ASSOCIATIVE CONVERSATION MODEL: GENERATING VISUAL INFORMATION FROM TEXTUAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we propose an Associative Conversation Model that generates visual information from textual information and uses it for generating sentences in order to utilize visual information in a dialogue system without image input. In research on Neural Machine Translation, there are studies that generate translated sentences using both images and sentences, and these studies show that visual information is useful for generating translated sentences. However, it is not possible to use sentence generation algorithms using images for the dialogue systems since many text-based dialogue systems only accept text input. Our approach generates (associates) visual information from input text and generates response text using context vector fusing associative visual information and sentence textual information. A comparative experiment between our proposed model and a model without association showed that our proposed model is generating useful sentences by associating visual information related to sentences. Furthermore, analysis experiment of association showed that our proposed model generates (associates) visual information effective for sentence generation.

1 INTRODUCTION

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

As a model that can extract knowledge from conversations, the encoder-decoder model has been proposed (Sutskever et al., 2014) (Vinyals & Le, 2015). It consists of an encoder that encodes the input information into a context vector and a decoder that generates sentences using the context. Vinyals & Le (2015) showed that it is possible to extract commonsense knowledge and to conduct conversation by learning pairs of dialogues with the model.For example, Vinyals & Le (2015) reported that when asked who is skywalker, NCM responded “he is a hero. ” .

Their conversation model (NCM) has a problem that it is not possible to respond properly to the input texts that require visual information. For example, Vinyals & Le (2015) reported that when asked how many legs a spider have, NCM responded “three, i think. ” . Further, the image or video may contain more detailed information than texts. Consider, for example, a scene in a news program including a closed caption “one player won the figure skating match” and showing images with skating players with gold medals. Here, in the video, more detailed information such as a gold medal that does not exist directly in the text is presented. We thought that if such detailed visual information could be extracted from the image, more specific and useful texts could be generated, including ”gold medals” which can not be obtained with text alone. In recent years, studies have been reported in which translated sentences are generated by adding image features to the context vector encoded by the encoder-decoder model (Calixto et al., 2017) (Elliott & Kádár, 2017) (Nakayama & Nishida,

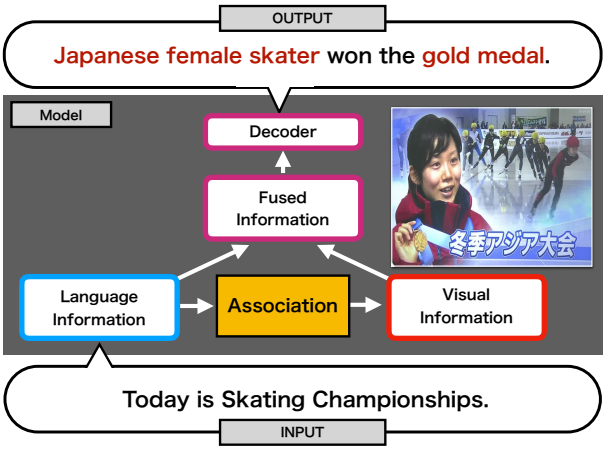

Figure 1: Generating a response by association. Generates corresponding visual information from the textual information, and generates a response text using a vector obtained by fusing two pieces of information.

2017) (Saha et al., 2016) (Toyama et al., 2016). These studies showed that visual information works effectively for generating translation.

Meanwhile, visual information is not considered in many text-based dialogue systems, because what is given to the input is only the utterance text. How can the visual information be used without accepting visual information as the input to the dialogue system?

Based on the discussion above, we propose an Associative Conversation Model that associates the input text with the visual information and generates the response using both the text and the associated visual information. In our proposed method, we attempted to generate response texts using visual information without inputting images. The contribution of this research is as follows:

- We made it possible to generate visual information related to sentence textual information through end-to-end learning of dialogue.
- We made it possible to generate sentences using visual information without directly inputting visual information by association.
- Our proposed model can generate response texts including useful information compared with a model without association by associating visual information related to input text. This is a useful method for constructing the text-based dialogue systems that automatically extract information from the text and the video data (e.g. TV news) to generate sentences.

