ASSOCIATIVE CONVERSATION MODEL: GENERATING VISUAL INFORMATION FROM LANGUAGE INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we propose an Associative Conversation Model that generates visual information from language information and uses it for generating sentences in order to utilize visual information in a dialog system without image input. In research on Neural Machine Translation, there are studies that generate translated sentences using both images and sentences. In these studies, it is shown that visual information is useful for generating translated sentences. However, in many dialog systems, visual information was not considered and only text input was accepted. Therefore, when applying a sentence generation algorithm using images to a conversation system, it is not possible to input visual information such as images. Our approach generates (associates) visual information from input sentence and generates response sentence using context vector fusing associative visual information and sentence language information. In comparative experiments with models without association, we found that the proposed method was able to generate useful sentences by associating visual information related to sentences.

1 INTRODUCTION

In human conversation, communication is carried out on the premise that each other has knowledge of a common language. This is the same when humans use the dialog system, and we conduct dialog on the premise that the dialog system has common sense. Therefore, it is essential that the dialog system has common sense. However, it is nearly impossible for humans to organize and describe a huge amount of common sense considering the time and cost aspect. In the case of a human being, it is possible to acquire the knowledge through conversation with a person. Therefore, it is desirable for dialog systems to acquire knowledge through conversation as well. This has the advantage that human beings need not explicitly give knowledge to the system.

Thus, we are working on a task where a model extracts knowledge from human conversation and generates response sentences. There is the Encoder-Decoder model that can extract knowledge from a conversation (Sutskever et al., 2014). The structure of the Encoder-Decoder model consists of an Encoder that encodes input information into a context vector and a Decoder that generates sentences using the context vector . Vinyals & Le (2015) showed that it is possible to conduct conversation by giving pairs of dialogs to the Encoder-Decoder model and extracting commonsense knowledge by learning.

However, their conversation model (Neural Conversational Model) has a problem that it is not possible to respond properly to input sentences that require visual information. For example, they reported NCM responded “three, i think.” to their question “how many legs does a spider have?”. Therefore, this paper addresses the task of adding visual knowledge to the response sentence by using the visual information related to the input sentence when the conversation model generates the sentence. There is research that generated translated text by adding an image feature amount to context vector encoded by Encoder-Decoder model (Toyama et al., 2016) (Calixto et al., 2017) (Elliott & Kádár, 2017) (Nakayama & Nishida, 2017) (Saha et al., 2016). These studies showed that visual information works effectively for generating translation.

However, many practical dialog systems, visual information was not considered, and the input was often text only. Therefore, when applying a sentence generation algorithm with image input to

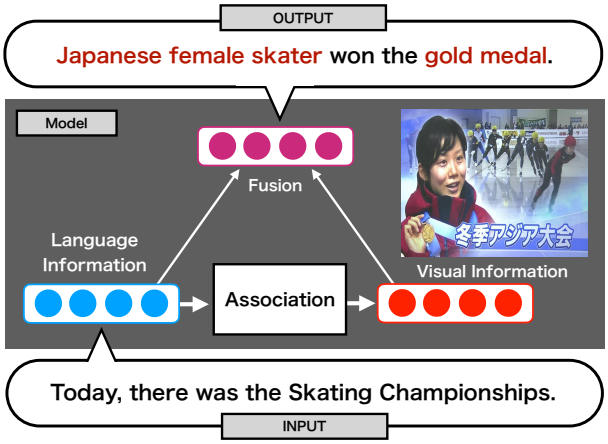


Figure 1: Generating a response by association. Generates corresponding visual information from the language information, and generates a response sentence using a vector obtained by fusing two pieces of information.

conversation, it was impossible to input visual information such as images. Then, how can the visual information be used without the dialog model accepting visual information for input?

In order to solve these problems, we propose an Associative Conversation Model for generating response sentences using visual information associated with sentence language information. In our proposed method, we attempted to generate response sentences using visual information without inputting visual information. The contribution of this research is as follows:

- We made it possible to generate visual information related to sentence language information through end-to-end learning of dialog.
- We made it possible to generate sentences using visual information without directly inputting visual information by association.
- Our proposed model can generate response sentences including useful information compared with a model without association by associating visual information related to input sentence.

