

Conclusion-Supplement Answer Generation for Non-Factoid Questions

Makoto Nakatsuji, Sohei Okui

NTT Resonant Inc.

Granparktower, 3-4-1 Shibaura, Minato-ku, Tokyo 108-0023, Japan

nakatsuji.makoto@gmail.com, okui@ntr.co.jp

Abstract

This paper tackles the goal of conclusion-supplement answer generation for non-factoid questions, which is a critical issue in the field of Natural Language Processing (NLP) and Artificial Intelligence (AI), as users often require supplementary information before accepting a conclusion. The current encoder-decoder framework, however, has difficulty generating such answers, since it may become confused when it tries to learn several different long answers to the same non-factoid question. Our solution, called an *ensemble network*, goes beyond single short sentences and fuses logically connected conclusion statements and supplementary statements. It extracts *the context* from the conclusion decoder’s output sequence and uses it to create supplementary decoder states on the basis of an attention mechanism. It also assesses *the closeness* of the question encoder’s output sequence and the separate outputs of the conclusion and supplement decoders as well as *their combination*. As a result, it generates answers that match the questions and have natural-sounding supplementary sequences in line with the context expressed by the conclusion sequence. Evaluations conducted on datasets including “Love Advice” and “Arts & Humanities” categories indicate that our model outputs much more accurate results than the tested baseline models do.

Introduction

Question Answering (QA) modules play particularly important roles in recent dialog-based Natural Language Understanding (NLU) systems, such as Apple’s Siri and Amazon’s Echo. Users chat with AI systems in natural language to get the answers they are seeking. QA systems can deal with two types of question: factoid and non-factoid ones. The former sort asks, for instance, for the name of a thing or person such as “What/Who is X ?”. The latter sort includes more diverse questions that cannot be answered by a short fact. For instance, users may ask for advice on how to make a long-distance relationship work well or for opinions on public issues. Significant progress has been made in answering factoid questions (Wang, Smith, and Mitamura 2007; Yu et al. 2014); however, answering non-factoid questions remains a challenge for QA modules.

Long short term memory (LSTM) sequence-to-sequence models (Sutskever, Vinyals, and Le 2014; Vinyals and Le 2015; Bahdanau, Cho, and Bengio 2014) try to generate short replies to the short utterances often seen in chat systems. Evaluations have indicated that these models have the possibility of supporting simple forms of general knowledge QA, e.g. “Is the sky blue or black?”, since they learn commonly occurring sentences in the training corpus. Recent machine reading comprehension (MRC) methods (Nguyen et al. 2016; Rajpurkar et al. 2016) try to return a single short answer to a question by extracting answer spans from the provided passages. Unfortunately, they may generate unsatisfying answers to regular non-factoid questions because they can easily become confused when learning several different long answers to the same non-factoid question, as pointed out by (Jia and Liang 2017; Wang et al. 2018).

This paper tackles a new problem: conclusion-supplement answer generation for non-factoid questions. Here, the conclusion consists of sentences that directly answer the question, while the supplement consists of information supporting the conclusion, e.g., reasons or examples. Such conclusion-supplement answers are important for helping questioners decide their actions, especially in NLU. As described in (Ennis 1991), users prefer a supporting supplement before accepting an instruction (i.e., a conclusion). Good debates also include claims (i.e., conclusions) about a topic and supplements to support them that will allow users to reach decisions (Rinott et al. 2015). The following example helps to explain how conclusion-supplement answers are useful to users: “Does separation by a long distance ruin love?” Current methods tend to answer this question with short and generic replies, such as, “Distance cannot ruin true love”. The questioner, however, is not likely to be satisfied with such a trite answer and will want to know how the conclusion was reached. If a supplemental statement like “separations certainly test your love” is presented with the conclusion, the questioner is more likely to accept the answer and use it to reach a decision. Furthermore, there may be multiple answers to a non-factoid question. For example, the following answer is also a potential answer to the ques-

tion: “distance ruins most relationships. You should keep in contact with him”. The current methods, however, have difficulty generating such conclusion-supplement answers because they can become easily confused when they try to learn several different and long answers to a non-factoid question.

To address the above problem, we propose a novel architecture, called the *ensemble network*. It is an extension of existing encoder-decoder models, and it generates two types of decoder output sequence, conclusion and supplement. It uses two viewpoints for selecting the conclusion statements and supplementary statements. (Viewpoint 1) *The context* present in the conclusion decoder’s output is linked to supplementary-decoder output states on the basis of an attention mechanism. Thus, the context of the conclusion sequence directly impacts the decoder states of the supplement sequences. This, as a result, generates natural-sounding supplementary sequences. (Viewpoint 2) *The closeness* of the question sequence and conclusion (or supplement) sequence as well as the closeness of the question sequence with *the combination* of conclusion and supplement sequences is considered. By assessing the closeness at the sentence level and sentence-combination level in addition to at the word level, it can generate answers that include good supplementary sentences following the context of the conclusion. This avoids having to learn several different conclusion-supplement answers assigned to a single non-factoid question and generating answers whose conclusions and supplements are logically inconsistent with each other.

Community-based QA (CQA) websites tend to provide answers composed of conclusion and supplementary statements; from our investigation, 77% of non-factoid answers (love advice) in the Oshiete-goo (<https://oshiete.goo.ne.jp>) dataset consist of these two statement types. The same is true for 82% of the answers in the Yahoo non-factoid dataset¹ related to the fields of social science, society & culture and arts & humanities. We used the above-mentioned CQA datasets in our evaluations, since they provide diverse answers given by many responders. The results showed that our method outperforms existing ones at generating correct and natural answers. We also conducted an love advice service² in Oshiete goo to evaluate the usefulness of our ensemble network.