2 Associative Conversation Model

2.1 Overview of architecture

Figure 1 shows our overall model. The objective of this model is to enable for the decoder to generate a more appropriate response text which cannot be obtained only by textual information, by adaptively referring to the context vector based on visual information. However, in this task, only the textual description is given as input. Therefore, we adopted the following approach so that visual information can be utilized from the textual information.

- First, visual association is performed from the input text, and the visual information corresponding to the text is generated.
  (For example, Figure 1 shows that the input text “Today is Skating Championships” is used to generate the visual information corresponding to a scene where a player in Skating Championships won the gold medal.)
- Next, by fusing the textual information and the associated visual information is obtained the information reflecting either or both as necessary.

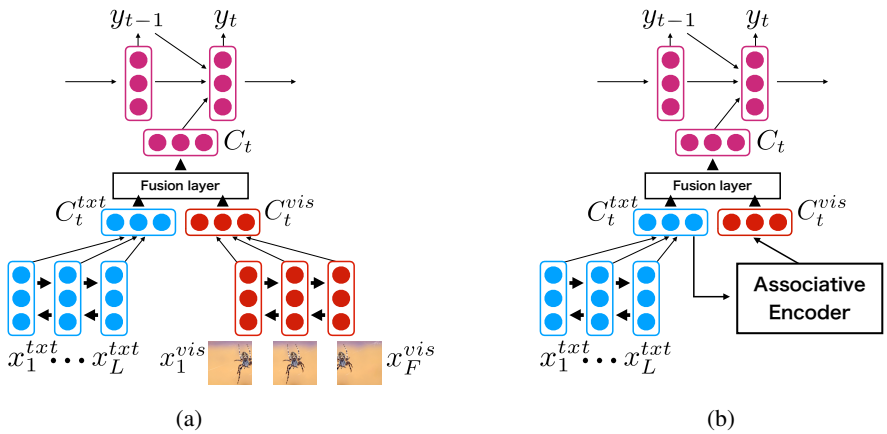

Figure 2: (a) Step1 : A model that performs preliminary learning for context vector extraction.
(b) Step3 : Associative Conversation Model.

(Figure 1 shows that the fused information is generated from the textual information "skating Japan Championships" and the visual information "gold medal, skater".)

- Finally, a response text is generated based on the fused information.
(Figure 1 shows that the fused information was used to decode the final response "Japanese female skater won the gold medal".)

Our approach is simple. When generating sentences from the input texts and videos, an encoder for text and another for video can be used to acquire context vectors for text and for video, respectively. In our task, however, the videos cannot be directly obtained from the input because only the textual information is given as input. Therefore, our idea is to replace the encoder for video with a mechanism for generating the context vectors of the video from texts. The configuration of this associative conversation model is simple, and is composed of the encoder for text and the decoder equipped with attention mechanism, as well as the visual association encoder from texts. Associative encoder is an RNN which generates visual context vectors from textual context vectors (see Sec 2.2.2). We used LSTMs for encoder, decoder, and the associative encoder, respectively (Hochreiter & Schmidhuber, 1997). The input to the associative conversation model is textual descriptions only, and the output is a response text to the input. This model generates the response text by inputting to the decoder the synthesized context vector obtained by fusing the textual context vector and the visual context vector generated from the input text. The associative conversation model is able to learn the pairs of dialogue texts by end-to-end, like a normal encoder-decoder model. Actually, however, two prior learnings are performed before training the associative conversation model. One is the training of multimodal encoder-decoder model which generates a response text from the video and text. The other is the training of the associative encoder which predicts the visual context vector from the textual context vector obtained by the multimodal encoder-decoder model.

2.2 Learning Method

Learning consists of the following three steps. **In order to learn the associative conversation model in the final step3, it is necessary to learn two models in step1 and step2.**

Step 1 : Extraction of context vectors between textual and visual information
Figure 2a shows a network model used in step1. End-to-end learning is performed on the model that inputs the utterance text X_{txt} and the video X_{vis} corresponding to the utterance and outputs the response sentence Y. In this learning, the following four components are trained simultaneously.

