Identifying Reply-to Relation in Textual Group Chat using Unlabeled Dialogue Scripts and Next Sentence Prediction

Junjie Shan, Yoko Nishihara, Yihong Han

Ritsumeikan Global Innovation Research Organization
Ritsumeikan University
Information Science and Engineering

2023/2/15

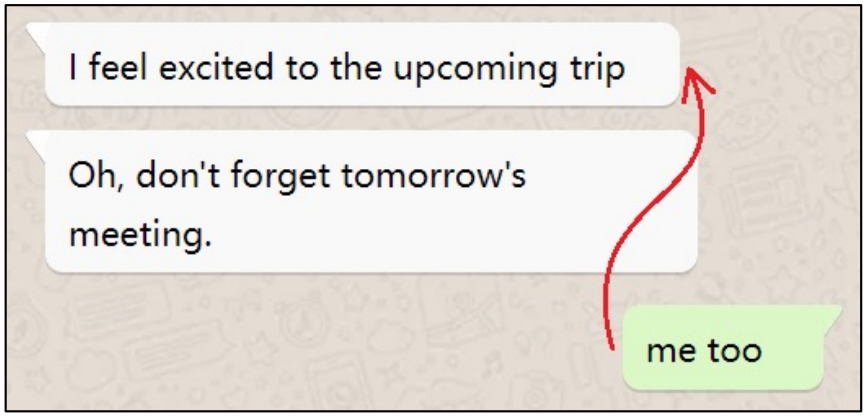
RITSUMEIKAN

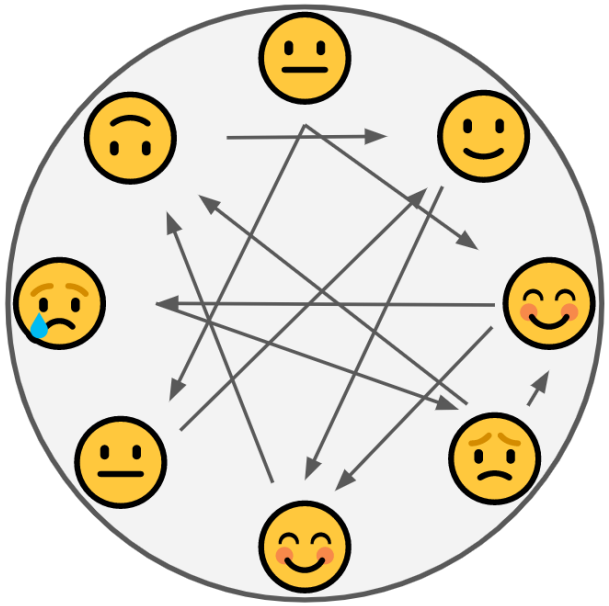
Contents

- Background
- Issues
- Contributions
- Proposed Method
  - "Reply-to" Messages Sampling from Dialogue Scripts
  - "Reply-to" relation's identification via NSP models
    - Orig-NSP
    - NSP-FA-IL
    - NSP-IL
- Evaluation
- Results & Discussions
- Conclustion & Future Works

Background

- Instant message (IM) tools have become a part of daily life
- Many researches aim at supporting communications on IMs
  - Relationship analyzation & sustainment
  - Topic recognition & provision
- Need to identify "Reply-to" relation first
  - Especially in the Group Chat





Issues

- Dataset for "Reply-to" relations' identification is hard to collect
  - Need to check "Reply-to" relations manually

- "Reply-to" relations in group chats are more complex
  - "Reply-to" past messages OR start of new topic?
  - Multiple "Reply-to" targets from one message

Message D replies to all of the previous messages A, B and C.

Contributions

1. Collect "Reply-to" messages dataset from dialogue scripts automatically
   Provided a method to sample messages with & without "Reply-to" relations from dialogue scripts

2. Identify "Reply-to" relations with multiple "Reply-to" targets
   Provided & evaluated the method of using Next Sentence Prediction to identify whether each message pair has a "Reply-to" relation.

