UT Dialogue System
at NTCIR-12 STC

Shoetsu Sato\textsuperscript{1}, Shonosuke Ishiwatari\textsuperscript{1}, Naoki Yoshinaga\textsuperscript{2}, Masashi Toyoda\textsuperscript{2}, Masaru Kitsuregawa\textsuperscript{2,3}

The University of Tokyo, \textsuperscript{2}IIS, the University of Tokyo, \textsuperscript{3}NII, Japan
A lot of dialogue systems that can chat have appeared

Siri (Apple)

Cortana (Microsoft)

しゃべってコンシェル (NTT Docomo)

http://www.idownloadblog.com/
http://techcrunch.com/2015/01/05/facebook-wit-ai/
http://ameblo.jp/cos-120/entry-11748747974.html
Recent approaches for chatting dialogue systems

- Data-driven approaches using dialogue data from social media are promising [Ritter+,'10]

U: Utterance
R: Response

U: また残業か・・・
R: 生き残ろうな・・・

U: あの人がどう思う？
R: ああいう人間ほんと嫌い

U: 魚介嫌いでした？
R: そんなこと無いですよ。
Challenge we have tackled in STC task

- The diversity of domains (topics, speaking styles, etc...) makes it difficult to learn.

U: Utterance
R: Response

U: あの人どう思う？
R: ああいう人間ほんと嫌い

U: また残業か・・・
R: 生き残ろうな・・・

U: 魚介嫌いでした？
R: そんなこと無いですよ。
Goal: building a domain-aware dialogue model

Idea: Divide conversation data into domain-consistent subsets to train multiple specific LSTM-based dialogue models

Evaluation: response selection from candidates

Does domain consistence compensate for reduction of training data per a model?
RELATED WORK
Recent promising approach to generate responses

- We employed recent promising Long-Short Term Memory based Recurrent Neural Network (LSTM-RNN) dialogue model [Vinyals+, ‘14]
Generate a response that elicits a specific emotion in the addressee’s mind.

- Joy: 早く良くなるといいですね
- Sadness: あまり近寄らないで下さい
Overview of the related work and target of our method

Point

- General LSTM based methods employed a single model trained from all data
- It is impossible to enumerate all domains in human dialogues

Purpose

- Capture the difference of domains automatically as clusters and train multiple models
Our approach: K-cluster model (1/2)

- **Cluster** the dialogues for each of the *unlabeled domain*, and train multiple models

U: また残業か・・・
R: 生き残ろうな・・・

U: あの人がどう思う？
R: ああいう人間ほんと嫌い

U: 魚介嫌いでした？
R: そんなこと無いですよ。
Our approach: K-cluster model (2/2)

Utterance

就職したくない・・・

Find the nearest domain by human utterance and utterances in training subsets

U: また残業か・・・

U: あの人がどう思う？

U: 魚介嫌いでした？

Response

市民、労働は義務です。
How to automatically handle the domains in each utterance (1/2)

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

---

就職したくない
仕事楽しい
仕事超楽しい
会社に住もう

おはよう
おはようございます
朝だ・・・

フォローありがとう
フォロバします
How to automatically handle the domains in each utterance (2/2)

- Represent each utterance as a vector built from word embeddings [Mikolov+, ‘13]
- The density of word embeddings would solve sparseness problems in short texts compared with Bag-of-Words
In response selection task from many candidates, our model’s **high computational cost** causes a problem.

To reduce the number of candidates into **500**, we employed a fast SVM classifier [Yoshinaga+, ‘10].
3 experiments we did

- **Experiment 1: Small response selection task**
  - Evaluate how our method effects in response selection
  - Select response from 20 candidates *without filtering*

- **Experiment 2: Filtering performance**
  - Evaluate to what extent our filter can select proper candidates

- **Experiment 3: NTCIR-12 formal run**
  - Evaluate whole performance of our system (clustered-LSTM, and filter)
Experiment 1: small response selection task

- **Dataset: Twitter**
  Utterance-response (tweet-reply) pairs crawled from Twitter: **100K** for training, **1K** for test
  - Provided for *NTCIR-12 Short Text Conversation Japanese Task* [Shang+, ‘16]