Related work

The encoder-decoder framework learns how to transform one representation into another. Contextual LSTM (CLSTM) incorporates contextual features (e.g., topics) into the encoder-decoder framework (Ghosh et al. 2016; Serban et al. 2016). It can be used to make the context of the question a part of the answer generation process. Hi-eRarchical Encoder Decoder (HRED) (Serban et al. 2016) extends the hierarchical recurrent encoder-decoder neural network into the dialogue domain; each question can be encoded into a dense context vector, which is used to recurrently decode the tokens in the answer sentences.

Such sequential generation of next statement tokens, however, weakens the original meaning of the first statement (question). Recently, several models based on the Transformer (Vaswani et al. 2017), such as for passage ranking (Nogueira et al. 2019; Liu, Duh, and Gao 2018) and answer selection (Shao et al. 2019), have been proposed to evaluate question-answering systems. There are, however, few Transformer-based methods that generate non-factoid answers.

Recent neural answer selection methods for non-factoid questions (dos Santos et al. 2015; Qiu and Huang 2015; Tan et al. 2016) learn question and answer representations and then match them using certain similarity metrics. They use open datasets stored at CQA sites like Yahoo! Answers since they include many diverse answers given by many responders and thus are good sources of non-factoid QA training data. The above methods, however, can only select and extract answer sentences, they do not generate them.

Recent machine reading comprehension methods try to answer a question with exact text spans taken from provided passages (Yu et al. 2018; Rajpurkar et al. 2016; Yang, Yih, and Meek 2015; Joshi et al. 2017). Several studies on the MS-MARCO dataset (Tan et al. 2017; Nguyen et al. 2016; Wang et al. 2018) define the task as using multiple passages to answer a question where the words in the answer are not necessarily present in the passages. Their models, however, require passages other than QA pairs for both training and testing. Thus, they cannot be applied to CQA datasets that do not have such passages. Furthermore, most of the questions in their datasets only have a single answer. Thus, we think their purpose is different from ours; generating answers for non-factoid questions that tend to demand diverse answers.

There are several complex QA tasks such as those present in the TREC complex interactive QA tasks³ or DUC⁴ complex QA tasks. Our method can be applied to those non-factoid datasets if an access fee is paid.

Model

This section describes our conclusion-supplement answer generation model in detail. An overview of its architecture is shown in Figure 1.

Given an input question sequence $\mathbf{Q} = \{\mathbf{q}_1, \dots, \mathbf{q}_i, \dots, \mathbf{q}_{N_q}\}$, the proposal outputs a conclusion sequence $\mathbf{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_t, \dots, \mathbf{c}_{N_c}\}$, and supplement sequence $\mathbf{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_t, \dots, \mathbf{s}_{N_s}\}$. The goal is to learn a function mapping from \mathbf{Q} to \mathbf{C} and \mathbf{S} . Here, \mathbf{q}_i denotes a one-of- K embedding of the i -th word in an input sequence of length N_q . \mathbf{c}_t (\mathbf{s}_t) denotes a one-of- K embedding of the t -th word in an input sequence of length N_c (N_s).

Encoder

The encoder converts the input \mathbf{Q} into a question embedding, \mathbf{O}_q , and hidden states, $\mathbf{H} = \{\mathbf{h}_i\}_i$.

¹<https://ciir.cs.umass.edu/downloads/nfL6/>

²<http://oshiete.goo.ne.jp/ai>

³<https://cs.uwaterloo.ca/~jimmylin/ciqa/>

⁴<http://www-nlpir.nist.gov/projects/duc/guidelines.html>

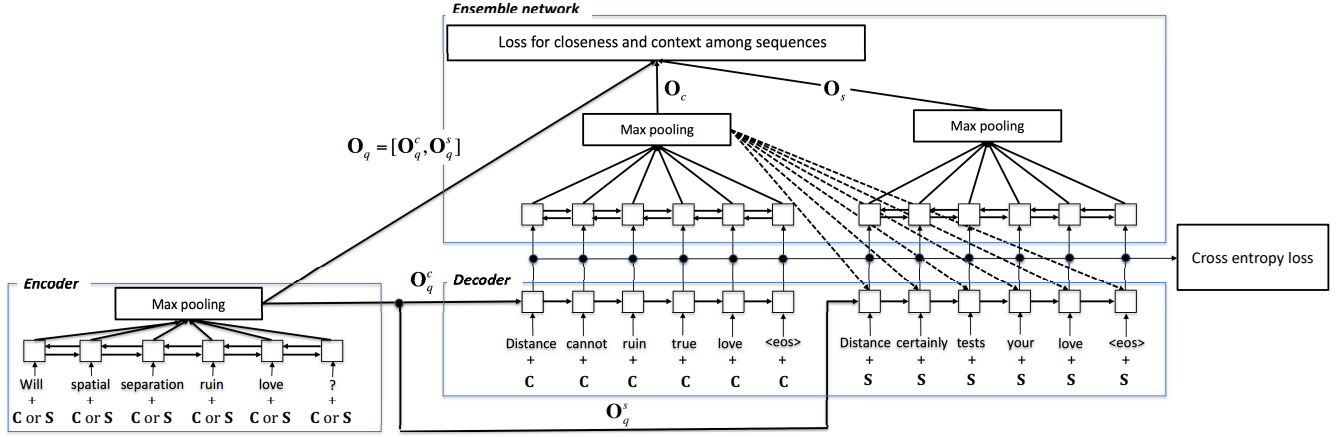


Figure 1: Neural conclusion-supplement answer generation model.

