The best of both worlds: Combining machine and human intelligence to crowdsource dialog data

Motivation
Current dialog systems often require large-scale domain-specific corpora as training inputs, yet it is difficult to collect domain-specific data to bootstrap and prototype conversational agents. To facilitate this, a systematic process for collecting both user and agent utterances is necessary.

Research Objective
We present a workflow to assist the process of collecting data, with these objectives in mind:

• How do we measure the quality of training data during the data collection phase?
  - Our framework alternates between humans-in-the-loop annotation and machine learning to identify when sufficient data have been collected.

• How do we efficiently collect data while maintaining data quality?
  - Our framework combines both crowdsourced ratings and machine-learning techniques to remove noisy data.

• How can we use crowdworkers to generate different interaction styles for the agent?
  - Our framework allows crowdworkers to generate and rate agent utterances for the purpose of training a dialog agent to interact using different interaction styles.

Dataset
- **Scenario:** Customer support (Q&A) for a real estate agent.
- **Intents:**
  - 29 intents
  - Examples: Which neighborhood the agent covers, neighborhood safety, services the agent provides.
  - Highly noisy customer and agent utterances (e.g., out-of-scope, not English, nonsensical words).

Training data quality metrics

**Stopping Criterion 1 (Pairwise semantic similarity):** Measure the proportion of utterance pairs within an intent which surpasses a threshold level of pairwise similarity, and use this as a stopping criterion.

**Stopping Criterion 2 (Connected component clustering):** Connected components build the paths between any existing subgraphs and a single vertex to eventually reach a final stable graph in which any vertex belongs to one of many components. Stop when the ratio of utterances in a main cluster has surpassed a threshold.

**Stopping Criterion 3 (Intent classification):** Train an intent classifier and evaluate the accuracy of each intent. We stop collecting example utterances when the accuracy of an intent is above a threshold.

**Result**
**Offline evaluation:** Even with the same number of customer training examples, an intent classifier achieved better performance from training data collected from the quality control phase as compared to data collection phase.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Customer Utterances</th>
<th>Agent Utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Collection Phase</td>
<td>21692</td>
<td>7480</td>
</tr>
<tr>
<td>Quality Control Phase</td>
<td>16410</td>
<td>2943</td>
</tr>
<tr>
<td>Fine Tuning Phase</td>
<td>1667</td>
<td></td>
</tr>
</tbody>
</table>

Agent data evaluation (Ongoing):
- Developed 3 different versions of the agent trained with data from the data collection phase, quality control phase, and fine tuning phase.
- We hypothesize the agent trained with the data from the fine tuning phase will achieve the best agent behavior overall.

**References**