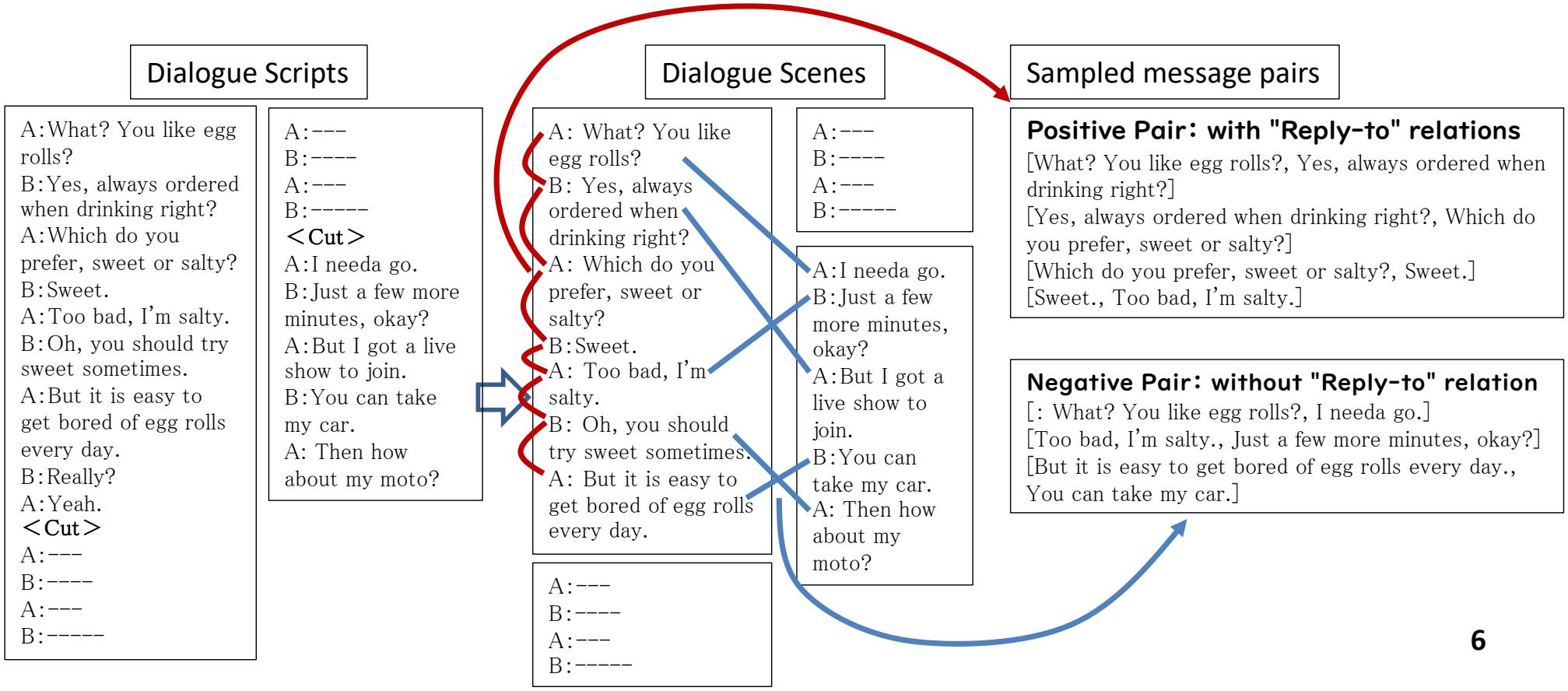
Proposed Method（1）

- Automatic sampling of "Reply-to" messages
  Sampling of messages' "Reply-to" relations from "dialogue scripts" with similar textual features to chat messages.
- Comparison of "dialogue scripts" and "chat messages":