Since the question includes several pieces of background information on the question, e.g. on the users' situation, as well as the question itself, it can be very long and composed of many sentences. For this reason, we use the BiLSTM encoder, which encodes the question in both directions, to better capture the overall meaning of the question. It processes both directions of the input, $\{q_1, \dots, q_{N_q}\}$ and $\{q_{N_q}, \dots, q_1\}$, sequentially. At time step t , the encoder updates the hidden state by:

$$h_i = [h_i^f, h_i^b]^T \text{ s.t.}$$

$$h_i^f = f(q_{i-1}, h_{i-1}^f), h_i^b = f(q_{i+1}, h_{i+1}^b),$$

where $f()$ is an LSTM unit, and h_i^f and h_i^b are hidden states output by the forward-direction LSTM and backward-direction LSTM, respectively.

We also want to reflect sentence-type information such as conclusion type or supplement type in sequence-to-sequence learning to better understand the conclusion or supplement sequences. We achieve this by adding a sentence type vector for conclusion C or for supplement S to the input gate, forget gate output gate, and cell memory state in the LSTM model. This is equivalent to processing a composite input $[q_i, C]$ or $[q_i, S]$ in the LSTM cell that concatenates the word embedding and sentence-type embedding vectors. We use this modified LSTM in the above BiLSTM model as:

$$h_i = [h_i^f, h_i^b]^T \text{ s.t.}$$

$$h_i^f = f([q_{i-1}, C], h_{i-1}^f), h_i^b = f([q_{i+1}, S], h_{i+1}^b).$$

When encoding the question to decode the supplement sequence, S is input instead of C in the above equation.

The BiLSTM encoder then applies a max-pooling layer to all hidden vectors to extract the most salient signal for each word. As a result, it generates a fixed-sized distributed vector representation for the conclusion, O_q^c , and another for the supplement, O_q^s . O_q^c and O_q^s are different since the encoder is biased by the corresponding sentence-type vector, C or S .

As depicted in Figure 1, the BiLSTM encoder processes each word with a sentence-type vector (i.e. C or S) and the

max-pooling layer to produce the question embedding O_q^c or O_q^s . These embeddings are used as context vectors in the decoder network for the conclusion and supplement.

Decoder

The decoder is composed of a conclusion decoder and supplement decoder. Here, let h_t' be the hidden state of the t -th LSTM unit in the conclusion decoder. Similar to the encoder, the decoder also decodes a composite input $[c_t, C]$ in an LSTM cell that concatenates the conclusion word embedding and sentence-type embedding vectors. It is formulated as follows:

$$h_t' = f'([c_{t-1}, C], h_{t-1}') \text{ s.t.}$$

$$c_{t-1} = \underset{c}{\operatorname{argmax}} \operatorname{softmax}_c(h_{t-1}'),$$

where $f'()$ denotes the conclusion decoder LSTM, $\operatorname{softmax}_c$ the probability of word c given by a softmax layer, c_t the t -th conclusion decoded token, and c_t the word embedding of c_t . The supplement decoder's hidden state h_t'' is computed in the same way with h_t' ; however, it is updated in the ensemble network described in the next subsection.

As depicted in Figure 1, the LSTM decoder processes tokens according to question embedding O_q^c or O_q^s , which yields a bias corresponding to the sentence-type vector, C or S . The output states are then input to the ensemble network.

Ensemble network

The conventional encoder-decoder framework often generates short and simple sentences that fail to adequately answer non-factoid questions. Even if we force it to generate longer answers, the decoder output sequences become incoherent when read from the beginning to the end.

The ensemble network solves the above problem by (1) passing *the context* from the conclusion decoder's output sequence to the supplementary decoder hidden states via an attention mechanism, and (2) considering *the closeness* of the

encoder’s input sequence to the decoders’ output sequences as well as the closeness of the encoder’s input sequence to the combination of decoded output sequences.

(1) To control the *context*, we assess all the information output by the conclusion decoder and compute the conclusion vector, \mathbf{O}_c . \mathbf{O}_c is a sentence-level representation that is more compact, abstractive, and global than the original decoder output sequence. To get it, we apply BiLSTM to the conclusion decoder’s output states $\{\tilde{\mathbf{y}}_t^c\}_t$; i.e., $\{\tilde{\mathbf{y}}_t^c\}_t = \{\mathbf{U} \cdot \text{softmax}(\mathbf{h}_t^c)\}_t$, where word representation matrix \mathbf{U} holds the word representations in its columns. At time step t , the BiLSTM encoder updates the hidden state by:

$$\mathbf{h}_t^c = [\mathbf{h}_t^{c,f}, \mathbf{h}_t^{c,b}]^T \text{ s.t.}$$

$$\mathbf{h}_t^{c,f} = f(\tilde{\mathbf{y}}_{t-1}^c, \mathbf{h}_{t-1}^{c,f}), \quad \mathbf{h}_t^{c,b} = f(\tilde{\mathbf{y}}_{t+1}^c, \mathbf{h}_{t+1}^{c,b}),$$