- Textual encoder inputting X_{txt} and outputting C_{txt} which is the textual context vector
- Visual encoder inputting X_{vis} and outputting C_{vis} which is the visual context vector

- Fusion layer inputting C_{txt} and C_{vis} and outputting the fused context vector C
- Decoder with attention mechanism, inputting C and outputting the response text Y

The trained model is used to extract the correspondence between the textual and visual information (Fig. 2a).

Step 2 : Learning association
In this step, the associative encoder is trained, which inputs the textual context vector C_{txt} extracted in step1 and outputs the visual context vector C_{vis} corresponding to the input utterance text.

Step 3 : Generating responses using association
In step3, learning is performed in the network where the visual encoder in step1 is replaced with the associative encoder trained in step2. This model is a generation-based model which inputs X_{txt}, and outputs a response text Y, by using the fused context vector C obtained from C_{txt} and C_{vis} in the fusion layer. That is, the structure of the Associative Conversation Model is the same as the network of step1, except that it uses an associative encoder instead of the visual encoder. In the fusion layer, the decoder with attention, the weights learned in step1 are initialized and and trained again.

2.2.1 STEP1 : EXTRACTION OF CONTEXT VECTORS THE CORRESPONDENCE BETWEEN THE TEXTUAL AND VISUAL INFORMATION

The training data used in this step are the utterance text, the video corresponding to the utterance text, and the response text for the utterance. We used videos rather than images because the texts often include the expression of actions. We created the dataset based on TV programs as they contain the utterance texts, their corresponding videos, and their response texts. First, the closed caption texts were extracted from TV news programs as the utterance texts X_{txt}. Also, the scenes corresponding the temporal intervals where the utterances occurred were extracted as the sequence of images X_{vis} from the TV news programs. For more information on the dataset, see Sec 3. Here, the utterance X_{txt} is represented by the sequence of the corresponding word x^{txt}. Similarly, the video X_{vis} corresponding to the utterance text is represented by the sequence of image features x^{vis}. In this work, the image features acquired by the already learned CNN were used instead of learning the image features directly from the input video. Therefore, the input to the prior learning models are the sequence of word vectors $X_{txt} = (x_1^{txt}, ..., x_L^{txt})$ and the sequence of image feature vectors $X_{vis} = (x_1^{vis}, ..., x_F^{vis})$. Also, the output is the sequence of words vectors $Y = (y_1, ..., y_T)$. Here, L, F, and T represent the length of each input text, the number of the input images, and the length of the output text, respectively. The model used in step1 is the multimodal encoder-decoder model consisting of a textual encoder, a visual encoder, and a decoder with attention (Bahdanau et al., 2014) . In step1, the textual and visual encoders encode the input text and video, respectively, to obtain the context vectors C_t^{txt} and C_t^{vis}. Then, C_t which is the fused context vector is input to the decoder to generate a response text. For extracting context vector C_t, the attention mechanism was used.

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

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

Attention mechanism can extract a context vector that strongly reflects some parts of the sequence corresponding to the words of the response text. For example, when the word "spider" of the text and the image of the spider were paid attention to, the textual and visual context vectors of the spider are generated, respectively. In this case, there is a correspondence relations between the textual and visual context vectors. So, in step2, the associative encoder which predicts the visual context vectors from the textual context vectors is trained. The purpose of step1 is to extract the textual and visual context vectors having correspondence relations to be used for learning in step2. The usual attention-based encoder-decoder model calculates the context vectors by Eq. , , . h_i is the intermediate vector of the bidirectional LSTM computed from the input sequence $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$.

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

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

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

Meanwhile, the multimodal encoder-decoder model acquires two context vectors C_t^{txt} and C_t^{vis}. Thus, the multimodal encoder-decoder model uses the following equations instead of Eq.(3) (4) and (5).

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

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

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

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

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

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

Also, there are some responses that do not always require visual information in the dialogue. Therefore, the decoder needs to be able to "measure" which of the textual and visual information should be paid attention to and how strongly it should be focused. The degree to which the textual and visual context vectors C_t^{txt} and C_t^{vis} should be referred to can be learnt by introducing the weight matrices and by multiplying them to C_t^{txt} and C_t^{vis}, respectively. Therefore, the context vector eventually passed to the decoder is the fused context vector having how strongly the textual and visual information should be referred to.

