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

Yasuhirou 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

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:

$$v_c = \text{Encoder}(c)$$

and

$$v_r = \text{Encoder}(r),$$

where $c$ and $r$ are respectively the embedded words of each comment and a reply. The encoders for $c$ and $r$ are encoded using the following equations:

$$v_c = \text{Encoder}(c)$$

and

$$v_r = \text{Encoder}(r),$$

where $c$ and $r$ respectively denote the embedded vectors of the comment and the reply. In the final MLPs, the probability value is calculated by

$$p(\text{flag} = 1|z(v_c, v_r)) = \sigma(\theta^T z(v_c, v_r) + b)$$

and

$$\mathcal{L} = -\sum \log p(\text{flag} = 1|z(v_c, v_r)),$$

where $v_c$ and $v_r$ 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:

$$z(v_c, v_r) = [v_c ; v_r],$$

$$z(v_c, v_r) = [v_c - v_r ; v_c \odot v_r]$$

and

$$z(v_c, v_r) = v_c \cdot v_r,$$

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. 5.

$$z = \sigma(\theta^T x + b)$$

where $x$ denotes the input vector, and $\sigma$ is the sigmoid function. 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 $\pi_i$, and $b_j$. The attention mechanism that reflects the information of the comments and the replies can be written as:

$$e_{ij} = \pi_i^T b_j,$$

$$v_c = \sum_{j=1}^{l_c} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_c} \exp(e_{ik})} b_j, \quad \forall i \in [1, \ldots, l_c]$$

and

$$v_r = \sum_{j=1}^{l_r} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_r} \exp(e_{kj})} a_j, \quad \forall j \in [1, \ldots, l_r],$$

where $v_c$ and $v_r$ were combined by Eq. 5. Then $z$ 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 top 100 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.