where $\mathbf{h}_t^{c,f}$ and $\mathbf{h}_t^{c,b}$ are the hidden states output by the forward LSTM and backward LSTM in the conclusion encoder, respectively. It applies a max-pooling layer to all hidden vectors to extract the most salient signal for each word to compute the embedding for conclusion \mathbf{O}_c . Next, it computes the context vector $\mathbf{c}\mathbf{x}_t$ at the t -th step by using the $(t-1)$ -th output hidden state of the supplement decoder, \mathbf{h}_{t-1}'' , weight matrices, \mathbf{V}_a and \mathbf{W}_a , and a sigmoid function, σ :

$$\mathbf{c}\mathbf{x}_t = \alpha_t \mathbf{O}_c \text{ s.t. } \alpha_t = \sigma(\mathbf{V}_a^T \tanh(\mathbf{W}_a \mathbf{h}_{t-1}'' + \mathbf{O}_c)).$$

This computation lets our ensemble network extract a conclusion-sentence level context. The resulting supplement sequences follow the context of the conclusion sequence. Finally, \mathbf{h}_t'' is computed as:

$$\begin{aligned} \mathbf{z}_t &= \sigma(\mathbf{W}_z[\mathbf{y}_{t-1}, \mathbf{T}] + \mathbf{U}_z \mathbf{h}_{t-1}'' + \mathbf{W}_z \mathbf{c}\mathbf{x}_t + \mathbf{b}_z) \quad (1) \\ \tilde{\mathbf{l}}_t &= \tanh(\mathbf{W}_l[\mathbf{y}_{t-1}, \mathbf{T}] + \mathbf{U}_l \mathbf{h}_{t-1}'' + \mathbf{W}_l \mathbf{c}\mathbf{x}_t + \mathbf{b}_l) \\ \mathbf{l}_t &= \mathbf{i}_t * \tilde{\mathbf{l}}_t + \mathbf{f}_t * \mathbf{l}_{t-1} \\ \mathbf{h}_t'' &= \mathbf{o}_t * \tanh(\mathbf{l}_t) \end{aligned}$$

z can be i , f , or o , which represent three gates (e.g., input \mathbf{i}_t , forget \mathbf{f}_t , and output \mathbf{o}_t). \mathbf{l}_t denotes a cell memory vector. \mathbf{W}_z^a and \mathbf{W}_l^a denote attention parameters.

(2) To control the *closeness* at the sentence level and *sentence-combination* level, it assesses all the information output by the supplement decoder and computes the supplement vector, \mathbf{O}_s , in the same way as it computes \mathbf{O}_c . That is, it applies BiLSTM to the supplement decoder’s output states $\{\tilde{\mathbf{y}}_t^s\}_t$; i.e., $\{\tilde{\mathbf{y}}_t^s\}_t = \{\mathbf{U} \cdot \text{softmax}(\mathbf{h}_t^s)\}_t$, where the word representations are found in the columns of \mathbf{U} . Next, it applies a max-pooling layer to all hidden vectors in order to compute the embeddings for supplement \mathbf{O}_s . Finally, to generate the conclusion-supplement answers, it assesses the *closeness* of the embeddings for the question \mathbf{O}_q to those for the answer sentences (\mathbf{O}_c or \mathbf{O}_s) and *their combination* \mathbf{O}_c and \mathbf{O}_s . The loss function for the above metrics is described in the next subsection.

As depicted in Figure 1, the ensemble network computes the conclusion embedding \mathbf{O}_c , the attention parameter weights from \mathbf{O}_c to the decoder output supplement states (dotted lines represent attention operations), and the supplement embedding \mathbf{O}_s . Then, \mathbf{O}_c and \mathbf{O}_s are input to the loss function together with the question embedding $\mathbf{O}_q = [\mathbf{O}_q^c, \mathbf{O}_q^s]$.

Loss function of ensemble network

Our model uses a new loss function rather than generative supervision, which aims to maximize the conditional probability of generating the sequential output $p(\mathbf{y}|\mathbf{q})$. This is because we think that assessing the closeness of the question and an answer sequence as well as the closeness of the question to two answer sequences is useful for generating natural-sounding answers.

The loss function is for optimizing the closeness of the question and conclusion and that of the question and supplement as well as for optimizing the closeness of the question with the combination of the conclusion and supplement. The training loss \mathcal{L}_s is expressed as the following hinge loss, where \mathbf{O}^+ is the output decoder vector for the ground-truth answer, \mathbf{O}^- is that for an incorrect answer randomly chosen from the entire answer space, M is a constant margin, and \mathbb{A} is set equal to $\{[\mathbf{O}_c^+, \mathbf{O}_s^-], [\mathbf{O}_c^-, \mathbf{O}_s^+], [\mathbf{O}_c^-, \mathbf{O}_s^-]\}$:

$$\mathcal{L}_s = \sum_{\mathbf{O}_a \in \mathbb{A}} \max\{0, M - (\cos(\mathbf{O}_q, [\mathbf{O}_c^+, \mathbf{O}_s^+]) - \cos(\mathbf{O}_q, \mathbf{O}_a))\}$$

The key idea is that \mathcal{L}_s checks whether or not the conclusion, supplement, and their combination have been well predicted. In so doing, \mathcal{L}_s can optimize not only the prediction of the conclusion or supplement but also the prediction of the combination of conclusion and supplement.

The model is illustrated in the upper part of Figure 1; $(\mathbf{O}_q, \mathbf{O}_c, \mathbf{O}_s)$ is input to compute the closeness and sequence combination losses.

