say a system is a summarizer under a optimization function $s: \mathrm{X} \times \mathrm{Y}$ a R as shown in Eq. (1).

$$\arg\max_{\mathbf{y} \in Y} s(\mathbf{x}, \mathbf{y}) \tag{1}$$

To find an optimal summarization **y**($\mathbf{y}_1$, $\mathbf{y}_2$, …, $\mathbf{y}_j$, …, $\mathbf{y}_N$), we can set a window of size C for $\mathbf{y}_t$, i.e., $\mathbf{y}_C \equiv \mathbf{y}_{[t-C+1,...,t]}$, and the conditional log-probability of the summarizing system can be given by following:

$$s(\mathrm{x}, \mathrm{y}) = \log p(\mathrm{y} \mid \mathrm{x}; \theta) \approx \sum_{j=0}^{N-1} \log p(\mathrm{y}_{j+1} \mid \mathrm{x}, \mathrm{y}_c; \theta) \tag{2}$$

where $\boldsymbol{\theta}$ is the parameter vector of the model. The model is optimized by a K-size mini-batch stochastic gradient descent method minimizing the negative likelihood function as following.

$$NLL(\boldsymbol{\theta}) = -\sum_{k=1}^{K} \sum_{j=1}^{N-1} \log p(\mathbf{y}_{j+1}^k \mid \mathbf{x}^k, \mathbf{y}_C; \boldsymbol{\theta}) \tag{3}$$

$$\theta^{t+1} = \theta^t - \eta \frac{\partial NLL(\boldsymbol{\theta})}{\partial \theta} \tag{4}$$

where $\theta^{t=0}$ is given by random number in (-0.2, 0.2), $\eta \in (0.0, 1.0)$ is a learning rate.

The NNLM is given by follows.

$$p(\mathbf{y}_{j+1} \mid \mathbf{x}, \mathbf{y}_C; \boldsymbol{\theta}) \propto \exp(\mathbf{Vh} + \mathbf{Wenc}(\mathbf{x}, \mathbf{y}_C)) \tag{5}$$

$$h = \tanh(U\tilde{y}_c) \tag{6}$$

$$\mathbf{enc}(\mathbf{x}, \mathbf{y}_C) = \mathbf{p}^{\mathrm{T}} \bar{\mathbf{x}} \tag{7}$$

$$\bar{\mathrm{x}}_{\mathrm{i}} = \sum_{q=i-Q}^{i+Q} \tilde{\mathrm{x}}_q / Q \tag{8}$$

$$\tilde{\mathrm{x}} = [F\mathrm{x}_1, ..., F\mathrm{x}_{\mathrm{M}}] \tag{9}$$

$$p \propto \exp(\tilde{\mathrm{x}} \mathrm{P} \tilde{\mathrm{y}}_{\mathrm{c}}') \tag{10}$$

$$\tilde{y}_c' = [Gy_{i-C+1}, ..., Gy_i] \tag{11}$$

$$\tilde{y}_c = [Ey_{j-C+1}, ..., Ey_j] \tag{12}$$

where **V, U, W, P, G, E, F** are weights for connections. In fact, the weights for input words **F**, and the historic output words **G** indicate the attention altitudes for summarizations. In other words, "attention" means that the important words appeared in the historic stochastic output of the system $\mathbf{y}_C$ were utilized as a cooperative input to the system as shown in Fig.1. $\mathbf{enc}(\mathbf{x}, \mathbf{y}_C)$ in Eq. (7) indicates the encoder part of the system, which compresses the input information **x** and the historic output $\mathbf{y}_C$ by the synaptic connections **F** and **G** of middle layers $\tilde{\mathrm{x}}$ and $\tilde{y}_C'$, and the stochastic output function $p(\mathbf{y}_{j+1} \mid \mathbf{x}, \mathbf{y}_C; \boldsymbol{\theta})$ in Eq. (5) is also called the decoder part of system.