2 Associative Conversation Model

2.1 Overview of Architecture

Figure 1 shows our overall model. The purpose of this study is to show that the Decoder appropriately refers to the context vector based on visual information as necessary, It is possible to generate a more appropriate response sentence which cannot be obtained only by language information. However, in this task, the information given as input is only language information (sentences). Therefore, we aimed to solve this problem and the spider's foot problem by taking the following approaches.

- First, visual association is performed from the language information of the input sentence, and visual information corresponding to the language information is generated.
- Next, by fusing visual information associated with language information, we obtain information reflecting both as necessary.
- Finally, a response sentence is generated based on the fused information.

Therefore, we introduced the Encoder that uses language information to generate a visual context vector . The structure of Associative Conversation Model consists of Associative Encoder in addition to Encoder and Decoder with Attention Mechanism. We used LSTM for Encoder, Decoder, and Associative Encoder (Hochreiter & Schmidhuber, 1997). Input to the Associative Conversation Model is only sentences, and output is response sentences to input sentences. Fusing the context

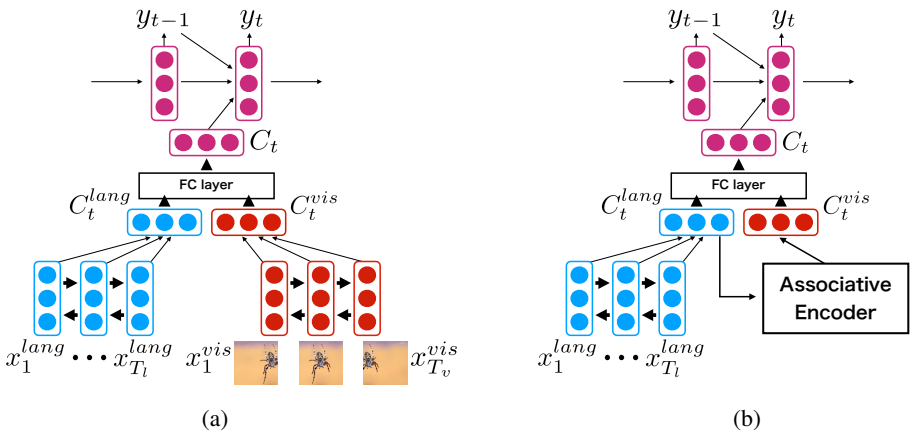


Figure 2: (a) Step1 : A model that performs preliminary learning for context vector extraction.
(b) Step3 : Associative Conversation Model.

vector of the visual information associated with the input sentence and the context vector of the input sentence, and inputting it into the decoder, the response sentence is generated. We call the process of generating the corresponding visual information from the meaning of the sentence associative, and call the generated visual information as associative representation. The reason to call associative representation rather than context vector is to distinguish it from visual information obtained from actual images by preliminary learning (see Sec. 2.2). It is possible to respond using visual information without associating an image etc. by association.

2.2 Learning method

Learning consists of the following three steps. **Step1 and Step2 are the learning of two models necessary for learning Associative Conversation Models in Step3.**

Step 1 : Extraction of context vectors between language information and visual information
Fig. 2a is the network of models used in this step. In this step, use a multimodal Encoder-Decoder model consisting of a sentence Encoder, a video Encoder, and a sentence Decoder including Attention(Bahdanau et al., 2014). The model of Step 1 encodes a video and a sentence with each Encoders, and fuse the context vector of the sentence and the video to predict response sentences. Then, using this learned model, correspondence between language information and visual information is extracted (Fig. 2a).

Step 2 : Learning association
In this step, use an RNN. The model of Step 2 associates visual information from language information. In other words, the model learns predicting video context vector (associative representation) from the sentence context vector.