Training

The training loss \mathcal{L}_w is used to check \mathcal{L}_s and the cross-entropy loss in the encoder-decoder model. In the following equation, the conclusion and supplement sequences are merged into one sequence \mathbf{Y} of length T , where $T = N_c + N_s$.

$$\mathcal{L}_w = \alpha \cdot \mathcal{L}_s - \ln \prod_{t=1}^T p(\mathbf{y}_t | \mathbf{Q}, \mathbf{y}_1, \dots, \mathbf{y}_{t-1}). \quad (2)$$

α is a parameter to control the weighting of the two losses. We use adaptive stochastic gradient descent (AdaGrad) to train the model in an end-to-end manner. The loss of a training batch is averaged over all instances in the batch.

Figure 1 illustrates the loss for the ensemble network and the cross-entropy loss.

Evaluation

Compared methods

We compared the performance of our method with those of (1) *Seq2seq*, a seq2seq attention model proposed by (Bahdanau, Cho, and Bengio 2014); (2) *CLSTM*, i.e., the CLSTM model (Ghosh et al. 2016); (3) *Trans*, the Transformer (Vaswani et al. 2017), which has proven effective for common NLP tasks. In these three methods, conclusion sequences and supplement sequences are decoded separately and then joined to generate answers. They give more accurate results than methods in which the conclusion sequences

and supplement sequences are decoded sequentially. We also compared (4) *HRED*, a hierarchical recurrent encoder-decoder model (Serban et al. 2016) in which conclusion sequences and supplement sequences are decoded sequentially to learn the context from conclusion to supplement; (5) *NAGMWA*, i.e., our neural answer generation model without an attention mechanism. This means that *NAGMWA* does not pass \mathbf{cx}_t in Eq. (1) to the decoder, and conclusion decoder and supplement decoder are connected only via the loss function \mathcal{L}_s . In the tables and figures that follow, *NAGM* means our full model.

Dataset

Our evaluations used the following two CQA datasets:

Oshiete-goo The Oshiete-goo dataset includes questions stored in the “love advice” category of the Japanese QA site, Oshiete-goo. It has 771,956 answers to 189,511 questions. We fine-tuned the model using a corpus containing about 10,032 question-conclusion-supplement (q-c-s) triples. We used 2,824 questions from the Oshiete-goo dataset. On average, the answers to these questions consisted of about 3.5 conclusions and supplements selected by human experts. The questions, conclusions, and supplements had average lengths of 482, 41, and 46 characters, respectively. There were 9,779 word tokens in the questions and 6,317 tokens in answers; the overlap was 4,096.

nfL6 We also used the Yahoo nfL6 dataset, the largest publicly available English non-factoid CQA dataset. It has 499,078 answers to 87,361 questions. We fine-tuned the model by using questions in the “social science”, “society & culture”, and “arts & humanities” categories, since they require diverse answers. This yielded 114,955 answers to 13,579 questions. We removed answers that included some stop words, e.g. slang words, or those that only refer to some URLs or descriptions in literature, since such answers often become noise when an answer is generated. Human experts annotated 10,299 conclusion-supplement sentences pairs in the answers.

In addition, we used a neural answer-sentence classifier to classify the sentences into conclusion or supplement classes. It first classified the sentences into supplements if they started with phrases such as “this is because” or “therefore”. Then, it applied a BiLSTM with max-pooling to the remaining unclassified sentences, $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{N_a}\}$, and generated embeddings for the un-annotated sentences, \mathbf{O}^a . After that, it used a logistic sigmoid function to return the probabilities of mappings to two discrete classes: conclusion and supplement. This mapping was learned by minimizing the classification errors using the above 10,299 labeled sentences. As a result, we automatically acquired 70,000 question-conclusion-supplement triples from the entire answers. There were 11,768 questions and 70,000 answers. Thus, about 6 conclusions and supplements on average were assigned to a single question. The questions, conclusions, and supplements had average lengths of 46, 87, and 71 characters, respectively. We checked the performance of the clas-

Table 1: Results when changing α .

α	Oshiete-goo			nfL6		
	0	1	2	0	1	2
ROUGE-L	0.251	0.299	0.211	0.330	0.402	0.295
BLEU-4	0.098	0.158	0.074	0.062	0.181	0.023

Table 2: Results when using sentence-type embeddings.

	Oshiete-goo		nfL6	
	NAGM	w/o ste	NAGM	w/o ste
ROUGE-L	0.299	0.235	0.402	0.349
BLEU-4	0.158	0.090	0.181	0.067

sifier; human experts checked whether the annotation results were correct or not. They judged that it was about 81% accurate (it classified 56,762 of 70,000 sentences into correct classes). There were 15,690 word tokens in questions and 124,099 tokens in answers; the overlap was 11,353.

Methodology

We conducted three evaluations using the Oshiete-goo dataset; we selected three different sets of 500 human-annotated test pairs from the full dataset. In each set, we trained the model by using training pairs and input questions in test pairs to the model. We repeated the experiments three times by randomly shuffling the train/test sets.

For the evaluations using the nfL6 dataset, we prepared three different sets of 500 human-annotated test q-c-s triples from the full dataset. We used 10,299 human-annotated triples to train the neural sentence-type classifier. Then, we applied the classifier to the unlabeled answer sentences. Finally, we evaluated the answer generation performance by using three sets of machine-annotated 69,500 triples and 500 human-annotated test triples.

