UT Dialogue System at NTCIR-12 STC

Shoetsu Sato[1], Shonosuke Ishiwatari[1], Naoki Yoshinaga[2],
Masashi Toyoda[2], and Masaru Kitsuregawa[2,3]
{shoetsu, ishiwatari, ynaga, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

[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: **フォロー**し**ました！**
R: ありがとう**ございます！**

U : Utterance
R : Response

U: また**残業**か・・・
R: 生き残ろう**な**・・・

U: **ラーメン食べ**たい気分
R: 今日の夜行こう**ぜ**

Related work

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

But it is impossible to enumerate all domains in human dialogues by hand

Our approach

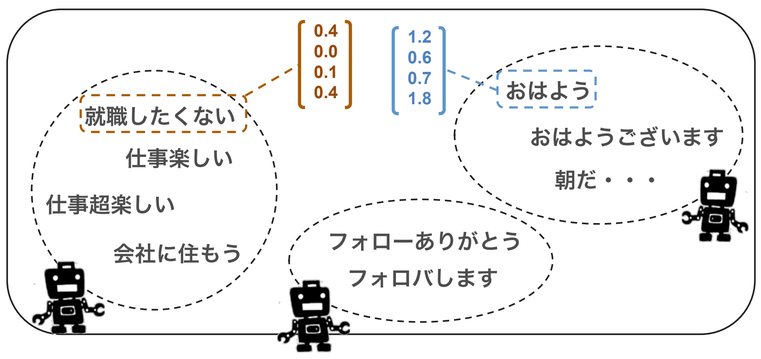
Cluster conversation data to **automatically** capture the difference of domains and train specific models

☺ Domain-consistent responses

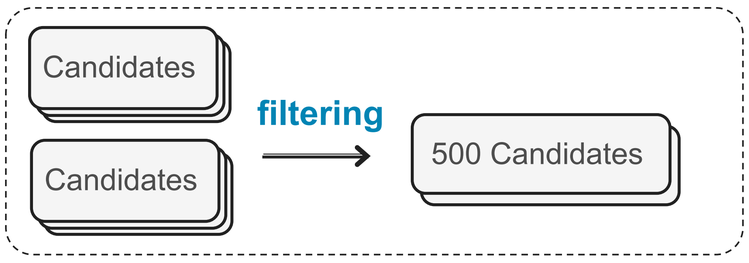
☹ Smaller size of the training data per a model

Proposed Method

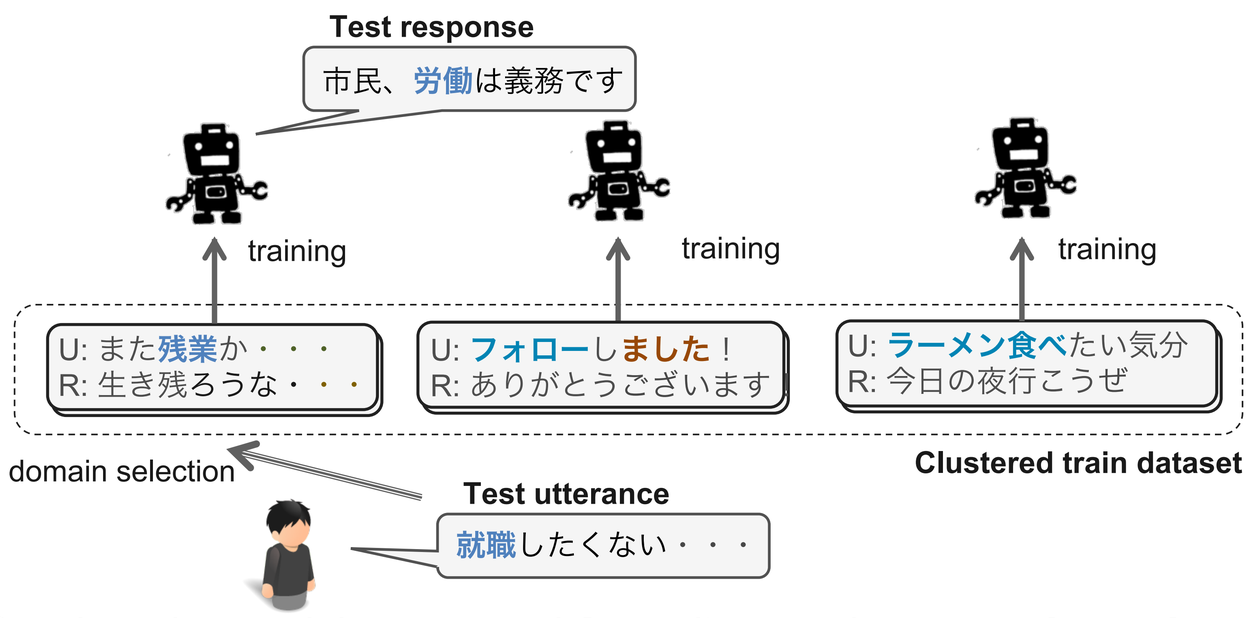
① Apply **k-means clustering** to the utterance vectors and **regard clusters as subsets** of the training data