Step 3 : Generating responses using association
Fig. 2b is the network of models used in this step. In this step, we use a multimodal Encoder-Decoder model using associative Encoder instead of video Encoder. In other words, the network architecture in Step3 is a Decoder of a sentence including a sentence Encoder, an Associative Encoder, and Attention. This is Associative Conversation Model. Association is performed using an Associative Encoder from language information of input sentence to generate associative representation. Then, it combines language information and associative representation, and generates response sentence (Fig. 2b).

There are two pre-train steps before the learning Associative Conversation Model. One is the Learning Associative Encoder. The Learning Associative Encoder predicts the context vector of an image from the sentence context vector by using the context vector of the sentence extracted in advance and the video context vector. The other is the extraction of context vectors of sentence and video. We used a multimodal Encoder-Decoder model which performs end-to-end learning with two Encoders

and one Decoder. The input data is a sentence and a video (a sequence of images). Enter a sentence in the sentence Encoder and enter the image sequence in the video Encoder. Then encode the sentence by Attention Mechanism and obtain the context vector of the sentence (Bahdanau et al., 2014). In the same way, we encode the sequence of images and obtain the video context vector. Then, we fuse these context vectors, input them into the Decoder, and generate sentences. After learning this, the model extracts the context vector of the sentence and the video context vector and uses it for the learning Associative Encoder.

2.2.1 STEP1 : EXTRACTION OF CONTEXT VECTORS BETWEEN LANGUAGE INFORMATION AND VISUAL INFORMATION

We are challenging the model to extract knowledge from human utterances and videos at that time. The important point here is that human utterances and videos at that time may not correspond to each other. For example, when we are talking about a spider, we talk in situations where a spider are not in front of us. Therefore, we did not use data with correspondence between visual information and language information in advance (e.g., MS COCO) (Lin et al., 2014). Instead, we used the following data which **may NOT** correspond to language information and visual information.

- Subtitle and the video of the scene where the subtitle was displayed

Language information has expressions representing actions. Therefore we used video rather than image. For details on data sets, see Sec. 3 .

The inputs to the model in Step1 are a video (sequence of images) and a sentence. The output is a response sentence to the input sentence. In addition, we input image features, instead of using images directly. We use features obtained by pre-trained CNN (without fully connected layers). Therefore, the input to this preliminary learning model is a series of words $X_{lang} = (x_1^{lang}, ..., x_{T_l}^{lang})$ and the image feature sequence $X_{vis} = (x_1^{vis}, ..., x_{T_v}^{vis})$. In addition, the output is the sequence of words $Y = (y_1, ..., y_T)$. Here, T_l, T_v and T represent the length of the input sentence, the number of input images and the length of the output sentence, respectively.

The network architecture of the model of Step1 is a multimodal Encoder-Decoder model consisting of a sentence Encoder, a video Encoder, and a Decoder with Attention. In Step 1, the sentence Encoder encodes the input sentence, the video Encoder encodes the input video, and each Encoders obtain context vectors C_t^{lang} and C_t^{vis}. Then, C_t which fused context vector is input to the Decoder to generate a response word.

$$s_t = LSTM(C_t, s_{t-1}, y_{t-1}) \tag{1}$$

$$P(y|s_{t-1}, C_t) = softmax(W_s s_{t-1} + W_c C_t + b_s) \tag{2}$$

For extracting context, we use Attention mechanism (Fig. 2a) (Bahdanau et al., 2014). In normal seq2seq (Sutskever et al., 2014), the context vector is compressed to a fixed length vector, but the attention mechanism solves it. Furthermore, in the Attention mechanism, it is possible to extract context vectors strongly reflecting the information of a part of the sequence corresponding to the word of the response sentence. An Attention Mechanism is a point where you can associate a certain scene in a video with words of an input sentence. It is suitable for learning the Associative Conversation Model in two points that can record a pair of sentence context vector and video context vector as the training data for Associative Encoder. The Attention mechanism receives the input series $X = (x_1, ..., x_T)$ and calculates the context vector C_t^{att}. h_i is the intermediate vector of bidirectional LSTM computed by receiving the input sequence $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$.

