Report on Japanese subtask for NTCIR-13 STC-2 from mnmlb

Sotaro Takeshita
The University of Electro-Communications
takeshita@sd.is.uec.ac.jp

Ryuji Tamaki
The University of Electro-Communications
tamaki@sd.is.uec.ac.jp

Yasuhiro Minami
The University of Electro-Communications
minami.yasuhiro@is.uec.ac.jp

Takeru Kazama
The University of Electro-Communications
kazama@sd.is.uec.ac.jp

Masato Nakamura
The University of Electro-Communications
nakamura@sd.is.uec.ac.jp

ABSTRACT

This paper reports a Japanese subtask for NTCIR-13 STC-2 for which we made a dialogue system and introduced neural network-based retrieval models (LSTM, ESIM and CNN) to rank the dialogue replies in the training dataset. We used data from Yahoo! News comments data and introduced LSTM and ESIM to effectively capture sequential information from the given comments. To evaluate the effectiveness, we compared systems using LSTM or ESIM with systems that use CNN. We also introduced an n-gram-based statistical filter into our systems to reduce the number of reply candidates.

Team Name

mnmlb

Subtasks

NTCIR-13 STC Japanese Subtask

Keywords

neural network conversation dual encoder

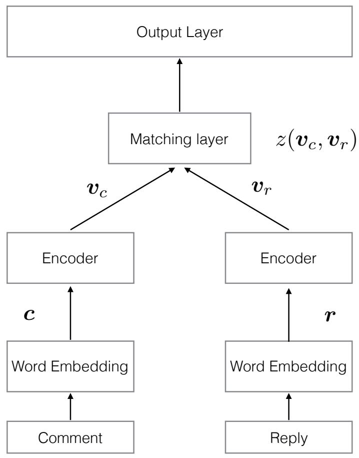

Figure 1: Model Architecture

1. INTRODUCTION

For natural interactions between humans and computers, the importance of dialog systems that can treat natural conversations will probably continue to increase. Recently such systems are attracting much attention in the natural language processing field because researchers can use a great deal of conversational data obtained by the recording logs of micro-blogging services and chat applications to train these systems. Although we have many training datasets, building a dialog system is challenging due to the difficulty of identifying the meaning of ambiguous dialogues.

This competition is one trial to create dialogue systems using a large amount of training and test data and their evaluations. The training data contain 894,998 comment and reply pairs with which we trained our models. In the test phase, we were given a set of 100 test comments to generate a reply for each comment.

For this competition, we built five retrieval-based systems that select a reply that maximizes the score function for a given comment. For the scoring functions, we used neural networks that predict whether a given comment-reply tuple is a correct combination. After training the neural networks, we scored all the combinations of the given test comments and replies in the training data and chose the reply candidate with the highest score.

We found two problems in our preliminary experiments. First, since the number of reply candidates is around one million, an enormous amount of time is required to calculate their scores. Second, some reply candidates contain inappropriate words, for example, violent or sexual words. To solve these two problems, we implemented filters with simple rules and trigram-based ranking.

2. SYSTEM ARCHITECTURE

2.1 Ranking Model

Figure 1 shows that for our scoring model, we employed the framework of a dual encoder [8] that encodes two sentences with a neural network in the encoder. We calculated the degree of interaction between a given comment and a reply. The comments and replies are encoded in a parallel encoder layer, merged, and converted into a single vector in the matching layer. Next the vector is mapped into a single value using multilayer perceptrons (MLPs) that are expected to learn the degree of interaction between the comment and the reply. Finally, the range of the value is restricted from zero to one by a sigmoid function so that it can be treated as a probability value. In the encoder layer, the comments and replies are encoded using the following equations:

$$\boldsymbol{v}_c = \text{Encoder}(\boldsymbol{c}) \quad (1)$$

and

$$\boldsymbol{v}_r = \text{Encoder}(\boldsymbol{r}), \quad (2)$$

where $\boldsymbol{c}$, and $\boldsymbol{r}$ are respectively the embedded words of each comment and a reply. The encoders for $\boldsymbol{c}$ and $\boldsymbol{r}$ of all the dual encoders in this paper share their weights. In the final MLPs, the probability value is calculated by

$$p(\text{flag} = 1|z(\boldsymbol{v}_c, \boldsymbol{v}_r)) = \sigma(\boldsymbol{\theta}^{\mathrm{T}} z(\boldsymbol{v}_c, \boldsymbol{v}_r) + b) \quad (3)$$

