Submodularity-Inspired Data Selection for Goal-Oriented Chatbot Training Based on Sentence Embeddings

Mladen Dimovski¹, Claudiu Musat², Vladimir Ilievski¹, Andreea Hossmann², Michael Baeriswyl²
¹IC School, EPFL, Switzerland ²Artificial Intelligence Group, Swisscom

Slot Filling

Are there any French restaurants in downtown Toronto?

**BILSTM model**

- Task: tag the relevant keywords in the user query (sequence-tagging problem)

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>0</th>
<th>B-Cuisine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are</td>
<td>there</td>
<td>any</td>
<td>French</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>B-Location</td>
<td>I-Location</td>
</tr>
<tr>
<td>restaurants</td>
<td>in</td>
<td>downtown</td>
<td>Toronto</td>
</tr>
</tbody>
</table>

Data Availability

- Existing BiLSTM models perform reasonably well if given enough training data.
- What if we can afford to label only small amount of data?

**MAIN IDEA**

Rank the raw unlabeled sentences according to their usefulness and select only the best ones for labeling.

<table>
<thead>
<tr>
<th>BAD CHOICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I want to eat salmon!</td>
</tr>
<tr>
<td>I want to eat pizza!</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GOOD CHOICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can you propose a good restaurant serving beef in the city center?</td>
</tr>
<tr>
<td>What is the best rated Chinese restaurant in Lausanne?</td>
</tr>
</tbody>
</table>

Sentence embeddings and sentence similarity

- Use a recently developed technique, sent2vec, that produces continuous vector representations of sentences.
- Define the similarity between two sentences $\mathbf{x}$ and $\mathbf{y}$ as:
  $$\text{sim}(\mathbf{x}, \mathbf{y}) = \exp(-\frac{1}{2}||\mathbf{e}(\mathbf{x}) - \mathbf{e}(\mathbf{y})||^2)$$
  where $\mathbf{e}(\mathbf{x})$ is the embedding of the sentence $\mathbf{x}$ and $\mathbf{z}$ is the inverse of the average distance between all pairs of embeddings.
- Hypothesis: closeness in the embedding space is in line with the human perception of sentence similarity.

**What was the movie that featured Over the Rainbow?**

- Find me the movie with the song Over the Rainbow.
- What movie was the song Somewhere Out There featured in?
- What movie features the song Hakuna Matata?

Data selection methods

- Use a well-chosen submodular function to evaluate the usefulness of each subset of sentences $\mathbf{X}$.

$$A function \; F : 2^V \to \mathbb{R} \; is \; called \; submodular \; if \; the \; value \; of \; an \; element \; diminishes \; as \; the \; context \; in \; which \; it \; is \; considered \; grows. \; Formally, \; F \; is \; submodular \; if \; F(\mathbf{y}) \leq F(\mathbf{x} \cup \mathbf{y}) \; for \; every \; \mathbf{x} \subseteq \mathbf{Y} \; and \; every \; s \in \mathbf{V} \setminus \mathbf{Y}.$$ 

**Baselines**

- Random Data Selection
- Classic Active Learning
- Randomized Active Learning

Uncertainty of the model for a sentence $\mathbf{x}$ of $k$ words:

$$\psi(x) = \frac{1}{k} \sum_{i=1}^{k} (1 - \text{pm}(s_i | x))$$

**Coverage Score**

$$C(\mathbf{X}) = \sum_{\mathbf{x}, \mathbf{y} \subseteq \mathbf{V}} \text{sim}(\mathbf{x}, \mathbf{y})$$

**Linear-Penalty Marginal Gain**

$$D(\mathbf{s}(\mathbf{x})) = \sum_{\mathbf{y} \subseteq \mathbf{V}} \text{sim}(\mathbf{s}(\mathbf{y}), \mathbf{s}(\mathbf{x})) - \alpha \sum_{\mathbf{x} \subseteq \mathbf{X}} \text{sim}(\mathbf{s}(\mathbf{x}), \mathbf{s}(\mathbf{x}))$$

**Ratio-Penalty Marginal Gain**

$$F(\mathbf{s}(\mathbf{x})) = \sum_{\mathbf{y} \subseteq \mathbf{V}} \frac{\text{sim}(\mathbf{s}(\mathbf{y}), \mathbf{s}(\mathbf{x}))}{\sum_{\mathbf{x} \subseteq \mathbf{X}} \text{sim}(\mathbf{s}(\mathbf{x}), \mathbf{s}(\mathbf{x}))}$$

**Results**

- We show that the space of raw, unlabeled sentences contains information that we can use to choose the sentences to label.
- We create a submodularity-inspired ranking function, the ratio-penalty marginal gain, for selecting the most useful sentences to label.
- We apply this data selection method to the problem of slot filling and prove that the model’s performance can be considerably better with training samples chosen in an intelligent way.

For more information, see: M. Dimovski et al., Submodularity-Inspired Data Selection for Goal-Oriented Chatbot Training Based on Sentence Embeddings, IJCAI 2018