$$C_t^{att} = \sum_{i=0}^{T} \alpha_{t,i} h_i \tag{3}$$

$$\alpha_{t_i} = softmax(e_{t_i}) \tag{4}$$

$$e_{t_i} = w_e^T tanh(W_e s_{t-1} + V_e h_i + b_e) \tag{5}$$

On the other hand, the multimodal Encoder-Decoder model acquires two context vectors C_t^{lang} and C_t^{vis}. Therefore, the multimodal Encoder-Decoder model uses the following equations instead of

Eq.(3) (4) and (5).

$$C_t^{lang} = \sum_{i=0}^{T_{lang}} \alpha_{t,i}^{lang} h_i^{lang} \tag{6}$$

$$C_t^{vis} = \sum_{j=0}^{T_{vis}} \alpha_{t_j}^{vis} h_j^{vis} \tag{7}$$

$$\alpha_{t_i}^{lang} = softmax(e_{t_i}^{lang}) \tag{8}$$

$$\alpha_{t_j}^{vis} = softmax(e_{t_j}^{vis}) \tag{9}$$

$$e_{t_i}^{lang} = w_{e_{lang}}^T tanh(W_{e_{lang}} s_{t-1} + V_{e_{lang}} h_i^{lang} + b_{e_{lang}}) \tag{10}$$

$$e_{t_i}^{vis} = w_{e_{vis}}^T tanh(W_{e_{vis}} s_{t-1} + V_{e_{vis}} h_j^{vis} + b_{e_{vis}}) \tag{11}$$

There are also responses that do not require visual information in the conversation. Therefore, the decoder side needs to have a "scale" on how strongly to refer to sentence information (sentence context) or visual information (video context vector). By multiplying by the weight matrix it is possible to learn the reference condition of the context vector C_t^{vis} and the context vector C_t^{lang} (Eq. 12). Therefore, the context vector eventually passed to Decoder is a vector having a reference condition of sentence context vector (language information) and video context vector (visual information).

$$C_t = W_{lang} C_t^{lang} + W_{vis} C_t^{vis} + b_c \tag{12}$$

2.2.2 STEP2 : LEARNING ASSOCIATION

In Step2, the Associative Encoder learns correspondence between visual information and language information. We thought that learning can extract the most probable visual information for the context vector of sentences. In other words, Associative Encoder expects to acquire common sense association of language and vision (e.g., "apple" word and apple visual information).

In addition, the Associative Encoder decides the next video context vector by referring to past context vectors in the sentence. We present and explain examples. The visual information (video context vector) that corresponds to the context vector of the word "squeeze" is diverse. For example, the phrases "squeeze a lemon", "squeeze a person's hand", "squeeze toothpaste out" would be considered. In this example, there are phrases that explain concretely "squeeze" in the sentence. For example, "lemon", "person's hand", and "toothpaste". In this paper, the Association is to generate corresponding visual information from language information. Therefore, there is a demand to associate as to what is. Here, it is assumed that a context vector "squeeze" is obtained by the Attention Mechanism. If the context vector "lemon" was obtained before this, it would be possible to generate visual information of "squeeze a lemon" from the context vector "lemon" and "squeeze". This is a temporal feature. Therefore, RNN is used for the Associative Encoder. Therefore, the Associative Encoder is a regression model with the input sequence of context vectors of word $C^{lang} = (C_1^{lang}, ..., C_t^{lang}, ..., C_T^{lang})$ and output sequence of context vectors of image $C^{vis} = (C_1^{vis}, ..., C_t^{vis}, ..., C_T^{vis})$ (Eq. 13). Here, T is the length of the output sentence. We used LSTM which is capable of learning long-term dependence.

$$C_t^{vis} = AssociativeEncoder(C_t^{lang}) \tag{13}$$

2.2.3 STEP3 : GENERATING RESPONSES USING ASSOCIATION

In Step 3, the Associative Conversation Model learns response sentence using association instead of video(Fig. 2b). The input to the Associative Conversation Model is the sequence of words (sentence) $X_{lang} = (x_1^{lang}, ..., x_{T_l}^{lang})$. The output is the sequence of words (sentence) $Y = (y_1, ..., y_T)$.