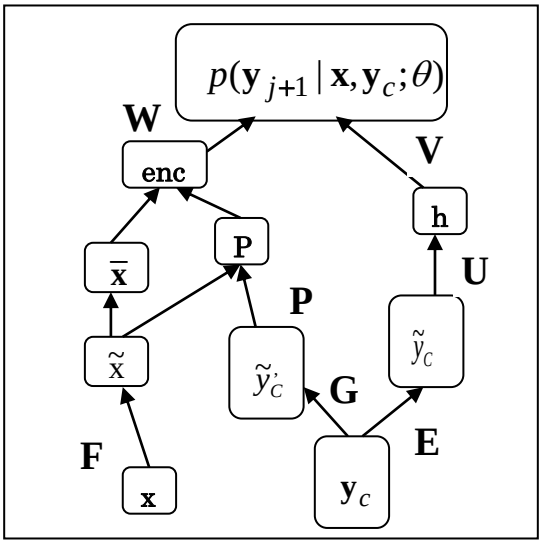

Fig. 1 A network diagram of attention-based summarization system (ABS) [3].

The principle of ABS can be simply abstracted as "long sentences **x** is compressed to shorter sentences **y** by adjusting the connection weights of middle layers of

the system **V, W, P, G, E, F** addressing some key words using teacher signals".

2.2 RAM

A recursive attention model (RAM) for summarization is proposed as shown in Fig. 2.

The output of hidden layers $\mathbf{h}_t^{enc}$ and $\mathbf{h}_t^{dec}$ are feedback to units $\mathbf{h}_{t-1}^{enc}$ and $\mathbf{h}_{t-1}^{dec}$ in the encoder $\mathbf{enc}(\mathbf{x}, \mathbf{y}_c)$ and the decoder $p(\mathbf{y}_{t+1} \mid \mathbf{x}, \mathbf{y}_t; \boldsymbol{\theta})$, respectively, at t step output. So we named the novel attention model with "recursive" which addresses the important words even in the middle layers of the encoder and the decoder.

The description of RAM is given by follows.

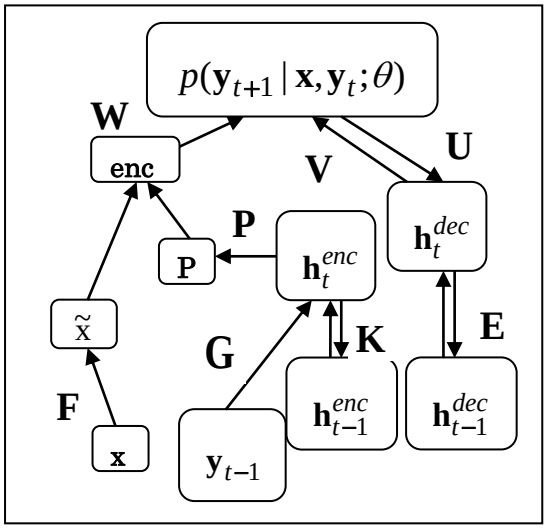

Fig.2 A network diagram of the proposed recursive attention model (RAM).

$$p(y_{t+1} \mid x, y_t; \theta) \propto \exp(V h_t^{dec} + W enc(\tilde{\mathrm{x}}, p)) \quad (13)$$

$$\mathbf{h}_t^{dec} = \tanh(\mathbf{U} p(\mathbf{y}_{t+1} \mid \mathbf{x}, \mathbf{y}_t; \theta) + \mathbf{E}\mathbf{h}_{t-1}^{dec}) \quad (14)$$

$$enc(x, y_c) = p^T \tilde{\mathrm{x}} \quad (15)$$

$$\tilde{\mathrm{x}} = [\mathrm{Fx}_1, ..., \mathrm{Fx}_{\mathrm{M}}] \quad (16)$$

$$\mathbf{p} = \exp(\mathbf{P}\mathbf{h}_t^{enc}) \quad (17)$$

$$\mathbf{h}_t^{enc} = \tanh(\mathbf{G}\mathbf{y}_{t-1} + \mathbf{K}\mathbf{h}_{t-1}^{enc}) \quad (18)$$

Symbols used in Eq. (13) to Eq. (18) indicate the same meaning in Eq. (5) to Eq. (12).

