**UT Dialogue System at NTCIR-12 STC**

Shoetsu Sato¹, Shonosuke Ishiwatari¹, Naoki Yoshinaga², Masashi Toyoda², and Masaru Kitsuregawa²,³

¹ The University of Tokyo, ² IIS, the University of Tokyo, ³ NII, Japan

**Background**

In data-driven approaches for chat-dialogue modeling, the diversity of domains (topics, speaking styles, emotions..) makes it difficult to learn.

- **U:** Utterance
- **R:** Response

**Related work**

Classify training data by several emotion types each response elicits and train multiple models [Hasegawa+, '13]

\[
\text{But it is impossible to enumerate all domains in human dialogues by hand}
\]

**Our approach**

Cluster conversation data to automatically capture the difference of domains and train specific models

- Domain-consistent responses
- Smaller size of the training data per a model

**Proposed Method**

1. **Apply k-means clustering** to the utterance vectors and regard clusters as subsets of the training data.

2. Narrow the number of the candidates to reduce computation by the pre-trained classifier [yoshinaga+, '10]

3. Train multiple LSTM-based dialogue models by each domain-specific training data subset.

4. Select the model to respond from distance between cluster's centroids and the utterance vector and response from candidates.

**Experiments**

**Effectiveness of clustering**

**Data**

100K (tweet-reply) pairs for train, 1K for test

**Evaluation method**

- **Utterance**
  - 発表つらいんだけど
  - わかる
  - **Correct Response**
  - 自分の研究を知ってもらう良い機会だよ
  - 今完全に鬱だよ
  - その店美味しいよね
  - くわせでる

- **Ranked 20 Candidates**
  - 1. わかる
  - 2. **自分の研究を知ってもらう良い機会だよ**
  - 3. 今完全に鬱だよ
  - 4. その店美味しいよね
  - 5. くわせでる

**Results**

- **Best Result (35.4%)**
- **Baseline (30.8%)**
- **Random Baseline (15.0%)**

**Difference between baseline and proposed method**

**Utterance**

- **Baseline**
  - あ、見るのは忘れてた。おめでとう！
- **Proposed**
  - ありがたいです。

**Our method less frequently select typical responses by extracting them as other domains**

**NTCIR-12 STC formal-run**

**Evaluation**

1. Evaluate filters trained on different size of training data, by recall whether top-N filtered candidates including the correct response.

2. Selected responses are assigned score of 0 (inappropriate), 1 (appropriate in some context), and 2 (appropriate) by human, and evaluated the proportion of 1 and 2, or only 2 for the top-1 or top-5 selected responses.

**R1 : Responses selected by our system from filtered candidates**

**R2 : Responses only pre-filtered**

**Filtering performance**

**Accuracy on the NTCIR-12 STC task**