In the Associative Conversation Model, the video Encoder of the preliminary learning model of Step1 is changed to an Associative Encoder. Therefore, the architecture of the Associative Conversation Model is Eq.(7) in Step1 replaced by Eq.(13). In other words, we do not change the mechanism of blending the language and visual information that was done in Step1, predicting the video context vector from the context vector of the input sentence by the Associative Encoder. We

input associative representation generated by the Associative Encoder and sentence context vector into Decoder, learn the Decoder, the Attention, and a fully connected layer (Eq. 12). The weights of Decoder, Attention and fully connected layer are initialized, and the weight of the preliminary learning model of Step1 is not used. In the sentence Encoder and Associative Encoder, weight update is not performed and parameters of the preliminary learning model of Step1 are used. The reason for re-training the Decoder, the Attention and a fully connected layer is because the associative representation has a different property from the context vector of the animation generated in Step1. The associative representation is different from the video context vector generated in Step1. The Associative Encoder associates with reference to context vectors of past sentences. Therefore, Decoder use representations not obtained in Step1.

Furthermore, in order to associate with reference to the context vector of past sentences, we obtained the representations not obtained in Step1. We thought that by re-training parts other than encoding of context, it is possible to generate response sentences using associative visual common sense.

3 DATASET

We use the following data collected independently for the learning our model.

- Subtitle of TV news and video of the scene where the subtitle was displayed

In addition, assuming that the period between "。" (Japanese period which is placed at the end of a sentence) , " ! " and " ? " Is one sentence, it is divided into sentences by this delimiting method and counted as a dialog for each sentence. Therefore, the response sentence of a person is the next input sentence, and the image is cut out as an image at a frame rate of 5 fps. Next, the image was input to the learned Convolutional Neural Network, and the output of the last convolution layer was used as the image feature amount. We used VGG16 for the Convolutional Neural Network (Simonyan & Zisserman, 2014). The news is a Japanese television news 163 programs acquired in from December, 2016 to March, 2017. The number of dialog is 38K dialog and the vocabulary size is 19K words.

4 EXPERIMENTS

In this section, we describe the result of comparison between associative conversation model and model without association. We use the Encoder-Decoder model with attention that caused the pair of conversation sentences to be learned in the baseline. The input and output of the baseline model are only sentences. We learned both the proposed model and the baseline model with the same data and investigated the effect of association. In addition, we visualized associative visual information and investigated whether associative conversation model acquired visual common sense.

4.1 MODEL CONFIGURATION

In the configuration of the multimodal Encoder-Decoder model in Step 1, one layer of LSTM was used for two Encoders and Decoders. We set the hidden layer dimension number of Encoder to 512 dimensions, the dimension of fused context vector to 1024 dimensions, and the number of dimension of Decoder hidden layer to 1024 dimensions.

The associative Encoder in Step 2 used 4 LSTMs. The hidden layer dimension number was 1024 in all 4 layers, the optimization method was adam, the batch size was 64.

In the associative conversation model in Step 3, we set the hidden layer dimension number of Encode of the sentence to 512 dimension, the dimension of synthetic meaning expression to 1024 dimension, and the number of dimension of hidden layer of Decoder to 1024 dimension. Also, in step 3, weights of the Encoder of the sentence and biases are not updated by using the weights in Step 1. Likewise, associative Encoder does not update weights. In addition, weights and biases of Decoding, Attention and Combination Semantic Expression are not weighted in Step 1 and weights are initialized and learning is performed (See Eq.(1), Eq.(2), Eq.(10), Eq.(11), Eq.(12)) .