and

$$\mathcal{L} = -\Sigma \log p(\text{flag} = 1|z(\boldsymbol{v}_c, \boldsymbol{v}_r)), \quad (4)$$

where $\boldsymbol{v}_c$ and $\boldsymbol{v}_r$ respectively denote the embedded vectors for the comments and replies. To concatenate the two vectors of the comment and the reply in the matching layer, we used three equations:

$$z(\boldsymbol{v}_c, \boldsymbol{v}_r) = [\boldsymbol{v}_c \ ; \ \boldsymbol{v}_r], \quad (5)$$

$$z(\boldsymbol{v}_c, \boldsymbol{v}_r) = [|\boldsymbol{v}_c - \boldsymbol{v}_r| \ ; \ \boldsymbol{v}_c \odot \boldsymbol{v}_r] \quad (6)$$

and

$$z(\boldsymbol{v}_c, \boldsymbol{v}_r) = \boldsymbol{v}_c \cdot \boldsymbol{v}_r, \quad (7)$$

where ;, $\odot$, and $\cdot$ denote the vector concatenation, the element-wise product, and the inner product. For encoding and output, we investigated the following three neural network architectures:

1. vanilla LSTM [3],
2. LSTM with attention mechanism (ESIM) [2],
3. Convolutional Neural Network [6].

Their details are discussed in the next section.

2.1.1 LSTM

Long short-term memory neural networks (LSTMs) are one type of recurrent neural network (RNN) that model sequential data. They reduce the vanishing gradient problem possessed by RNNs by replacing their hidden layer units in LSTM blocks. LSTMs individually encode each comment and reply to calculate the input values of the matching layer. The comment and reply vectors were combined using Eq. **??**. Then the loss was calculated using Eq. 4.

2.1.2 ESIM

ESIM, which is previously proposed attention-based architecture [2], uses BiLSTM [7] that performs two LSTMs in two different ways. First, LSTM takes input from the previous layer forward, provides output to the second LSTM, and encodes it backward.

In our model, we individually used a BiLSTM to encode the comments and the replies. Here we denote their outputs as $\overline{a}_i$ and $\overline{b}_j$. The attention mechanism that reflects the information of the comments and the replies can be written as:

$$e_{ij} = \overline{\boldsymbol{a}}_i^{\mathrm{T}} \overline{\boldsymbol{b}}_j, \quad (8)$$

$$\boldsymbol{v}_c = \Sigma_{j=1}^{l_a} \frac{\exp(e_{ij})}{\Sigma_{k=1}^{l_b} \exp(e_{ik})} \overline{\boldsymbol{b}}_j, \quad \forall_i \in [1, \cdots, l_a] \quad (9)$$

and

$$\boldsymbol{v}_r = \Sigma_{i=1}^{l_b} \frac{\exp(e_{ij})}{\Sigma_{k=1}^{l_a} \exp(e_{kj})} \overline{\boldsymbol{a}}_i, \quad \forall_j \in [1, \cdots, l_b], \quad (10)$$

where $\boldsymbol{v}_c$ and $\boldsymbol{v}_r$, were combined by Eq. 5. Then $\boldsymbol{z}$ is fed into a multilayer perceptron (MLP) classifier with three layers and a sigmoid function in the output layer. Finally, Eq. 4 is used to calculate the loss.

2.1.3 CNN

CNNs are a type of a neural network that achieved huge success in image recognition tasks. They are also used in the natural language processing field and have achieved novel results in document classification tasks. In our system, we used CNNs that were cited in a previous work [4]. We used CNNs with six different kernel sizes (1, 2, 3, 4, 5, 6) to extract different granularity information. After the convolutional layer, we used global average pooling to choose the critical information as the shape of the scalar value. All of the scalar values for each comment and reply from each CNN were combined using Eq. 6. Finally, we obtained the concatenated vector through the output layer.

2.2 Candidates Filtering

In addition to neural network models, we implemented filters for obtaining reply candidates to address the problems we mentioned in the introduction. We denote the problems here again:

- Number of replies is too large and requires enormous processing time.
- Reply candidates contain some obviously inappropriate expressions.

To solve these problems, we implemented the following two filters:

- one that removes the candidates that contain the same word more than twice.
- another that removes the candidates that contain more than two sentences.

