Generating Responses based on Information Visually-Induced by Text Utterance

Yoichi Ishibashi
Nara Institute of Science and Technology
ishibashi.yoichi.ir3@is.naist.jp

Hisashi Miyamori
Kyoto Sangyo University
miya@cse.kyoto-su.ac.jp

Abstract

In research on Neural Machine Translation, there have been studies that generate translated sentences by using both images and sentences, the results of which indicate that visual information improves translation performance. However, it is not possible to use sentence generation models (for dialogue systems) using images, as many text-based dialogue systems only accept text input. Herein, we propose an Associative Conversation Model that generates visually-induced feature vectors from textual vectors and uses it for generating sentences to utilize visually-associated information in a dialogue system without image input. Experimental results show that the proposed model improves context-dependent and informative scores by generating image feature vectors related to sentences.

1 Introduction

In research regarding Neural Machine Translation, the encoder–decoder model has been proposed (Sutskever et al., 2014) . It consists of an encoder that encodes the input text into a textual vector and a decoder that generates sentences by using the vector. Vinyals and Le (2015) showed that it is possible to extract knowledge and to conduct conversation by learning dialogue pairs with the model. For example, Vinyals and Le (2015) reported that, when asked who Skywalker is, their conversation model (*NCM*) responded, "*he is a hero.*"

NCM has a problem that it is not possible to generate properly to the input utterances that may require visual information for answering. For example, Vinyals and Le (2015) reported that, when asked how many legs a spider has, NCM responded, "*three, i think.*" However, images or videos may contain more detailed information than text. We thought that if such detailed image features could be extracted from the image, more specific and useful texts could be generated, including which cannot be obtained using text alone. In recent years, there have been studies in which translated sentences are generated by adding image features to the textual vector encoded by the encoder (Calixto et al., 2017) (Elliott and Kádár, 2017) (Nakayama and Nishida, 2017) (Saha et al., 2016) (Toyama et al., 2016). These studies showed that visual features work effectively for generating translation. Meanwhile, visual features are not considered in many text-based dialogue systems, because only the utterance text is given as input. How can visual features be used without accepting images as the input to the dialogue system?

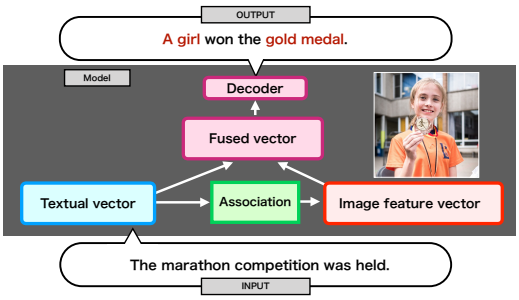

Figure 1: Generating a response by visual association. The textual vector is used to induce the visually-corresponding feature vector, and a response text is generated using the vector obtained by fusing the textual and visually-induced feature vectors.

Based on the discussion above, we propose an Associative Conversation Model that induces the visually-corresponding feature vector from the input text and generates a response using both the textual and visually-induced feature vectors. The contribution of this research is as follows:

- We made it possible to generate responses using the visually-induced feature vectors, without direct image input.

- Our proposed model can generate response texts, including richer and more useful information, by inducing visually-corresponding feature vectors from the input text vector.

2 Related Work

In recent years, encoder–decoder models which generate sentences from multi-modal inputs have been investigated (Calixto et al., 2017) (Elliott and Kádár, 2017) (Nakayama and Nishida, 2017) (Saha et al., 2016) (Toyama et al., 2016) (Hori et al., 2019). For example, Elliott and Kádár (2017) showed that machine translation performance is improved by multitask learning models that use text and image features without inputting images. However, these studies mainly examined machine translation, rather than dialog tasks. Hori et al. (2019) generated responses in a dialog by using text and video. However, their model required image inputs and audio features. We use video for training our model but use only text for inference.