The baseline model consists of one Encoder and one Decoder with Attention. Baseline model does not include associative Encoder nor video Encoder. Both Encoder and Decoder used one layer of

Associated image		
Input	The vicinity of Honshu is expected to be covered widely by mobile high pressure.	Well, today is the figure skating All Japan Championship.
Output (ACM)	In the morning of tomorrow, there are many places to be **sunny** from West Japan, East Japan to the Japan Sea side of West Japan.	A player who has won the **gold medal** won three women's singles.
Output (Baseline)	Then, the weather in various places of tomorrow.	Aiming for 4 consecutive championships in the women's singles, athletes of the Japanese championships participated in the tournament.
Generated word	Sunny	Gold medal

Figure 3: This result shows that the visual information associated with the input sentence by the proposed method corresponds to the input sentence. In addition, we can see that association of them produces words with valuable information.

LSTM. We set Encoder's hidden layer dimension number to 512 dimension and Decoder hidden layer dimension number to 1024 dimension.

In Step 1 of the baseline and Step 1 and Step 3 of the proposed method, adagrad was used as the optimization method. The batch size was set to 64. In addition, I made a 512-dimensional embedding expression of a word and entered it into sentence Encoder. In step 1 of the proposed method, an image is input to the VGG 16, and the output of the last convolved layer is made a one-dimensional vector. This vector is 25088 dimensions. This dimension is compressed to 512 dimensions and entered into the video Encoder.

4.2 Results

The figure (3) shows the proposed method and the sentence generated by the baseline. This result shows that the proposed method generates more useful sentences with more specific information than the baseline. For example, for the input sentence "The vicinity of Honshu is expected to be covered widely by mobile high pressure.", The proposed method generates a specific weather forecast called "sunny". On the other hand, the baseline does not say "Then, the weather in various places of tomorrow." And concrete weather forecasts. In addition to this example, I also found an example where the suggested method with association produces useful sentences with more specific information than baselines without association. From these results, we speculated that association works effectively to generate sentences with concrete information.

4.3 Analysis of Association

Fig.(3) is the analysis result of association. The above image is an actual image with the highest degree of similarity to the associative representation generated by association of the proposed method from the input sentence. We input the video of the learning data (sequence of images) to the model of step 1 and we measured the meaning expression and the associative representation of the ani-

mation generated by the video Encoder with cosine similarity. Although it is indirect, the semantic representation of the movie with the highest similarity among them and the corresponding input image were made to be associative visual information. In other words, the image in Fig.(3) is visualized visual information associated with the sentence using the image in the learning data. From the result of Fig.(3) , you can see that the sentence and associated visual information are associated. For example, in the example on the left of Fig.(3) , the topic of "Japan's wide range is covered with high pressure" is reminiscent of "Japan covered with high pressure" It generates the word "sunny". Actually, if the high pressure is covered widely, the weather will be fine, so the output sentence of the proposed method contains useful information. Also, although the example on the right of Fig.(3) misunderstood figure skating as speed skating, the skating player has won the gold medal for the topic of "skate championship" We associate the scene and generate the word "gold medal" from the associative expression.

There is also a problem. In this research, there was a case that the input sentence was misunderstood and associated with visual information. For example, in the example to the right of Fig.(3) , it is reminiscent of a scene where skate player is acquiring gold against topics about figure skating. This will be improved by increasing the learning data. Also, because I did learning in the news sentence, I could not respond well to spoken conversation. As for the solution to this, it is conceivable to learn with a general dialogue sentence instead of a news sentence when relearning the model using association in step 3. That is, using the knowledge extracted by Encoder in steps 1 and 2, in step 3, relearning is performed using a general dialogue sentence. If knowledge can be extracted, learning can be performed more efficiently by performing step 3 only in accordance with the task (eg, QA).

5 Related Work

(Lin et al., 2014) assumed a conversation on social media about a shared image and proposed a model that performs natural conversation from images and sentences. We assume a case where we do not use images when we talk. In the sentence generation algorithm, there is an attempt to obtain a useful output sentence by simultaneously giving visual information and language information (Toyama et al., 2016) (Calixto et al., 2017) (Elliott & Kádár, 2017) (Nakayama & Nishida, 2017) (Saha et al., 2016). Approaches using visual information only during training are similar to our approach, with the approach of Toyama et al. (2016) and Elliott & Kádár (2017).