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

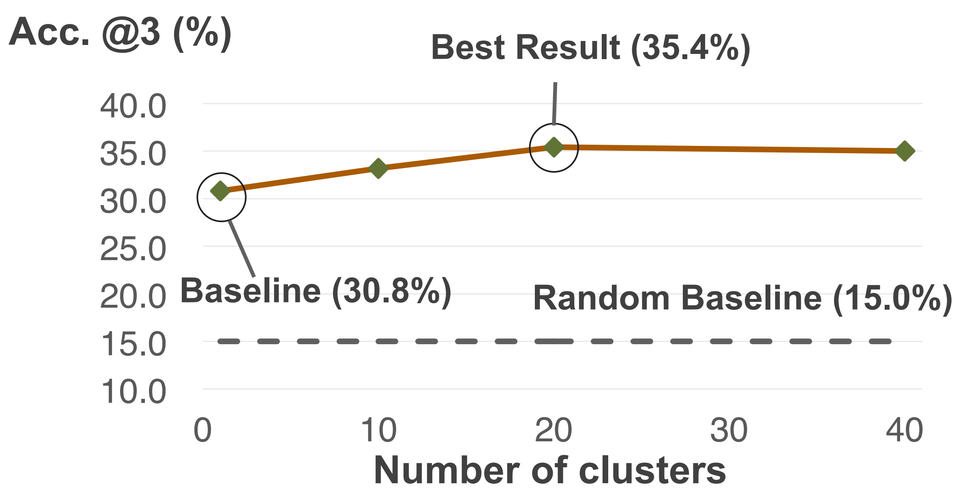
Ranked 20 Candidates

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 その店美味いよね
⋮
⑳ くぁwせｄｒｆｔｇ

Correct Response ②

We defined it as success if the **top-3 responses** include the correct response

Results



Difference between baseline and proposed method

Utterance あ、見るの忘れてた。おめでとう！
Baseline 今年は1年ありがとうございました
Proposed ありー！見なおしてくれてありがとう！

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

●NTCIR-12 STC formal-run

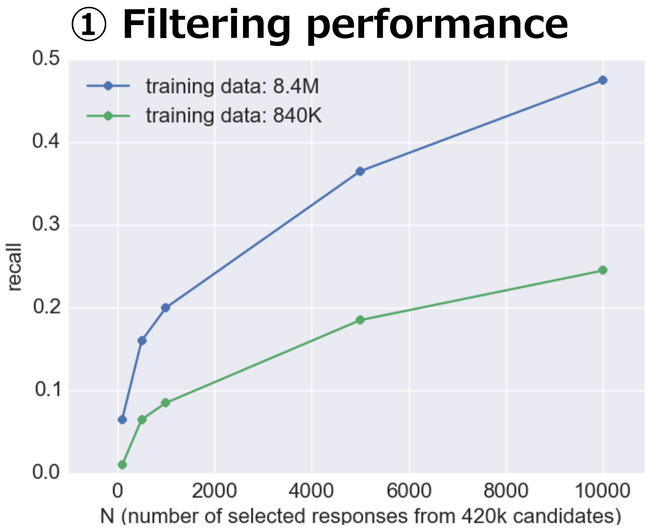
Evaluation

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

① Filtering performance



② Accuracy on the NTCIR-12 STC task