After training, we input the questions in the test triples to the model to generate answers for both datasets. We compared the generated answers with the correct answers. The results described below are average values of the results of three evaluations.

The softmax computation was slow since there were so many word tokens in both datasets. Many studies (Yin et al. 2016; Yang et al. 2016; Vinyals and Le 2015) restricted the word vocabulary to one based on frequency. This, however, narrows the diversity of the generated answers. Since diverse answers are necessary to properly reply to non-factoid questions, we used bigram tokens instead of word tokens to speed up the computation without restricting the vocabulary. Accordingly, we put 4,087 bigram tokens in the Oshiete-goo dataset and 11,629 tokens in the nfL6 dataset.

To measure performance, we used human judgment as well as two popular metrics (Sutskever, Vinyals, and Le 2014; Yang et al. 2016; Bahdanau, Cho, and Bengio 2014) for measuring the fluency of computer-generated text: ROUGE-L (Lin 2004) and BLEU-4 (Papineni et al. 2002). ROUGE-L is used for measuring the performance for evaluating non-factoid

Table 3: ROUGE-L/BLEU-4 for Oshiete-goo.

	<i>Seq2seq</i>	<i>CLSTM</i>	<i>Trans</i>	<i>HRED</i>	<i>NAGMWA</i>	<i>NAGM</i>
ROUGE-L	0.238	0.260	0.278	0.210	0.291	0.299
BLEU-4	0.092	0.121	0.087	0.042	0.147	0.158

Table 4: ROUGE-L/BLEU-4 for nfL6.

	<i>Seq2seq</i>	<i>CLSTM</i>	<i>Trans</i>	<i>HRED</i>	<i>NAGMWA</i>	<i>NAGM</i>
ROUGE-L	0.291	0.374	0.338	0.180	0.383	0.402
BLEU-4	0.081	0.141	0.122	0.055	0.157	0.181

QAs (Song et al. 2017), however, we also think human judgement is important in this task.

Parameter setup

For both datasets, we tried different parameter values and set the size of the bigram token embedding to 500, the size of LSTM output vectors for the BiLSTMs to 500×2 , and number of topics in the CLSTM model to 15. We tried different margins, M , in the hinge loss function and settled on 0.2. The iteration count N was set to 100.

We varied α in Eq. (2) from 0 to 2.0 and checked the impact of L_s by changing α . Table 1 shows the results. When α is zero, the results are almost as poor as those of the seq2seq model. On the other hand, while raising the value of α places greater emphasis on our ensemble network, it also degrades the grammaticality of the generated results. We set α to 1.0 after determining that it yielded the best performance. This result clearly indicates that our ensemble network contributes to the accuracy of the generated answers.

A comparison of our full method *NAGM* with the one without the sentence-type embedding (we call this method *w/o ste*) that trains separate decoders for two types of sentences is shown in Table 2. The result indicated that the existence of the sentence type vector, **C** or **S**, contributes the accuracy of the results since it distinguishes between sentence types.

Results

Performance The results for Oshiete-goo are shown in Table 3 and those for nfL6 are shown in Table 4. They show that *CLSTM* is better than *Seq2seq*. This is because it incorporates contextual features, i.e. topics, and thus can generate answers that track the question’s context. *Trans* is also better than *Seq2seq*, since it uses attention from the question to the conclusion or supplement more effectively than *Seq2seq*. *HRED* failed to attain a reasonable level of performance. These results indicate that sequential generation has difficulty generating subsequent statements that follow the original meaning of the first statement (question).

NAGMWA is much better than the other methods except *NAGM*, since it generates answers whose conclusions and supplements as well as their combinations closely match the questions. Thus, conclusions and supplements in the answers are consistent with each other and avoid confusion made by several different conclusion-supplement an-

Table 5: Human evaluation (Oshiete-goo).

<i>CLSTM</i>				<i>NAGM</i>			
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
21	18	27	34	47	32	11	10

Table 6: Human evaluation (nfL6).

<i>CLSTM</i>				<i>NAGM</i>			
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
30	3	27	40	50	23	16	11

swers assigned to a single non-factoid questions. Finally, *NAGM* is consistently superior to the conventional attentive encoder-decoders regardless of the metric. Its ROUGE-L and BLEU-4 scores are much higher than those of *CLSTM*. Thus, *NAGM* generates more fluent sentences by assessing the context from conclusion to supplement sentences in addition to the closeness of the question and sentences as well as that of the question and sentence combinations.

Human evaluation Following evaluations made by crowdsourced evaluators (Li et al. 2016), we conducted human evaluations to judge the outputs of *CLSTM* and those of *NAGM*. Different from (Li et al. 2016), we hired human experts who had experience in Oshiete-goo QA community service. Thus, they were familiar with the sorts of answers provided by and to the QA community.

The experts asked questions, which were not included in our training datasets, to the AI system and rated the answers; one answer per question. The experts rated the answers as follows: (1) the content of the answer matched the question, and the grammar was okay; (2) the content was suitable, but the grammar was poor; (3) the content was not suitable, but the grammar was okay; (4) both the content and grammar were poor. Note that our evaluation followed the DUC-style strategy⁵. Here, we mean “grammar” to cover grammaticality, non-redundancy, and referential clarity in the DUC strategy, whereas we mean the “content matched the questions” to refer to “focus” and “structure and coherence” in the DUC strategy. The evaluators were given more than a week to carefully evaluate the generated answers, so we consider that their judgments are reliable. Each expert evaluated 50 questions. We combined the scores of the experts by summing them. They did not know the identity of the system in the evaluation and reached their decisions independently.

