Utterance Selection based on Sentence Similarities and Dialogue Breakdown Detection on NTCIR-12 STC Task

Hiroaki SUGIYAMA
NTT Communication Science Labs., NTT corp., Japan

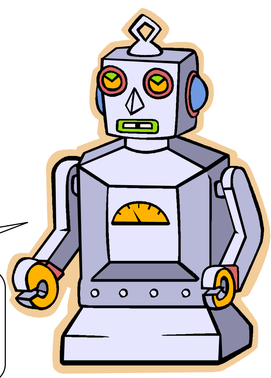
[STC 15][NTTCS]

Objective

Utterance selection for chat-oriented systems from a number of candidates

I like classical music, especially Beethoven.

I think Beethoven is thick sounds with presence.



Utterance candidates

Hello! How are you t
I usually eat ethnic food, but I haven't cooked it yet.
I think Beethoven is thick sounds with presence.
The coffee is so bad.
I go to arcades because there is almost n
I like sweet cakes with many strawberries!

Proposed approach

User utterance
I like classical music, especially Beethoven.

Microblog Corpus

Sentence similarity based retrieval

- Retrieve sentences that resemble the user utterance using word2vec.

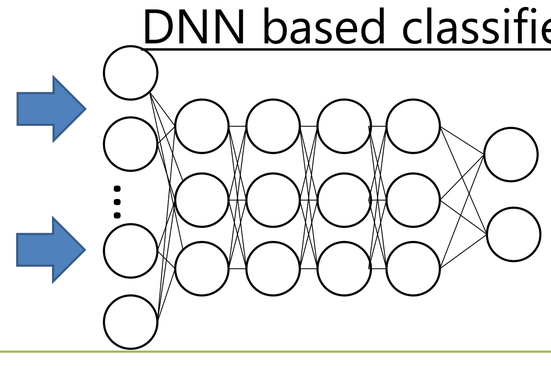
$$Sim(s_1, s_2) = v_{s_1} \cdot v_{s_2} \qquad v_s = \frac{\sum_{w \in W_s} v_w}{|W_s|}$$

v_w: word2vec based vector

Dialogue breakdown detection based reranking

- Detects inappropriate utterances that cause dialogue breakdown.
- We leverage the estimated appropriateness to reranking of the candidates.

User utterance
I like classical music, especially Beethoven.
Response utterance
This coffee is so bad.



Appropriateness
0.163
(not appropriate)

Reranking
1. *I think Beethoven is thick sounds with presence.*
2. *Hello! How are you today?*
3. *The coffee is so bad.*

System utterance
I think Beethoven is thick sounds with presence.

Experiments

Settings

- Microblog corpus: 100M tweet pairs
- Test data: 202 tweets are used for user utterances
  - Systems output ten tweets for each input tweet.
  - Humans label 0, 1, 2 to the system outputs.

Approach	Correct=2 1-best	Correct=2 5-best	Correct=1&2 1-best	Correct=1&2 5-best
W/o DBD	**0.0921**	**0.0698**	0.2639	0.2318
W/ DBD	0.0876	0.0677	**0.2946**	**0.2333**

Implements

- **Word2vec**
  - Develop with 150M tweets (2013)
- **Dialogue breakdown detection (DBD)**
  - Six layer perceptron [Sugiyama 2015]
  - Trained with DBD corpus [Higashinaka, 2015]

Results and Discussion

☺ DBD reranking is effective to filter inappropriate utterances.
☹ DBD does not have enough sensitivity to distinguish labels 1 and 2, because the DBD was trained with dialogue system's utterances that are less appropriate than the tweets.

Future work

- Examine fine-tuned DBD system with the tweets.