3. Experiments and Results

There were 20 Japanese texts used in the experiments to confirm the effectiveness of the proposed method. As shown in Table 1, there are verbose words existing in the conversation, additionally, because the text data were obtained by the speech recognition software, so there are some nonsensical words in the sentences. The target summarizations (teacher signals) were made manually. As shown in Table 1, the teacher signals only included the main words of a speaker, who was a main speaker such as a person in charge, a salesperson, a doctor, etc. In other word, the conversation summarization needs to abstract the contents of 2 or more people speeches, however, the case was not confirmed in the experiment of this paper.

The output of the proposed system RAM during the training process was shown in Table 2 and Table 3. It can be confirmed that the summarization of data No.1 and data 2 were output stably after 350 iterations of training.

Conventional summarization method ABS [3] was used to compare the performance of the proposed method RAM. The measure of precision was Rouge-L [1], which utilizes "*Recall*" and "*Precision*" to evalutate the longest common subsequence (LCS) between texts.

$$Rouge-L = \frac{2(Recall \times Precision)}{Recall + Precision} \quad (19)$$

$$Recall = \frac{LCS(summary_{words}, reference_{words})}{Rerefence_{words}} \quad (20)$$

$$Precision = \frac{LCS(summary_{words}, reference_{words})}{summary_{words}} \quad (21)$$

If the number of same words between 2 sentences is zero, then Rouge-L value is zero,

meanwhile when the sentences are completely with same words, the value is 1.0.

20 cross-validation was used in the experiments, i.e., 19 data were used as training samples and 1 unknown data was used as test data. Learning epoch (iteration times) was set as 1000, and loss value was output per 50 epochs. Optimizer of the model used Adam, an advance learning rule of neural networks [9]. From Fig. 3 and Fig. 4, it can be confirmed that the loss function (Eq. (3)) converged according to the number of learning times, and the conventional method ABS [3] showed a better learning performance in Fig. 3 comparing to the case of the proposed method shown by Fig. 4. The robustness of different models, however, is usually in inverse proportion to the convergence of learning process.

The Rouge-L values of the conventional method ABS and the proposed method RAM are shown in Table 4. The number of experiments means the different experiments using different training samples and test data were repeated randomly. Both experiment results showed the priority of the proposed method RAM whereas the scores of the Rouge-L values were higher than the conventional method ABS [3].

Table 1　Example of training data

No.1	Input	5th ステージというのが、本当直前の直前。の、お打合せに なります。言ってみれば結婚式の、準備ということなのでリハーサルをしたりとか。ふうん。あとはお荷物。ですね。結婚式のお荷物って何だろうと思うんですけど、例えば、受付で 飾りたいようなグッズとか。うん。あの、披露宴で使いたいある方へのプレゼントを渡されるとか、そういったもののお 荷物は事前に預かりをさせて頂きます。一週間以内ぐらいの 間で、にはなってくるので、まあ、12 月の 10 じゅう、14 ですね。じゅう、じゅう、14。14.14 日のお日にちの、だいたいこの、2 週目ぐらいですかね。12 月入ったら、その準備が始まる、っていう風に思ってもらっ たらいいので。まあ、そういった、あの、その具合で、やっていこうかなと、思いますが。
	Output (teacher signal)	5th ステージは直前のお打合せ。リハーサル。荷物を事前に お預かり。受付で飾るグッズ、披露宴でプレゼント。お預かりは 1 週間以内ぐらいの間で。12 月に入ったら準備が始まる 。
No.2	Input	ええと、3 つ目。3 rd ステージなんですけれども。ここでは、実際の結婚式でまあ、どんな感じでしていくか。の、あの、まあ演出であったりとか、あとどういったアイテムを使って いきたいか、とか、そんなお話をしていくので。先ほどお伝えたそのテーマとか、そういったものに合わせてご提案ができればなと思ってます。だいたい時期の目安としては、だいたい 8 月から 9 月この辺りぐらいでの、サードステージができればなと、思っておりますので。うんうん。だいたい、じゃあそちらはお休みの土曜日日曜日と、まあ、毎回なお打合せ来るとなると大変だと思いますので、ちょっとそちらの予定も合わせてやっていこうかなと思います。大体 8 月 9 月ぐらい目安と思ってください。はい。
	Output (teacher signal)	3rd ステージは実際の結婚式でどんな感じでしていくか、演出、アイテムのお話をしていく。テーマと合わせてご提案。時期 の目安は 8 月から 9 月。