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

This fused context vector C_t is used to predict the next word in the decoder. We call this layer the fusion layer. After training the multi-modal encoder-decoder model, the training data is again input to the model to save the textual and visual context vectors.

2.2.2 STEP2 : LEARNING FOR VISUAL ASSOCIATION

In step2, the visual associative encoder is trained. The associative encoder is trained so that it can predict the visual context vector from the textual context vector. In addition, the associative encoder decides the next visual context vector by referring to the past textual context vectors. Let us consider the following example. The visual information (visual context vector) associated with the word "squeeze" is diverse. For example, the phrases "squeeze a lemon", "squeeze a person's hand", and "squeeze toothpaste out" would make us imagine the different visual situations representing "squeeze". In this example, the phrases "lemon", "person's hand", and "toothpaste" make the situation of "squeeze" more concrete. In this work, the association is to generate the corresponding visual information from the textual information. Namely, there is a demand that the model can imagine what to "squeeze". Here, suppose that a context vector "squeeze" is obtained by the attention mechanism. If the context vector "lemon" was obtained before this point, it will be possible to generate the visual information of "squeeze a lemon" from the context vectors of "lemon" and "squeeze". This indicates the utilization of temporal features. Therefore, RNN is used for the associative encoder. Accordingly, the associative encoder can be seen as a regression model with the input sequence of context vectors of words $C^{txt} = (C_1^{txt}, ..., C_t^{txt}, ..., C_T^{txt})$ and the output sequence of context vectors of images $C^{vis} = (C_1^{vis}, ..., C_t^{vis}, ..., C_T^{vis})$ (Eq. 13). Here, T is the length of the output text. We used LSTM which is capable of learning long-term dependence.

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

2.2.3 STEP3 : GENERATING RESPONSES WITH VISUAL ASSOCIATION

In step3, the associative conversation model learns response texts using the associative visual information instead of visual information directly obtained from video (Fig. 2b). The input to the associative conversation model is the sequence of words $X_{txt} = (x_1^{txt}, ..., x_L^{txt})$. The output is also the sequence of words $Y = (y_1, ..., y_T)$.

In the associative conversation model, the visual encoder obtained in the prior learning of step1 is replaced with the associative encoder. Therefore, the architecture of the associative conversation model can be obtained by replacing Eq.(7) in step1 with Eq.(13). In other words, the mechanism of blending the textual and visual information trained in step1 is left unchanged, and the visual context vector is predicted from the textual context vector by the associative encoder. The visual context vectors as well as the textual context vectors are given to the decoder, and the decoder, the attention, and the fusion layer (Eq. 12) are trained. The weights of the decoder, the attention and the fusion layer are initialized, and the weights obtained by prior learning of step1 is not used. The weights in the textual and associative encoders are not updated, and the same parameters used in the prior learning of step1 are given. The reason for re-training the decoder, the attention and the fusion layer is because the associative visual representation has different property from the visual context vector generated in step1. The associative visual representation is different from the visual context vector obtained in step1. We thought that re-training parts other than encoding of semantic representation enables to generate response texts using the associative visual elements in common knowledge.

3 DATASET

The model was learned using the following data collected independently: subtitles in TV news and the corresponding video where the subtitles were displayed. The visual encoder of the model of step1 encodes a sequence of images (video).

Note that the subtitles were delimited by “ ! ” , “ ? ”, or Japanese period to acquire sentences, and a pair of one sentence and the following sentence was regarded as one dialogue. Therefore, the response text of a certain utterance becomes the next input utterance. Also, the video was cut out as a frame sequence with a frame rate of 5 fps. Each frame was input to the prior learned convolutional neural network, and the output of the last pooling layer was used as the image features. VGG16 was used for the convolutional neural network (Simonyan & Zisserman, 2014). The recorded programs were 163 Japanese TV news broadcast in from December, 2016 to March, 2017, having 38K dialogues and 19K vocabulary words. Data was divided into 34K dialogue for training data and 4K dialogue for test data.