	Chat Messages	**Dialogue Scripts**	Articles or News Reports
Short & Brief	○	○	×
Sequential	×	○	○
Multiple Sending	Many	Less	×
Topic in same time	Distributed	Concentrated	Concentrated
Speaker	Multiple	Multiple	Single
Reply-to relation	Complex	Simple	None

Proposed Method (2)

- Sampling of "Reply-to" relations from dialogue scripts
  - Adjacent two messages (dialogues) → Positive Pair
  - Two messages from different dialogue scenes (scripts) → Negative Pair



Proposed Method (3)

- Identify "Reply-to" relation through Next Sentence Prediction method
- Next Sentence Prediction (NSP) :
  - A supplement of the BERT pre-training process
  - learn to predict whether the 2nd sentence in the pair logically or meaningfully follows the 1st sentence

[I like afternoon tea.], [I usually take some pizza and milk at 4 p.m.]	→ 1
[I like afternoon tea.], [This castle was built 500 years ago.]	→ 0

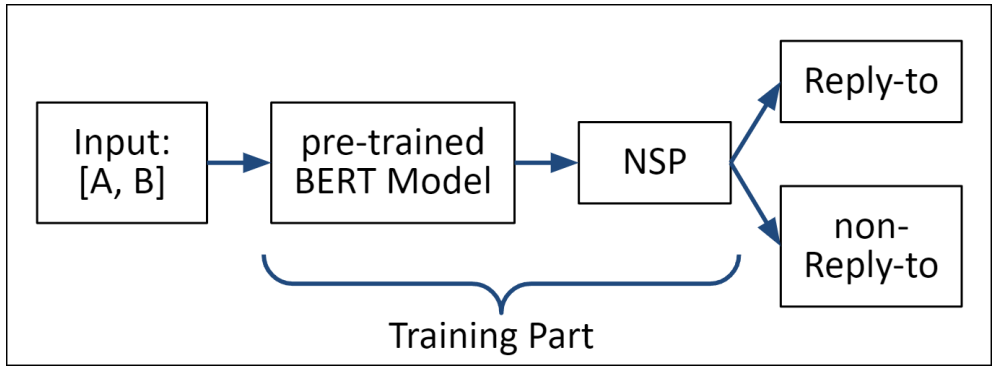
  - NSP task provides a mechanism that could receive sentence pairs directly on the pre-trained BERT model

Word ID : [CLS], [word list of 1st sentence], [SEP], [word list of 2nd sentence], [SEP]
Type ID : [0], [0], [0], [0], [0], [1], [1], [1], [1], [1],

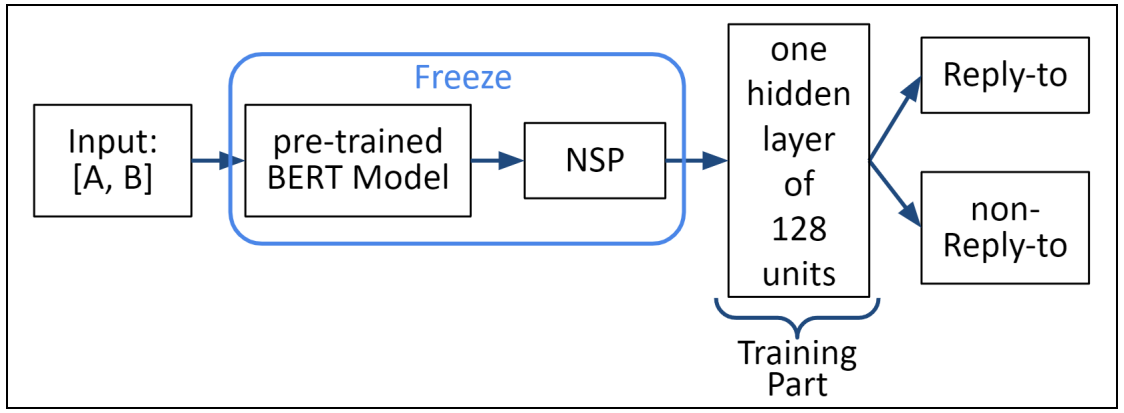
1st sentence　　2nd sentence

Proposed Method (4)