Table 2 The output of RAM during training using data No.1

Epoch	Output of RAM
50	3 。 <eos>
100	3 提案 。 時期 の から に お盆 。 に 入っ たら たら が が 。 。 。 。 。 。 月 月 月 お 打合せ も 打合せ も <eos>
150	3 合わせ テーマ と 合わせ て ご 提案 。 目安 は 8 月 から に 入っ たら たら 3 3 。 <eos>
200	3 て 合わせ て ご 提案 。 提案 。 8 月 から 9 月 から 9 月 。 <eos>
250	3 を 手助け する 資料 も ある 。の ご 報告 と 一緒 に 聞く と 良い 。 お盆 帰る とき に 渡し て も いい か 。 大丈夫 。 <eos>
300	3 設備 が 入場 の 演出 、 アイテム の お話 を し て いく 。 テーマ と 合わせ て ご 提案 。 時期 の 目安 は 8 月 から 9 月 。<eos>
350	3 rd ステージ が 入場 の 演出 、アイテム の お話 を し て いく 。テーマ と 合わせ て ご 提案 。 時期 の 目安 は 8 月 から 9 月 。<eos>
400	3 rd ステージ が 入場 の 演出 、アイテム の お話 を し て いく 。テーマ と 合わせ て ご 提案 。 時期 の 目安 は 8 月 から 9 月 。<eos>
450	3 rd ステージ が 入場 の 演出 、アイテム の お話 を し て いく 。テーマ と 合わせ て ご 提案 。 時期 の 目安 は 8 月 から 9 月 。<eos>
500	3 rd ステージ が 入場 の 演出 、 アイテム の お話 を し て いく 。テーマ と 合わせ て ご 提案 。 時期 の 目安 は 8 月 から 9 月 。<eos>

Table 3 The output of RAM during training using data No.2

Epoch	Output of RAM
50	5 。 <eos>
100	5 。 。 の の に 。 。 に に に に 趣味 も も <eos>
150	5 大丈夫 。 合わせ 合わせ <eos> 5 8 月 に 入っ たら 準備 が 始まる 。 <eos>
200	5 て ご 提案 。 時期 の で 。 12 月 に 入っ たら 準備 が 始まる 。 <eos>
250	5 合わせ て 7 月 ぐらい 。 月 から お出し する の に 入っ たら 準備 が 始まる 。 <eos>
300	5 th ステージ は 直前 の お 打合せ 。 披露宴 で プレゼント 。受付 で 飾る グッズ 、 お 預かり は 1 週間 以内 ぐらい の 間 で 。 12 月 に 入っ たら 準備 が 始まる。<eos>
350	5 th ステージ は 直前 の お 打合せ 。 披露宴 で プレゼント 。受付 で 飾る グッズ 、 お 預かり は 1 週間 以内 ぐらい の 間 で 。 12 月 に 入っ たら 準備 が 始まる。<eos>
400	5 th ステージ は 直前 の お 打合せ 。 披露宴 で プレゼント 。受付 で 飾る グッズ 、 お 預かり は 1 週間 以内 ぐらい の 間 で 。 12 月 に 入っ たら 準備 が 始まる。<eos>
450	5 th ステージ は 直前 の お 打合せ 。 披露宴 で プレゼント 。受付 で 飾る グッズ 、 お 預かり は 1 週間 以内 ぐらい の 間 で 。 12 月 に 入っ たら 準備 が 始まる。<eos>

Table 4 Rouge-L scores by different methods in two experiments.