Table 5 and Table 6 present the results. The numbers are percentages. Table 7 presents examples of questions and answers. For Oshiete-goo results, the original Japanese and translated English are presented. The questions are very long and include long background descriptions before the questions themselves.

These results indicate that the experts were much more satisfied with the outputs of *NAGM* than those of *CLSTM*. This is because, as can be seen in Table 7, *NAGM* generated longer and better question-related sentences than *CLSTM* did. *NAGM* generated grammatically good answers whose

⁵<http://www-nlpir.nist.gov/projects/duc/duc2007/quality-questions.txt>

Table 7: Example answers generated by *CLSTM* and *NAGM*. #1 is for Oshiete-goo and #2 for nL6.

ID	Questions	Answers generated by <i>CLSTM</i>	Answers generated by <i>NAGM</i>
1	Many people say that a quarrel is one example of a farewell in a love affair. Recently we also often quarrel with each other. But within a few minutes we are reconciled and even laughing about what we quarreled about. Are such "frequent quarrels" good things to maintain the relationship between us? Or should we stop quarreling with each other so as not to farewell?	Why not ask somehow to tell your feelings to your lover. He/she may be able to get to know you. (相手に気持ちを伝えることを、それとなく聞いてみてはいかがでしょうか。相手にも、あなたのことを知ってもらえるようになるかもしれません。)	I think that it is better not to do anything. You should think positively about it without thinking too pessimistically. I think that it is important to communicate the feelings to your lover; how you like or care about him/her especially when you are quarreling with each other. (何もしないほうが良いと思います。あまり悲観的な想像ばかりせず、前向きに考えてみてはいかがでしょうか。ケンカをした時こそ好きな気持ちや相手を大切に思う気持ちを伝えることが大事なのだと思いますよ。)
2	How I can find my love?	Feed him lots of fresh milk and change his diaper often. God hears prayers that are sincere.	Well, in my opinion, u can't get a lover without any effort...u gotta do smth! Otherwise don't care for the results, just go ahead and do it.

conclusion and supplement statements are well matched with the question and the supplement statement naturally follows the conclusion statement.

Generating answers missing from the corpus The encoder-decoder network tends to re-generate answers in the training corpus. On the other hand, *NAGM* can generate answers not present in the corpus by virtue of its ensemble network that considers contexts and sentence combinations.

Table 7 lists some examples. For example, answer #1 generated by *NAGM* is not in the training corpus. We think it was generated from the parts in italics in the following three sentences that are in the corpus: (1) "*I think that it is better not to do anything* from your side. If there is no reaction from him, it is better not to do anything even if there is opportunity to meet him next." (2) "I think it may be good for you to approach your lover. *Why don't you think positively about it without thinking too pessimistically?*" (3) "Why don't you tell your lover that you usually do not say what you are thinking. . . . *I think that it is important to communicate the feelings to your lover; how you like or care about him/her especially when you are quarreling with each other.*"

The generation of new answers is important for non-factoid answer systems, since they must cope with slight differences in question contexts from those in the corpus.

Online evaluation in "Love Advice" service Our ensemble network is currently being used in the love advice service of Oshiete goo (Nakatsuji 2018). The service uses only the ensemble network to ensure that the service offers high-quality output free from grammar errors. We input the sequences in our evaluation corpus instead of the decoder output sequences into the ensemble network. Our ensemble network then learned the optimum combination of answer sequences as well as the closeness of the question and those sequences. As a result, it can construct an answer that corresponds to the situation underlying the question. In particular, 5,702 answers created by the AI, whose name is Oshi-el (Oshi-el means teaching angel), using *our ensemble network* in reply to 33,062 questions entered from September 6th, 2016 to November 17th, 2019, were judged by users of

the service as *good answers*. Oshi-el output good answers at about twice the rate of the average human responder in Oshiete-goo who answered more than 100 questions in the love advice category. Thus, we think this is a good result.

Furthermore, to evaluate the effectiveness of the supplemental information, we prepared 100 answers that only contained conclusion sentences during the same period of time. As a result, users rated the answers that contained both conclusion and supplement sentences as good 1.6 times more often than those that contained only conclusion sentences. This shows that our method successfully incorporated supplemental information in answering non-factoid questions.

Conclusion

We tackled the problem of conclusion-supplement answer generation for non-factoid questions, an important task in NLP. We presented an architecture, ensemble network, that uses an attention mechanism to reflect *the context* of the conclusion decoder's output sequence on the supplement decoder's output sequence. The ensemble network also assesses *the closeness* of the encoder input sequence to the output of each decoder and the combined output sequences of both decoders. Evaluations showed that our architecture was consistently superior to conventional encoder-decoders in this task. The ensemble network is now being used in the "Love Advice," service as mentioned in the Evaluation section.

Furthermore, our method, *NAGM*, can be generalized to generate much longer descriptions other than conclusion-supplement answers. For example, it is being used to generate Tanka, which is a genre of classical Japanese poetry that consists of five lines of words⁶, in the following way. The first line is input by a human user to *NAGM* as a question, and *NAGM* generates second line (like a conclusion) and third line (like a supplement). The third line is again input to *NAGM* as a question, and *NAGM* generates the fourth line (like a conclusion) and fifth line (like a supplement).