- **Evaluation: Response selection**
  The proportion of test tweets where we succeeded to select the **correct (actually replied) response** from randomly chosen **20 candidates**
We defined it as success if the top-3 responses include the correct response.
Results of K-cluster model

- We compared 1, 10, 20, and 40 cluster models increasing number of clusters until the accuracy was saturated

![Graph showing accuracy vs number of clusters]

Accuracies:
- K=1: Baseline (30.8%)
- K=20: Best Result (35.4%)
- Random Baseline (15.0%)

Our best 20-cluster model outperformed baseline at 4.6%
Examples of clusters we labeled by hand from their frequent utterances

- 13 models out of 20 cluster got improved

Improvement rate vs frequency
### Examples of the selected responses

<table>
<thead>
<tr>
<th></th>
<th>Utterance</th>
<th>Proposed</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>あ、見の忘れてた。おめでとう!</td>
<td>ありー！見直してくれてありがとう!</td>
<td>今年は1年ありがとうございました</td>
</tr>
<tr>
<td>2</td>
<td>カントリーマァムのドリンクのやつが見つかりません。</td>
<td>ローソン限定じゃなかったっけ？</td>
<td>先輩、おはよーございます♪</td>
</tr>
</tbody>
</table>

Our proposed model tends to stop selecting typical responses
Experiment 2: Filtering performance

Evaluate the filter by recall, whether top-N filtered candidates include the **correct response**

Filtering effectively reduced the number of candidates
Experiment 3 : NTCIR-12 STC Japanese Task

- **Model:** 20 cluster model
  The best one evaluated at experiment 1, 20 cluster model trained from 100k utterance-response pairs

- **Evaluation:**
  - For the 204 provided test utterances, select responses from 500k candidates
  - responses are assigned scores of 0 (inappropriate), 1 (appropriate in some context), 2 (appropriate) by human annotators
Accuracies of selected top-1 responses at NTCIR-12 STC Japanese Task

Ours-R1: Filtering + 20-cluster LSTM model
Ours-R2: Filtering only

Our system selected better responses from filtered candidates
By response selection test we confirmed the effect of cluster-based domain-aware dialogue model

- Domain-consistent training subsets made better results in spite of reduction of training data
- By filtering candidates, our system could effectively select responses
RESULTS FOR EACH CLUSTER
## Results in each cluster (20-cluster model)

<table>
<thead>
<tr>
<th>domain (topics, wording, writing style)</th>
<th>#elems</th>
<th>#corr</th>
<th>improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>train</td>
<td>test</td>
<td>ours</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11801</td>
<td>108</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>11524</td>
<td>124</td>
<td>37</td>
</tr>
<tr>
<td>politics, economics, social matters</td>
<td>10294</td>
<td>130</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>9743</td>
<td>94</td>
<td>32</td>
</tr>
<tr>
<td>animation, comics</td>
<td>6747</td>
<td>56</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>6552</td>
<td>66</td>
<td>24</td>
</tr>
<tr>
<td>game</td>
<td>5677</td>
<td>50</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>5627</td>
<td>45</td>
<td>14</td>
</tr>
<tr>
<td>end with ‘?’ r ‘!’</td>
<td>5190</td>
<td>63</td>
<td>17</td>
</tr>
<tr>
<td>moaning (esp., sleepy, weary)</td>
<td>5064</td>
<td>52</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>4908</td>
<td>50</td>
<td>22</td>
</tr>
<tr>
<td>numbers</td>
<td>3803</td>
<td>31</td>
<td>5</td>
</tr>
<tr>
<td>eating</td>
<td>2630</td>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>frank acknowledgment (follow, RT)</td>
<td>2252</td>
<td>33</td>
<td>29</td>
</tr>
<tr>
<td>end with ’!!!’</td>
<td>1869</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>polite acknowledgement (follow, RT)</td>
<td>1553</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>greetings</td>
<td>1537</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>end with ‘...’</td>
<td>1326</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>polite morning greetings</td>
<td>1174</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>shouting with word lengthing or repetition</td>
<td>729</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>100000</td>
<td>1000</td>
<td>354</td>
</tr>
</tbody>
</table>