	Experiment No. 1	Experiment No. 2	Average value
ABS [3]	0.785	0.746	0.766
RAM (proposed)	**0.815**	**0.826**	**0.821**

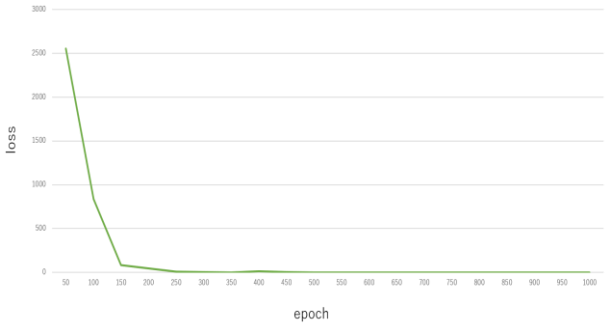

Fig. 3 The convergence of loss function of ABS [3] during learning.

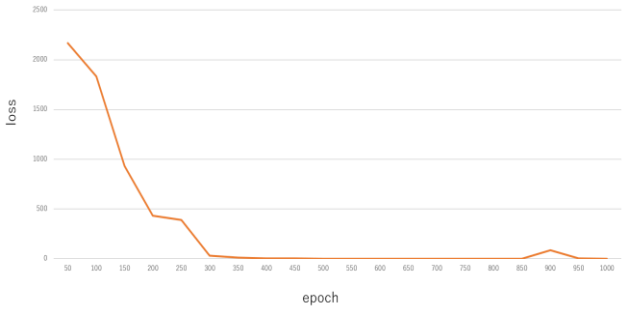

Fig. 4 The convergence of loss function of the proposed method RAM during learning.

4. Conclusions

To improve the summarization precision, the recursive attention model (RAM) was proposed in this paper. The modification of the proposed model was based on the attention-based summarization system (ABS) which adopted attention function to neural network language model (NNLM), and recurrent structure was added to ABS in the proposed RAM to address the related words in the input sentence and historic output of the model. Experiment results using Japanese conversation records showed the effectiveness of the proposed method.

As ABS has been improved by adopting recurrent neural networks to the encoder [6] [7], and hierarchical attention summarization methods were proposed recently [8], it is interesting to compare the performance of these different summarization models in the future.

Acknowledgement

We would like to thank Mr. Akira Kawakami and the anonymous reviewers for their valuable suggestions to this study, and Nihon Ceremony Ltd. for financial support. We also would like to acknowledge the early works of this study with Mr. Takuji Tsuruda. Apart of this study was supported by JSPS-KAKENHI No.19K12120.

References:

[1] C. W. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries”, Proceeding of the ACL-04 workshop, Vol.8, pp.74-81, 2004.

[2] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Kludanpur, “Recurrent neural network based language model”, Conference of International Speech Communication Association, pp.1045-1048, 2010.

[3] M. Rush, S. Chopra, and J. Weston, “A Neural Attention Model for Abstractive Sentence Summarization”, In Proceedings of Empirical Methods in Natural Language Processing 2015, pp. 379-389, 2015.

[4] P. Ganesh, S. Dingliwal, “Abstractive summarization of spoken and written conversation”, arXiv: 1902.01615 [cs.CL], 2019.

[5] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, “A neural probabilistic language model”, The Journal of Machine Learning Research, vol. 3, pp.1137–1155, 2003.

[6] S. Chopra, M. Auli, A. M. Rush, “Abstractive Sentence Summarization with Attentive Recurrent Neural Networks”, In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies (NAACL-HLT 2016), pp. 93-98, 2016.

[7] T. Kuremoto, T. Tsuruda, M. Obayashi, S. Mabu, “A Sentence Summarizer using Recurrent Neural Network and Attention-Based Encoder”, Proceedings of 2017 International Conference on Applied Mathematics, Modeling and Simulation (AMMS 2017), pp. 245-248, 2017.

[8] T. Kuremoto, T. Tsuruda, S. Mabu, “Summarizing Articles into Sentences by Hierarchical Attention Model and RNN Language Model”, Proceedings of The 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI 2019), pp. 1-6, 2019.

[9] D. P. Kingma, J. L. Ba, “Adam: A Method for Stochastic Optimization”, arXiv:1412.6980v9, 2017.