- Built and evaluated three structure settings of the NSP model to verify the effect of "Reply-to" relations' identification
  - 1, Orig-NSP: Original NSP model

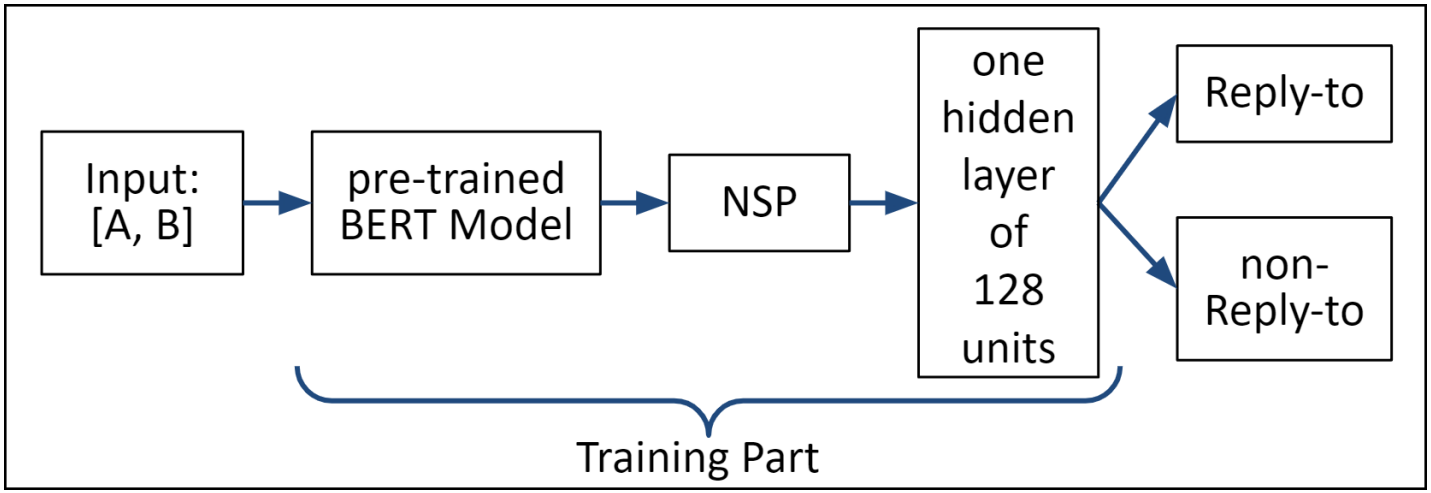


  - 2, NSP-FA-1L: NSP model for embedding sentence pair



提案手法（5）

– NSP-1L: NSP model + one hidden layer of 128 units



All three NSP models were initialized by the pre-trained Japanese BERT model published by Tohoku University:
https://github.com/cl-tohoku/bert-japanese/tree/v2.0

Training settings & results

- NSP model settings:
  - Training Data:
    Automatic sampled
    5094 message pairs
    from dialogue scripts
    - 1,698 Pos
      (with "Reply-to" relation)
    - 3,396 Neg
      (without "Reply-to" relation)

Items	Values
# of training data	5,094 message pairs
Token level	Character
Max length of input	128 (characters)
Batch Size	64
Epoch	10
Validation Rate	0.1
Learning Rate	5e-5 (0.00005)
Optimizer	Adam

- Training results

Model	Train Loss	Train Acc	Val Loss	Val Acc
Orig-NSP	0.0389	95.56%	0.6259	87.06%
NSP-FA-1L	0.4974	76.20%	0.4594	78.24%
NSP-1L	0.0807	97.25%	0.3888	**88.82%**

Evaluation

- Evaluation through actual Japanese group chat records
- Evaluation Procedures
  1. Collect actual messages record from chat group
  2. Manually check "Reply-to" relations between messages
  3. Predict "Reply-to" relations through trained NSP model
  4. Calculate Acc & F1 scores from prediction results and manually checked labels

B : How about the date ?
A : At those 4 places
B : It may take 2 nights to go around these spots.
C : My August is pretty full
A : so am I, maybe in late September I think
C : September is ok for me

B : There will be a mid-term report, late September.