After reducing the reply candidates using the above two simple rules, we also used a trigram occurrence frequency-based method to further reduce candidates. This method argues that most regular Japanese sentences tend to have fixed forms at their ends and sentences with a fixed form are more appropriate as replies for the given comments. This method uses the following two steps. In the first step, the trigrams at the end of all the sentences are sorted in descending order of their occurrence frequency. In the second step, all the sentences that have a trigram in the top 100 trigrams at the end of the sentences are sorted in descending order of their scores, where each sentence ’ s score is calculated by summing up the probabilities of all the trigrams and dividing them by the number of trigrams in the sentence. For each trigram in the top 100, we selected the 100 top sentences and obtained 10,000 candidate sentences.

3. EXPERIMENTS

The distributed dataset contains 894,998 comment and reply pairs that were divided into 794,998 training and 5000 validation data.

For this task, the best sentence that fit the given comment is selected from the training data. We designed a binary classification system that checks whether a reply fits the given comments. This system requires both negative and positive examples. However, the training data do not have negative data. We randomly sampled the replies from all the replies in the training data for every comment to make negative examples and concatenated with the original positive examples. Using this concatenated training data, we trained neural network models with learning binary classification systems. In the test phase, the neural network models select the appropriate reply for the given comment.

To separate Japanese sentences into words, we used software MeCab and neologd for the dictionary and set the number of LSTM hidden units to 200 in the ESIM model. Both LSTMs have 300 hidden units. For the output layer, we used a three-layer MLP whose layers have 300, 300, and 1 units, respectively. The dropout rate was set to 0.5. For the CNN model, we used two MLPs whose layers have 300 and 1 units.

3.1 Settings

For all the neural network model settings, we used cross entropy for the loss function. Adam [5] with a learning rate of 0.001 maximized the loss function. All embedding layers were initialized by the parameters of fastText [1] and trained on the same given dataset.

3.2 Result

All the generated replies are evaluated and labeled as L0, L1, or L2 by Rules 1 and 2 using Algorithms 1 and 2. The evaluation results are given in Tables 2 and 2. All the replies are scored zero to one with the labels from [9]. Labels L0, L1, and L2 were scored as 0, 1, and 3 and used to calculate the following scores: nG@1, nERR@2, and AccG@k ([9]). Table 5 shows some reply results generated by ESIM without filtering, which scored the best in our submitted runs. Although this is the best system, some replies are very long or contain some aggressive phraseologies since this model does not implement a filter.

Algorithm 1 Rule-1

```
if fluent & coherent = L1 then
  if context-dependent & informative = L2 then
    return L2
  else
    return L1
  end if
else
  return L0
end if
```

Table 1: Official STC results Rule 1

Model	Mean $nG@1$	Mean $nERR@2$
ESIM with filter	0.2949	0.3463
ESIM without filter	0.3690	0.4410
LSTM with filter	0.2230	0.2538
LSTM without filter	0.2584	0.2799
CNN with filter	0.2544	0.3066
Model	Mean $Acc_{L2}@1$	$Acc_{L2}@2$
ESIM with filter	0.0700	0.0710
ESIM without filter	0.1040	0.1210
LSTM with filter	0.0560	0.0450
LSTM without filter	0.0940	0.0750
CNN with filter	0.0680	0.0690
Model	Mean $Acc_{L1,L2}@1$	$Acc_{L1,L2}@2$
ESIM with filter	0.5400	0.5360
ESIM without filter	0.6540	0.6600
LSTM with filter	0.4020	0.3930
LSTM without filter	0.4200	0.3800
CNN with filter	0.4520	0.4640

Table 2: Official STC results Rule 2

Model	Mean $nG@1$	Mean $nERR@2$
ESIM with filter	0.2518	0.2829
ESIM without filter	0.3144	0.3804
LSTM with filter	0.1792	0.2018
LSTM without filter	0.2212	0.2415
CNN with filter	0.2144	0.2573
Model	Mean $Acc_{L2}@1$	$Acc_{L2}@2$
ESIM with filter	0.0700	0.0710
ESIM without filter	0.1040	0.1210
LSTM with filter	0.0560	0.0450
LSTM without filter	0.0940	0.0750
CNN with filter	0.0680	0.0690
Model	Mean $Acc_{L1,L2}@1$	$Acc_{L1,L2}@2$
ESIM with filter	0.4380	0.3880
ESIM without filter	0.5300	0.5360
LSTM with filter	0.3020	0.2910
LSTM without filter	0.3380	0.3050
CNN with filter	0.3600	0.3550

3.2.1 Comparing ESIM and CNN

To compare the results of ESIM with a filter and CNN with a filter, the examples are listed in Table 3 whose ESIM scores are 1.0 (maximum score) and CNN scores are 0.0 (minimum score). We confirm that ESIM refers more properly to important words in the comments than the CNN model.