4 EXPERIMENTS

In this section, we describe the result of comparison between an Associative Conversation Model and a model without association. We used a encoder-decoder model with attention that caused the pair of dialogue sentences for the baseline model. The baseline model generate a sentence Y as utterance text X_{txt}. Both the proposed model and the baseline model was trained with the same dialogue sentences, and we investigated the effect of association. In addition, we visualized the associated objects, and investigated whether associative conversation model acquired effective visual information for response generation. As a result of the experiments, we found that the association works effectively to generate sentences with useful information. We also found that our proposed model associates visual information related to input texts.

4.1 MODEL CONFIGURATION

In the step1, we used single layer LSTMs as both the encoders and the decoder. The hidden layer dimension of the textual encoder was set to 512, the hidden layer dimension of the visual encoder to 512, the dimension of the fused context vector to 1024, and the hidden layer dimension of the decoder to 1024, the dimensionality word embeddings was set to 512, the dimensionality image feature was set to 512, the batch size was set to 64. Adagrad was used for optimization (Duchi et al., 2011).

In the step2, the associative encoder is composed of 4-layer LSTM with the hidden layers dimension being 1024, the batch size was 64. Adam was used for optimization (Kingma & Ba, 2014).

In the step3, the model is composed of a textual encoder, decoder, and a associative encoder instead of visual encoder. The hidden layer dimension of the textual encoder was set to 512, the dimension of the fused context vector to 1024, and the hidden layer dimension of the decoder to 1024, the

Associated image		
Input	On the 14th and 15th is the University Entrance examination exam.	Well, today is All Japan Figure Skating Championships.
Output (ACM)	**Snow** and wind are expected to spread.	A player who has won the **gold medal** three women's singles.
Output (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.
Generated word	**Snow**	**Gold medal**

Figure 3: This result shows that the visual information associated with the input text by our model (ACM) corresponds to the input text. In addition, we can see that association of them produces words with valuable information. For example, The result on the left shows our model associates the snow scene with the utterance text and generated the word “snow”. Importantly, it is a fact that snow fell on the day of the exam and the word “snow” is not related to the word “exam” normally, so it is a word which can not be easily generated unless it is associated with the snow scene .

dimensionality word embeddings was set to 512, the batch size was set to 64. Adagrad was used for optimization

Also, in the step3, weights and biases of the decoder and attention were used the weights in step1 and not updated. The weights of associative encoder were not updated either. The weights of the decoder are initialized in the step3 (See Eq.(1), (2), (10), (11), and (12)) .

The baseline model is composed of single layer LSTMs as the encoder and the decoder. The hidden layer dimension of the textual encoder was set to 512, the dimension of the fused context vector to 1024, and the hidden layer dimension of the decoder to 1024, the dimensionality word embeddings was set to 512, the batch size was set to 64. Adagrad was used for optimization.

4.2 Results

4.3 Analysis of Association

Figure 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 text. We input the video of the learning data (sequence of images) to the model of step1 and we measured the meaning expression and the associative representation of the animation generated by the visual encoder with cosine similarity. Although it is indirect, the context vector 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 Figure 3 is visualized visual information associated with the sentence using the image in the learning data. From the result of Figure 3, you can see that the sentence and associated visual information are associated. For example, in

the example on the left of Figure 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 an example on the right of Figure 3 misunderstood figure skating as speed skating, the skating player has won the gold "skating 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 text was misunderstood and associated with visual information. For example, in the example to the right of Figure 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 step3. That is, using the knowledge extracted by encoder in step1 and 2, in step3, relearning is performed using a general dialogue sentence. If knowledge can be extracted, learning can be performed more efficiently by performing Step3 only in accordance with the task.

5 Related Work

In studies about conversation model based on maltimedia data, Mostafazadeh et al. (2017) used deep neural network models trained on social media data (pair of image and text), and show their approach improves the quality of response generation. Although they assumed Image-Grounded Conversations, we developed their idea and assumed the case of not using the images during the conversation. "In recent years, studies have been reported in which translated sentences are generated by adding image features to the context vector encoded by the encoder-decoder model (Calixto et al., 2017) (Elliott & Kádár, 2017) (Nakayama & Nishida, 2017) (Saha et al., 2016) (Toyama et al., 2016). These studies showed that visual information works effectively for generating translation." The approaches that use images only when training such as our approach are Toyama et al. (2016) and Elliott & Kádár (2017).