3 Associative Conversation Model

Figure 1 shows the overview of our model. We use video for training the model. The purpose of this research is to generate a response that has richer information by inducing the visually-corresponding vector from a textual vector. An important point to consider is how to induce visually-corresponding features from textual vectors. We would like to eventually train a network that has only text input and uses association, such as in figure 2b. This associative encoder in figure 2b needs to be trained in advance. Therefore, our method has three learning steps.

- Step1: We train the step1 model as illustrated in figure2a. This model has an LSTM textual encoder and an LSTM video encoder. VGG16 (Simonyan and Zisserman, 2014) is used for extracting image features X^{vis}. After training, we extract sequences of textual vectors C_t^{txt} and sequences of image feature vectors C_t^{vis} by using the pre-trained step1 model for training the associative encoder in step2. In this step, inputs are texts and images, both encoder and decoder use LSTMs, and we use attention (Bahdanau et al., 2014). Fusion layer fuses both C_t^{txt} and C_t^{vis} by using MLP.

- Step2: We train an associative encoder that predicts sequences of image features C^{vis} from those of textual vectors C^{txt}. The loss function for the associative encoder is given by

$$L = \frac{1}{T}\sum_{i=0}^{T}(C_i^{vis} - RNN(C_i^{txt}))^2 \quad (1)$$

  T represents the length of the output text at step1 ($Y = (y_1, ..., y_T)$).

- Step3: We train an associative conversation model (Fig.2b) that has the pre-trained associative encoder.

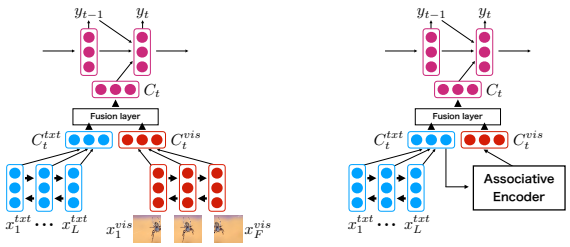

Figure 2: **(a)** Step1 model : A model that performs prior learning for extracting context vectors C^{txt} and C^{vis}. After this model learns C^{txt} and C^{vis} from video $X_{vis} = (x_1^{vis}, ..., x_F^{vis})$ and text $X_{txt} = (x_1^{txt}, ..., x_L^{txt})$, we extract C^{txt} and C^{vis} using this model. **(b)** Step3 model (Associative Conversation Model): This model learns to generate responses by using the visually-induced feature vectors instead of inputting images

4 Dataset

There are many datasets combining image and text for dialogue or question answering tasks (Antol et al., 2015) (Das et al., 2017) (Yagcioglu et al., 2018) (Suhr et al., 2017) (Bigham et al., 2010) (Johnson et al., 2017) (Saha et al., 2017) (Hudson and Manning, 2019). Although they use images, videos contain richer information than images.

In this research, we created two datasets combining dialogues and videos: a TV drama and a TV news dataset. Each frame was used as input to the pre-trained convolutional neural network (VGG16 (Simonyan and Zisserman, 2014)), and the output of the last pooling layer was used as the images features. The recorded programs used were 510 Japanese TV dramas for the Drama dataset, and 163 Japanese TV news episodes for the News dataset. The size of the TV Drama

dataset is 7.2 Million dialogues and the corresponding videos have 50 thousand vocabulary words. The size of the TV News dataset is 1.5 Million dialogues and the corresponding videos have 19 thousand vocabulary words.

5 Experiments

5.1 Quantitative Evaluation

To analyze the effect of association, we presented response texts generated by our model (*seq2seq+assc*) and the baseline model (*seq2seq*) to six different study participants and asked them to judge the score of the generated response for the input utterance. As the evaluation method, the method in NTCIR 13 STC-2 Japanese subtask was used (Shang et al., 2017). To evaluate the generated texts, a score of 0, 1, or 2 was given to each of the following four criteria: fluency, coherence, context-dependence, and informativeness.

The labels L0, L1, and L2 are given by the procedure called Rule-1 in NTCIR 13 STC-2 Japanese subtask (Shang et al., 2017). The procedure is given in listing 1.