3.2.2 Filtering

In this section, we analyze why candidate filtering was ineffective. One reason is that the filter summed up (Eq. 11) the trigram occurrence frequency instead of multiplying trigram probabilities (Eq. 12). This reduced the score of the short replies that contain rare trigrams more than those of long ones with rare trigrams. In other words, when comments were given with a proper noun, the model produced a reply with the same proper noun but with a long context. Since such replies tend to be too specific, the annotators might evaluate them with low scores:

Algorithm 2 Rule-2

```
if fluent & coherent = L1 then
  if context-dependent & informative = L2 then
    return L2
  else if context-dependent or informative = L0 then
    return L0
  else
    return L1
  end if
else
  return L0
end if
```

$$score = \sum_{i=0}^{l} p(TriGram_i) \tag{11}$$

$$score = \prod_{i=0}^{l} p(TriGram_i). \tag{12}$$

4. CONCLUSIONS

This paper reports the result of our methods for Japanese subtaskfor NTCIR-13 STC-2. We investigated Neural Network based reply generating systems with simple rule and tri-gram based reply candidates filtering. With the filtering methods, we achieved reducing the number of the reply candidates to calculate the matching score.

5. ADDITIONAL AUTHORS

6. REFERENCES

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016.

[2] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. Enhancing and combining sequential and tree LSTM for natural language inference. *CoRR*, abs/1609.06038, 2016.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[4] Yoon Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[6] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, R. E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten digit recognition with a back-propagation network. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 396–404. Morgan-Kaufmann, 1990.

[7] Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. Learning natural language inference using bidirectional LSTM model and inner-attention. *CoRR*, abs/1605.09090, 2016.

[8] Ryan Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. Training end-to-end dialogue systems with theubuntu dialogue corpus. *Dialogue and Discourse*, 10.5087/dad.2017.102, 2017.

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

Table 3: Compare ESIM with filter, CNN with filer

Comment	ESIM resply	ESIM score	CNN reply	CNN score
サブカルチャー の 配信 は 歓迎 。 囲碁 で も モノポリー で も 何 でも 見 た い人 は いる だろ う し 。 ビデオゲーム の 大会 も 配信 し たら いい ん じゃ ない か な 。	アジア を 始め 、 世界各国 に 配信 する ん だ と 思う よ	1.0	バルサ に 出し すぎ た か ？	0.0
釈放 って 、 何 だっ た ん だ 、 結局 。 本人 が 否定 し て い た 通り 、 やっ て い なかっ た の か ???	否定 も し て ない けど ね	1.0	カミナリ 良かっ た よ ね	0.0
7 か国 以外 で も 危ない 国 が ある はず 。 どうも 納得 でき ない 。	あれ は 危ない よ ね 1	1.0	被疑者 死亡 だ から 起訴 で き ない ん です よ	0.0
言う べき 事 は 言っ て ほしい けど 、 あまり 真剣 に 相手 し ない ほう が 良い かも ね 。 振り回さ れる ほう が 迷惑 だ し 。	迷惑 に なる よう な こと は しゃべっ て ない と 思う け ど	1.0	薬物 絡み は スポンサー や 共演者 など に 迷惑 掛かる か ら 無視 出来 ない でしょ う	0.0

Table 4: ESIM with filter

Comment	Submitted reply	score
柴崎 って 誰 ？ 知り ませ ん が … 。	まぁわかる柴崎の努力次第だろうな	0.8
失う もの が あっ た として も 、 自分 の 根 っ子 の 部分 は 曲げ ない 。 大事 な 事 だ と 感じる 。	相葉ちゃんはグッとスポーツで活躍しているから抜擢されたんですよ	0.0
こういう 記事 に 否定 的 な 意見 を 言う 人 に は うんざり 。 是非 と も 頑張っ て もらい たい 。	もっともな意見だと思います！	1.0
好き な ブランド の 一つ だ けど 、 最近 、 気品 が 感じ られ ない 。	今は FULL BKってブランドやってますよ！	0.6
本当に そう 思う なら 、 もっと 早い 段階 に 、 自分 の 言葉 で SMAP に対する 思い を 語っ て ほしかっ た 。 そ れ は 、 他 の ジャニーズメンバー に対して も 同じ 。	そもそも、SMAP 愛してなかったら 30 年も続けないと思いますよ	0.6