(Toyama et al., 2016) proposed a neural machine translation model that introduces a continuous latent variable containing an underlying semantic extracted from texts and images. Their model is the encoder-decoder with attention, but there is a mechanism to generate the latent variables. When for predicting a translation in decoder inputting the sentence to encoder, decoder generate a target word using a latent variable generated from the source text, target text, and an image.

(Elliott & Kádár, 2017) show their approach using multitask learning model, inputting the text and outputting the text and the image feature, improves translation performance. The network of structure is attention-based encoder-decoder, but has two decoders outputting text or image feature vector $\hat{v}$. This decoder outputting image feature vector is trained to predict the true image vector v.

There are important differences between our approach and their approaches. First, although their approach has given the image as a visual information when training, our approach has given the video (sequence of image features) to expand their approach. By this extension, we think our model is able to learn words expressing motion effectively. Secondly, we are dealing the dialogue task, not the translation task. Thirdly, we challenged our model extract knowledge from the noisy data that the visual information at that time and the sentence does not necessarily correspond (News of the video and subtitle). Therefore, our model includes a fusion layer learning to use the measure of visual information and textual information (Eq. 12). The difference between Elliott & Kádár (2017) and our approach is our approach is not multitask learning model, thus out model receives input of visual information. Hence it is possible to perform attention to visual information during training can be extracted visual information according to the response word.

6 Conclusions

In a study applying a sentence generation algorithm of translation sentence to a conversation model, there was a problem that it was not possible to respond well to an input text which requires visual information. However, it is not possible to use sentence generation algorithms using images for the dialogue systems since many text-based dialogue systems only accept text input. Based on

the discussion above, we propose an Associative Conversation Model that associates the input text with the visual information and generates the response using both the text and the associated visual information. Comparative experiments with models that do not use association show that association of visual information related to input texts produces response texts 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 textual information through end-to-end learning of dialogue. Our approach is a useful for constructing the text-based dialogue systems that automatically extract information from the text and the video data (e.g. TV news) to generate sentences.

REFERENCES

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

Iacer Calixto, Qun Liu, and Nick Campbell. 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*, pp. 1913–1924, 2017. doi: 10.18653/v1/P17-1175. URL https://doi.org/10.18653/v1/P17-1175.

John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011. URL http://dl.acm.org/citation.cfm?id=2021068.

Desmond Elliott and Ákos Kádár. Imagination improves multimodal translation. *CoRR*, abs/1705.04350, 2017. URL http://arxiv.org/abs/1705.04350.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL http://arxiv.org/abs/1412.6980.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *CoRR*, abs/1701.08251, 2017. URL http://arxiv.org/abs/1701.08251.

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. doi: 10.1007/s10590-017-9197-z. URL https://doi.org/10.1007/s10590-017-9197-z.

Amrita Saha, Mitesh M. Khapra, Sarath Chandar, Janarthanan Rajendran, and Kyunghyun Cho. 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*, pp. 109–118, 2016. URL http://aclweb.org/anthology/C/C16/C16-1011.pdf.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 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*, 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. *CoRR*, abs/1611.08459, 2016. URL http://arxiv.org/abs/1611.08459.

Oriol Vinyals and Quoc V. Le. A neural conversational model. *CoRR*, abs/1506.05869, 2015. URL http://arxiv.org/abs/1506.05869.

7 APPENDIX

Associated image		
Input	The Grand Sumo Tournament is the second day.	In a pitcher, three people including Otani 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	The vicinity of Honshu is expected to be covered widely by mobile high pressure.	In the morning of tomorrow, it is affected by the developing low pressure mainly in the western Japan and East Japan Sea side.
Output (ACM)	Tomorrow morning, it will be **sunny** in many places from western Japan to eastern Japan, and western Japan side of the Sea side.	Please also note the influence on the **traffic** caused by **windstorms**, heavy blizzards, and snowdrift.
Output (Baseline)	Then, the weather in various places of tomorrow.	Vigilance is necessary for windstorms and high waves.
Generated word	**Sunny**	**Traffic, Windstorms**

(b) Topic : Weather

Figure 4: Additional examples of analysis of association.