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
else
    return L0
end if
```

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:

### Table 1: Official STC results Rule 1

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean nG@1</th>
<th>Mean nERR@2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESIM with filter</td>
<td>0.2949</td>
<td>0.3463</td>
</tr>
<tr>
<td>ESIM without filter</td>
<td>0.3690</td>
<td>0.4410</td>
</tr>
<tr>
<td>LSTM with filter</td>
<td>0.2230</td>
<td>0.2538</td>
</tr>
<tr>
<td>LSTM without filter</td>
<td>0.2584</td>
<td>0.2799</td>
</tr>
<tr>
<td>CNN with filter</td>
<td>0.2544</td>
<td>0.3066</td>
</tr>
</tbody>
</table>

### Table 2: Official STC results Rule 2

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean nG@1</th>
<th>Mean nERR@2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESIM with filter</td>
<td>0.5400</td>
<td>0.5360</td>
</tr>
<tr>
<td>ESIM without filter</td>
<td>0.6540</td>
<td>0.6600</td>
</tr>
<tr>
<td>LSTM with filter</td>
<td>0.4020</td>
<td>0.3939</td>
</tr>
<tr>
<td>LSTM without filter</td>
<td>0.4260</td>
<td>0.3889</td>
</tr>
<tr>
<td>CNN with filter</td>
<td>0.4520</td>
<td>0.4640</td>
</tr>
</tbody>
</table>
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

\[
\text{score} = \sum_{i=0}^{1} p(\text{TriGram}_i) \quad (11)
\]

\[
\text{score} = \prod_{i=0}^{1} p(\text{TriGram}_i). \quad (12)
\]

4. CONCLUSIONS

This paper reports the result of our methods for Japanese subtask for 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


### Table 3: Compare ESIM with filter, CNN with filer

<table>
<thead>
<tr>
<th>Comment</th>
<th>ESIM reply</th>
<th>ESIM score</th>
<th>CNN reply</th>
<th>CNN score</th>
</tr>
</thead>
<tbody>
<tr>
<td>サブカルチャーの配信は歓迎。開催でもモノポリーでも何でも見える人はいるだろうし。ビデオゲームの大会も配信したらいいんじゃないかな。</td>
<td>アジアを始め、世界各国に配信するんだけどと思うよ</td>
<td>1.0</td>
<td>パルサに出すすぎたか？</td>
<td>0.0</td>
</tr>
<tr>
<td>押えって、何だったんだ、結局。本人が否定していた通り、やっていたかったのか？??</td>
<td>否定していないけどね</td>
<td>1.0</td>
<td>カミナリ良かったよね</td>
<td>0.0</td>
</tr>
<tr>
<td>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。</td>
<td>あれば危ないよね！</td>
<td>1.0</td>
<td>被疑者死亡だから起訴できないよ</td>
<td>0.0</td>
</tr>
<tr>
<td>言葉の事は、言ってはいいけど、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。</td>
<td>遊びになるようなことはしゃべってないとと思うけど</td>
<td>1.0</td>
<td>裸の嘘みたいスポンサーや共演者などに迷惑かけておらね</td>
<td>0.0</td>
</tr>
</tbody>
</table>

### Table 4: ESIM with filter

<table>
<thead>
<tr>
<th>Comment</th>
<th>Submitted reply</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>楽しくって誰？知りませんが…</td>
<td>まあわかる楽観の努力失敗だろうな</td>
<td>0.8</td>
</tr>
<tr>
<td>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。</td>
<td>相変わらずクツとスポーツで活躍しているから抜群されたんだよ</td>
<td>0.0</td>
</tr>
<tr>
<td>こういう記事に否定的な意見を言う人にはうんざり。是非とも頭張ってもらいたい。</td>
<td>もっとも意見だと思います！</td>
<td>1.0</td>
</tr>
<tr>
<td>お互いに相違考えながら、もっと早め段階に、自分の言葉でSMAPに対する思いを語ってはよかった。それなら、他のジャニーズメンバーに対しても同じ。</td>
<td>今はFULL BKってブランドやってますよ</td>
<td>0.6</td>
</tr>
</tbody>
</table>

### Table 5: ESIM without filter

<table>
<thead>
<tr>
<th>Comment</th>
<th>Submitted reply</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>楽しくって誰？知りませんが…</td>
<td>楽観は？</td>
<td>1.0</td>
</tr>
<tr>
<td>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。</td>
<td>ともだらな、いじられキャラなんでいう呪りがおしかしい。芸人があれば誰であれ他人が馬鹿にされているの見て面白いか？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</td>
<td>0.0</td>
</tr>
<tr>
<td>こういう記事に否定的な意見を言う人にはうんざり。是非とも頭張ってもらいたい。</td>
<td>否定的な意見がアホな人とは限らない！</td>
<td>1.0</td>
</tr>
<tr>
<td>好きなブランドの一つだけど、最近、気品が感じられなくなった。</td>
<td>気品はあったけど</td>
<td>1.0</td>
</tr>
<tr>
<td>本当にそう思うなら、もっと早め段階に、自分の言葉でSMAPに対する思いを語ってはよかった。それは、他のジャニーズメンバーに対しても同じ。</td>
<td>早い段階でSMAP出ない事は分かってたはずですがが…</td>
<td>0.8</td>
</tr>
<tr>
<td>ロマンがありますね今となっては感情ではないとわからないので残念ですあとこの方がずっともつべきだったのは手懐編ではなく、白球だったと思います</td>
<td>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いるの？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、子供みてると思いませんか。いい大人なのだだからもっと分別した行動をお勧めします。</td>
<td>0.0</td>
</tr>
<tr>
<td>見てみたいと分からないそれより立花冬季なんかやったら面白いか？島津義弘なんかもいいかも</td>
<td>立花いるしw</td>
<td>0.8</td>
</tr>
<tr>
<td>もうこの人はプロ意識に欠けている。</td>
<td>もうこの人の下手くそな文章に何も期待していません。</td>
<td>0.0</td>
</tr>
<tr>
<td>やっは山陽電鉄や！！</td>
<td>やっはクリボナベイルでしょう！！</td>
<td>0.0</td>
</tr>
<tr>
<td>この人の人間性やこれまでの悪徳ぶりを思うと素直に応援する気にはれない。</td>
<td>素直に応援したい。</td>
<td>1.0</td>
</tr>
<tr>
<td>Comment</td>
<td>Submitted reply</td>
<td>score</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------</td>
<td>-------</td>
</tr>
<tr>
<td>嵐崎って誰？知らないさんが…。</td>
<td>嵐崎の2点目とかクルトワなら止めると思うよ</td>
<td>0.8</td>
</tr>
<tr>
<td>失うものがあったとしても、自分の根っこ部分は曲げない。大事な事だと感じる。</td>
<td>舌を絡まなきゃ根は悪い人たちじゃないのか</td>
<td>0.6</td>
</tr>
<tr>
<td>ど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。</td>
<td>こういう意見今後どんどん増えていくと思うよ</td>
<td>1.0</td>
</tr>
<tr>
<td>好きなブランドの一つだけど、最近、気品が感じられない。</td>
<td>好きなんだしょうね</td>
<td>0.8</td>
</tr>
<tr>
<td>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はよかった。それは、他のジャニーズメンバーに対しても同じ。</td>
<td>もう解散したのに素晴らしい未来は無いでしょう？</td>
<td>0.8</td>
</tr>
</tbody>
</table>