1st Message (A)	2nd Message (B)	Reply-to	Predict
At those 4 places	There will be a mid-term report, Late September.	0	0
It may take 2 nights to go around these spots.	There will be a mid-term report, late September.	0	0
My August is pretty full	There will be a mid-term report, late September.	0	1
so am I, maybe in late September I think	There will be a mid-term report, late September.	1	1
September is ok for me	There will be a mid-term report, late September.	1	1

Evaluation Results

- Evaluation results of Accuracy & F1 score

Model	Accuracy	F1 Score
No-Training	49.8%	0.456
Orig-NSP	56.63%	0.519
NSP-FA-1L	**69.64%**	0.458
NSP-1L	62.77%	**0.558**

  - No-Training: original pre-trained BERT model without fine-tuning by sampled dialogue scripts' message pairs.
  - NSP-FA-1L obtained the highest accuracy of 69.64%
  - NSP-1L model got the highest F1 score of 0.558

Discussion

- Results Summary

Model	Val Acc	Test Acc	F1 Score
No-Training	-	49.8%	0.456
Orig-NSP	87.06%	56.63%	0.519
NSP-FA-1L	78.24%	**69.64%**	0.458
NSP-1L	**88.82%**	62.77%	**0.558**

  - All three trained models outperformed No-Training
    → Dialogue script data is effect for identifying "reply-to" relation
  - For all three trained models, Val Acc > Test Acc
    → Maybe caused by lack of training data (1698 Pos + 3396 Neg)
  - NSP-FA-1L obtained highest test Acc、but F1 score is almost the same as No-Training → Risk of over-fitting (only 1 layer trained)
  - NSP-1L outperformed Orig-NSP in both Acc & F1 score
    → Adding a smaller hidden layer is beneficial to support NSP fine-tuning to focus more on specific tasks

Discussion (2)

- Analysis of correct & incorrect identifications

	Aver. Length of First Msg. (A)	Aver. Length of Second Msg. (B)	Aver. Length of Input Pair
Correct	12.72	14.28	27.0
Incorrect	11.05	9.89	20.95

  - Aver length of 2nd Msg in incorrect < correct results
    → Identify "reply-to" relation between two independent msgs and ignore the contextual information
    →More incorrect identifications at short common response ("Yes", "Okay", "Sure", "Agree"…)

A : Oh, but the plane is in AA?
B : Or in BB?
A : I wonder if the plane is in CC too?
B : If so, I think AA might be a little confused.
B : But for the travel distance, the Zoo is difficult, isn't it?

C : Maybe

1st Msg (A)	2nd Msg (B)	Reply-to	Predict
Oh, but the plane is in AA?	Maybe	0	1
Or in BB?	Maybe	0	1
I wonder if the plane is in CC too?	Maybe	0	1
If so, I think AA might be a little confused.	Maybe	0	1
But for the travel distance, the Zoo is difficult, isn't it?	Maybe	1	1

Conclusion & Future Work

- A method for identifying multiple “Reply-to" targets of messages in group chat.
  - A method for automatically sampling "reply-to" relations between messages from "dialogue scripts" data, which textual features are similar to chat messages.
  - Identify "reply-to" relations from each two-message pair input through Next Sentence Prediction method.
- Built & trained three NSP models via collected 5094 message pairs from dialogue scripts, evaluated with actual group chat records.
  - Greater than 80% accuracy on validation set, higher than 60% accuracy on test set.
- Try to increase the training data, improve the proposed model and explore its effect more clearly.

Thank you very much!