⁶<https://www.tankakenkyu.co.jp/ai/>

References

- [Bahdanau, Cho, and Bengio 2014] Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.
- [dos Santos et al. 2015] dos Santos, C.; Barbosa, L.; Bogdanova, D.; and Zadrozny, B. 2015. Learning hybrid representations to retrieve semantically equivalent questions. In *Proc. ACL-IJCNLP'15*, 694–699.
- [Ennis 1991] Ennis, R. 1991. Critical thinking: A streamlined conception. In *Teaching philosophy*, 5–25.
- [Ghosh et al. 2016] Ghosh, S.; Vinyals, O.; Strophe, B.; Roy, S.; Dean, T.; and Heck, L. 2016. Contextual LSTM (CLSTM) models for large scale NLP tasks. *CoRR* abs/1602.06291.
- [Jia and Liang 2017] Jia, R., and Liang, P. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proc. EMNLP'17*, 2021–2031.
- [Joshi et al. 2017] Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *CoRR* abs/1705.03551.
- [Li et al. 2016] Li, J.; Monroe, W.; Ritter, A.; Jurafsky, D.; Galley, M.; and Gao, J. 2016. Deep reinforcement learning for dialogue generation. In *Proc. EMNLP'16*, 1192–1202.
- [Lin 2004] Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: In: Proc. ACL-04 Workshop*, 74–81.
- [Liu, Duh, and Gao 2018] Liu, X.; Duh, K.; and Gao, J. 2018. Stochastic answer networks for natural language inference. *CoRR* abs/1804.07888.
- [Nakatsuji 2018] Nakatsuji, M. 2018. Can ai generate love advice? neural conclusion-supplement answer construction for non-factoid questions. on-demand.gputechconf.com/gtc/2018/video/S8301/. In *GPU Technology Conference 2018 San Jose, CA*.
- [Nguyen et al. 2016] Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with NIPS 2016*.
- [Nogueira et al. 2019] Nogueira, R.; Yang, W.; Lin, J.; and Cho, K. 2019. Document expansion by query prediction. *CoRR* abs/1904.08375.
- [Papineni et al. 2002] Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proc. ACL'02*, 311–318.
- [Qiu and Huang 2015] Qiu, X., and Huang, X. 2015. Convolutional neural tensor network architecture for community-based question answering. In *Proc. IJCAI'15*, 1305–1311.
- [Rajpurkar et al. 2016] Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. *CoRR* abs/1606.05250.
- [Rinott et al. 2015] Rinott, R.; Dankin, L.; Perez, C. A.; Khapra, M. M.; Aharoni, E.; and Slonim, N. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proc. EMNLP'15*, 440–450.
- [Serban et al. 2016] Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proc. AAAI'16*, 3776–3784.
- [Shao et al. 2019] Shao, T.; Guo, Y.; Hao, Z.; and Chen, H. 2019. Transformer-based neural network for answer selection in question answering. *IEEE Access* PP:1–1.
- [Song et al. 2017] Song, H.; Ren, Z.; Liang, S.; Li, P.; Ma, J.; and de Rijke, M. 2017. Summarizing answers in non-factoid community question-answering. In *Proc. WSDM '17*, 405–414.
- [Sutskever, Vinyals, and Le 2014] Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Proc. NIPS'14*, 3104–3112.
- [Tan et al. 2016] Tan, M.; dos Santos, C. N.; Xiang, B.; and Zhou, B. 2016. Improved representation learning for question answer matching. In *Proc. ACL'16*, 464–473.
- [Tan et al. 2017] Tan, C.; Wei, F.; Yang, N.; Lv, W.; and Zhou, M. 2017. S-net: From answer extraction to answer generation for machine reading comprehension. *CoRR* abs/1706.04815.
- [Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need.
- [Vinyals and Le 2015] Vinyals, O., and Le, Q. V. 2015. A neural conversational model. *CoRR* abs/1506.05869.
- [Wang et al. 2018] Wang, Y.; Liu, K.; Liu, J.; He, W.; Lyu, Y.; Wu, H.; Li, S.; and Wang, H. 2018. Multi-passage machine reading comprehension with cross-passage answer verification. In *Proc. ACL'18*, 1918–1927.
- [Wang, Smith, and Mitamura 2007] Wang, M.; Smith, N. A.; and Mitamura, T. 2007. What is the jeopardy model? a quasi-synchronous grammar for qa. In *Proc. EMNLP-CoNLL'07*, 22–32.
- [Yang et al. 2016] Yang, Z.; Yuan, Y.; Wu, Y.; Cohen, W. W.; and Salakhutdinov, R. 2016. Review networks for caption generation. In *Proc. NIPS'16*, 2361–2369.
- [Yang, Yih, and Meek 2015] Yang, Y.; Yih, W.; and Meek, C. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proc. EMNLP'15*, 2013–2018.
- [Yin et al. 2016] Yin, J.; Jiang, X.; Lu, Z.; Shang, L.; Li, H.; and Li, X. 2016. Neural generative question answering. In *Proc. IJCAI'16*, 2972–2978.
- [Yu et al. 2014] Yu, L.; Hermann, K. M.; Blunsom, P.; and Pulman, S. 2014. Deep learning for answer sentence selection. *CoRR* abs/1412.1632.
- [Yu et al. 2018] Yu, A. W.; Dohan, D.; Luong, M.-T.; Zhao, R.; Chen, K.; Norouzi, M.; and Le, Q. V. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *Proc. ICLR'18*.