Listing 1: Rule-1

```
IF fluent & coherent = 1
   IF context-dependent & informative = 2
      THEN L2
   ELSE L1
ELSE
   L0
```

As with Shang et al. (2017), Accuracy $Acc_G@k$ was calculated based on the following equation.

$$Acc_G@k = \frac{1}{nk} \sum_{r=1}^{k} \sum_{i=1}^{n} \delta(l_i(r) \in G) \qquad (2)$$

$l_i(r)$ is the label assigned to the r th response candidate for the i th utterance. n is the number of labels assigned by evaluators to one response ($n = 7$). G is the set of labels regarded as "correct" ($G = \{L2\}$ or $G = \{L2, L1\}$). k is the number of candidate responses per utterance. In this experiment, as the model generates one response per input utterance, $k = 1$. Therefore, $Acc_{L2}@1$ is the average number of $L2$ labels given to the first response. This implies that if $Acc_{L2}@1$ is high, the model can generate context-dependent and informative responses. Here, 50 utterances of news subtitles and dialogue texts, including questions that ask general facts, were used as the evaluation data.

Table 1: Results of Human Evaluation (Drama)

Models	TV Drama data	
	Mean $Acc_{L2}@1$	Mean $Acc_{L1,L2}@1$
seq2seq	**0.01**	**0.251**
ours	0.00	0.180

Table 2: Results of Human Evaluation (News)

Models	TV News data	
	Mean $Acc_{L2}@1$	Mean $Acc_{L1,L2}@1$
seq2seq	0.066	0.429
ours	**0.109**	**0.503**

Table 1 and 2 show the evaluation results. First, we trained models using the TV Drama dataset. However, both $Acc_{L2}@1$ and $Acc_{L1,L2}@1$ of our model was lower than those of the baseline model, expressed as *seq2seq*. We speculated that this is because the objects in the video are not closely related to the utterance texts themselves in the TV Drama dataset. Therefore, we created the TV News dataset in which the objects and the utterance texts are considered to be much more related to each other. Both the proposed model and the baseline model (*seq2seq*) had higher accuracy for the news data. The result shows that our model had higher accuracy for both Mean $Acc_{L2}@1$ and Mean $Acc_{L2}@1$ compared to the baseline when using the TV News texts. This means that the proposed model generates informative responses that have context dependency because the objects are more related to the utterance texts in News data than in Drama data.

Input	The University Entrance exam will be held on 14th and 15th.	Well, today is All Japan Figure Skating Championships.
Output by Baseline	There will be a large-scale fire that is also in western Japan and eastern Japan.	Aiming for four consecutive championships in the women's singles, athletes of the Japanese championships participated in the tournament.
Output by ACM	It is highly expected to be **snowy** and windy.	A player who has won the **gold medal** in women's singles.
Image similar to associated feature vector		
Generated words	**Snowy**	**Gold medal**
Cos similarity	**0.333**	**0.344**

Figure 3: Example of comparison results on validity of text generation by visual association[1)]

[1)] Source: Image on the left: "NHK News 7" broadcast on

Input	The Grand Sumo Tournament is in the second day.	As for pitchers, three players including Otani have been selected from Nippon Ham.
Output by Baseline	No. 1 is No. 1 in 1 meter.	Baseball is out in the professional this season.
Output by ACM	Today, **Yokozuna** Hakuho will aim for the first victory.	This is an **athlete**.
Image similar to associated feature vector	**Sumo wrestler** 	**Pitcher** 
Generated words	**Yokozuna** (The highest rank wrestler in sumo)	**Athlete**
Cos similarity	**0.336**	**0.297**

Figure 4: Topic : Sports [2)]

Input	The vicinity of Honshu is expected to be covered widely by mobile high pressure.	In tomorrow morning, it will be isolated snowstorms around the Japan Sea side of western Japan and eastern Japan, due to the developing low pressure.	
Output by Baseline	Let's move on to the weather around the country on tomorrow.	Vigilance is necessary for windstorms and high waves.	
Output by ACM	Tomorrow morning, it will be **sunny** in many places from western Japan to eastern Japan, and the side on the Japan Sea of western Japan.	Please also be aware of the influence on **traffic** caused by **windstorms**, heavy blizzards, and snowdrift.	
Image similar to associated feature vector	**Honshu is covered widely by high pressures** 	**A person riding a bicycle likely to fall over due to a strong wind** 	
Generated words	**Sunny**	**Traffic**	**Windstorms**
Cos similarity	**0.749**	**0.338**	**0.280**