Toyama et al. (2016) proposed a model for modeling latent variables from text and images and generating sentences using the latent variables. In addition to EncoderDecoder with normal attention, their model has a mechanism to generate latent variables. When a sentence is input to Encoder and predict translation with Decoder, target word is generated by using latent variable generated from source sentence / target sentence / image as hidden layer of Decoder. Next, we incorporate the latent variable into the Decoder and finally output the target.

Elliott & Kádár (2017) shows that learning semantic representation using a multitasking learning framework with two types of output, sentence generation and image feature generation, is used to generate Decoder's translation sentence, thereby improving accuracy It was. The structure of their model consists of Encoder of sentence, Decoder of sentence, Decoder of image, input sentence and predict translation by Decoder of sentence. At the same time, hidden layers of Encoder are averaged and input to the Decoder of the image to predict image features.

There are three major differences between our approach and their approach. First, while their approach gives images as visual information during training, our approach extends it and gives animation. With this extension we believe that it will be possible for models to learn the correspondence between words representing motion and visual information. Secondly, we deal with dialog tasks, not translation tasks. Thirdly, we are assuming that we learn such as knowledge extraction from data which does not necessarily correspond to the sentence and the visual information at that time, such as general conversation. For this reason, our approach is equipped with a noise removal mechanism that learns the degree of usage of visual information and language information Eq. (12).

The other difference between (Elliott & Kádár, 2017) is that we do not adopt multitasking learning in our method and therefore input visual information. Therefore, since it is possible to attention to visual information during training, visual information can be extracted according to the response word.

6 CONCLUSIONS

In a study applying a sentence generation algorithm of a translation sentence to a conversation model, there was a problem that it was not possible to respond well to an input sentence which requires visual information. However, in many practical dialogue systems, there are many cases where there is no input of visual information. Therefore, our proposed method generates (associates) visual information from input sentences and generates response sentences using fused context vector with associative visual information and sentence language information. Comparative experiments with models that do not use association show that association of visual information related to input sentences produces response sentences that contain valuable information compared to models without association. Analysis of association also showed that our proposed method can generate visual information related to sentence language information through end-to-end learning of dialogue.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. *arXiv preprint arXiv:1702.01287*, 2017.

Desmond Elliott and Akos Kádár. Imagination improves multimodal translation. *arXiv:1705.04350*, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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

Amrita Saha, Mitesh M Khapra, Sarath Chandar, Janarthanan Rajendran, and Kyunghyun Cho. A correlational encoder decoder architecture for pivot based sequence generation. *arXiv preprint arXiv:1606.04754*, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

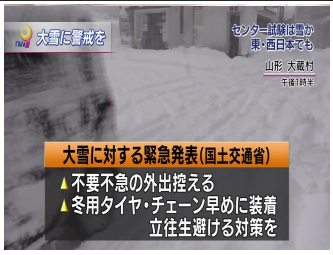
Joji Toyama, Masanori Misono, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Neural machine translation with latent semantic of image and text. *arXiv:1611.08459*, 2016.

Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

7 APPENDIX

Associated image		
Input	The Grand Sumo Tournament is the second day.	In a pitcher, 3 people including Otani pitcher has been selected from the Nippon Ham.
Output (ACM)	Today, **Yokozuna** Hakuho will aim for the first victory.	This is an **athlete**.
Output (Baseline)	No. 1 is No. 1 in 1 meter.	Baseball is out in the professional this season.
Generated word	Yokozuna (The highest rank in sumo)	Athlete

(a) Topic : Sports

Associated image		
Input	On the 14th and 15th is the University Entrance examination exam.	On the 14th and 15th is the University Entrance examination exam.
Output (ACM)	The expectation of **snowy** wind is spreading.	The expectation of snowy wind is **spreading**.
Output (Baseline)	There will be a large-scale fire that is also in western Japan and eastern Japan.	There will be a large-scale fire that is also in western Japan and eastern Japan.
Generated word	Snowy	Spreading

(b) Topic : Weather

Figure 4: Additional examples of analysis of association.