Table 5: ESIM without filter

Comment	Submitted reply	score
柴崎 って 誰 ？ 知り ませ ん が … 。	柴崎は？	1.0
失う もの が あっ た として も 、 自分 の 根 っ子 の 部分 は 曲げ ない 。 大事 な 事 だ と 感じる 。	そもそも 、 いじられキャラ なん ていう 括り が おかしい 。 芸人 で あれ 誰 で あれ 他人 が 馬鹿 に さ れ てる の 見 て 面白 え の か ？ ドッキリ とか モニタリング なんて の も 一緒 。 根 っ子 の 部分 で は イジメ と 同じ 。 漫才 とか 落語 等 の 芸 で 笑い を 取る の が 基本 だろ う よ 。	0.0
こういう 記事 に 否定 的 な 意見 を 言う 人 に は うんざり 。 是非 と も 頑張っ て もらい たい 。	否定 的 な 意見 が アホ な 人 と は 限ら ない ！	1.0
好き な ブランド の 一つ だ けど 、 最近 、 気品 が 感じ られ ない 。	気品 は あっ た けど	1.0
本当に そう 思う なら 、 もっと 早い 段階 に 、 自分 の 言葉 で SMAP に対する 思い を 語っ て ほしかっ た 。 そ れ は 、 他 の ジャニーズメンバー に対して も 同じ 。	早い 段階 で SMAP 出 ない 事 は 分かっ て た はず です が …	0.8
ロマン が あり ます ね 今 と なっ て は 憶測 で ない と わから ない の が 残念 です あと この 方 が ずっと もつ べ き だっ た の は 手榴弾 で は なく 、 白球 だっ た と 思い ます	最近 全く 同じ 書き込み を 見 まし た 。 何度 も 同じ 事 を 載せ て 悪意 を 感じ ます 。 何 の 目的 が 有っ て 繰り返し やっ てる の ？ そんなに 木村拓哉 が 憎い ん です か ？ いい加減 な 記者 の 言葉 や 憶測 で 言っ て いる 人 の 事 を 信じ て 、 子供 じみ てる と 思い ませ ん か 。 いい 大人 な の だ から もつ と 分別 し た 行動 を お勧め し ます 。	0.0
見 て み ない と 分から ない それ より 立花 宗重 なんか やっ たら 面白い かも 島津義弘 なんか も いい かも	立花 いる し w	0.8
もうこ の 人 は プロ意識 に 欠け てる 。	もうこ の 人 の 下手くそ な 文章 に 何 も 期待 し て い ませ ん 。	0.0
やっぱ 山陽電鉄 や ！ ！	やっぱ クリロナベイル でしょ ！ ！	0.0
この 人 の 人間性 や これ まで の 悪態 ぶり を 思う と 素直 に 応援 する 気 に は なれ ない 。	素直 に 応援 し たい 。	1.0

Table 6: CNN with filter

Comment	Submitted reply	score
柴崎 って 誰 ? 知り ませ ん が … 。	柴崎 の 2 点 目 とか クルトワ なら 止め てる と 思う よ	0.8
失う もの が あっ た として も 、 自分 の 根 っ子 の 部分 は 曲げ ない 。 大事 な 事 だ と 感じる 。	共産党 が 絡ま なきゃ 根 は 悪い 人たち じゃ ない の か な	0.6
こういう 記事 に 否定 的 な 意見 を 言う 人 に は うんざ り 。 是非 と も 頑張っ て もらい たい 。	こういう 意見 今後 どんどん 増え て いく と 思う よ	1.0
好き な ブランド の 一つ だ けど 、 最近 、 気品 が 感じ られ ない 。	好き な ん でしょ う ね	0.8
本当に そう 思う なら 、 もっと 早い 段階 に 、 自分 の 言葉 で SMAP に対する 思い を 語っ て ほしかっ た 。 そ れ は 、 他 の ジャニーズメンバー に対して も 同じ 。	もう 解散 し た のに 素晴らしい 未来 は 無い でしょ う ?	0.8