Figure 5: Topic : Weather [3)]

5.2 Qualitative Evaluation

Figures 3 , 4 and 5 show the texts generated by our model (or by the baseline). The images in those figures are the nearest neighbor images obtained from the visually-induced feature vector generated by associative encoder. These results show that association works effectively to generate texts with more specific information. For example, in the example on the left of figure 3, the proposed model generated a specific weather forecast with the word "*snowy*" for the input text "*The University Entrance exam will be held on 14th and 15th.*" Note that snow actually fell on the day of the entrance examination, and that the images showing that it was snowing at the venue of the exam were included in the training data. The important point here is that the word "*snowy*" cannot be easily generated from the input texts alone, but is a word that can be generated for the first time in association with the image of snow. However, the result generated by the baseline is "*There will be a large-scale fire that is also in western Japan and eastern Japan*" and contains erroneous information such as "*fire*". In the example on the left of figure 5, the proposed model induced the visually-corresponding features showing the placement of *high pressures* from the input phrase "*high pressure*", and the word "*sunny*" was generated from the visually-induced features. These results show that our association mechanism can induce the visually-corresponding feature vectors from the textual vectors and infer appropriate responses.

Other interesting examples are shown in figures 4 and 5. In the example on the left of figure 5, the proposed model associated an image showing the placement of *high pressures* from the input phrase "*high pressure*", and the word "*sunny*" was generated from the associated image. This is a good association because high pressure makes the weather sunny. These results show that our association mechanism can infer an image feature vector from a sentence vector; for instance, our model generates '*sunny*" from a generated image feature vector of a high pressure image.

However, our method still had several problems. In some examples, it was clear that the visually-induced feature vector was associated with a different topic from the input text. For example, a scene where a speed skating player won the gold medal was associated with figure skating 3.

6 Conclusions

Herein, we proposed an Associative Conversation Model that induces the visually-corresponding feature vectors from the input text and generates responses using both the textual and visually-induced feature vectors. Experimental results show that the proposed model produces the visually-induced feature vectors related to the input texts and can generate responses containing richer and more useful information.

NHK on 11th January 2017, Image on the right: "NHK News 7" broadcast on NHK on 23rd February 2017

[2)] Source: Image on the left: "NHK News 7" broadcast on NHK on 13th January 2017, Image on the right: "News Watch 9" broadcast on NHK on 17th February 2017

[3)] Source: Image on the left: "News Watch 9" broadcast on NHK on 27th February 2017, Image on the right: "NEWS CHECK 11" broadcast on NHK on 20th February 2017

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342. ACM.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. 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 1913–1924.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. *CoRR*, abs/1705.04350.

Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. 2019. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356. IEEE.

Drew A Hudson and Christopher D Manning. 2019. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Hideki Nakayama and Noriki Nishida. 2017. Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot. *Machine Translation*, 31(1-2):49–64.

Amrita Saha, Mitesh M. Khapra, Sarath Chandar, Janarthanan Rajendran, and Kyunghyun Cho. 2016. A correlational encoder decoder architecture for pivot based sequence generation. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 109–118.

Amrita Saha, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2017. Towards building large scale multimodal domain-aware conversation systems.

Lifeng Shang, Tetsuya Sakai, Hang Li, Ryuichiro Higashinaka, Yusuke Miyao, Yuki Arase, and Masako Nomoto. 2017. Overview of the NTCIR-13 short text conversation task. In *Proceedings of NTCIR-13*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Joji Toyama, Masanori Misono, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. 2016. Neural machine translation with latent semantic of image and text. *CoRR*, abs/1611.08459.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv preprint arXiv:1809.00812*.