Motivation: Inflexible Intent Schema
- Intents are usually predefined and inflexible to expand and transfer across domains and genres, where re-designing a semantic schema with intents for different domains or genres requires human effort for annotation and model re-training.

Approach: Learning Intent Representation
- Applying CDSSM to learn high-level semantic representations to bridge the semantic relation across domains and across genres for intent expansion and actionable item detection tasks respectively, (e.g. “find movie” and “find weather” belong to different domains, but they share the semantics about “find”.)

Result
- CDSSM is capable of generating more flexible intent embeddings to remove the domain constraint in dialogue systems for intent expansion. The intent embeddings can also be transferred to different genres, showing the robustness to genre mismatch.

Convolutional Deep Structured Semantic Models (CDSSM)

- **Model Architecture**
  - **Semantic Layer:** \( y \)
    - feed-forward neural network layers outputs the final non-linear semantic features
  - **Max Pooling Layer:** \( l_n \)
    - only retain the most prominent local features by applying the max operation over each dimension of \( i \) to keep the max activation of hidden topics across the whole word sequence
  - **Convolutional Layer:** \( l_c \)
    - contextual features \( c \) for each target word \( i \)
    - \( i = \text{tanh}(W^c_c) \)
  - **Word Hashing Layer:** \( l_h \)
    - one-hot word vector \( \rightarrow \) tri-letter vector (e.g. "email" \( \rightarrow \) "email", "ema", "mai", "ail", "ill")
  - **Word Sequence:** \( x \)
    - user utterance / intent

- **Training Procedure**
  - The objective maximizes the likelihood of associated intents given utterances using mini-batch SGD updates:
    - **Predictive Model**
      \[ \Lambda(\theta_1) = \log \prod_{(U, I)} P(I^+ | U) \]
    - **Generative Model**
      \[ \Lambda(\theta_2) = \log \prod_{U^+} P(U^+ | I) \]

- **Intent Expansion**
  - For each utterance vector \( y_i \), the semantic similarity can be estimated using vectors for both seen and unseen intents.
  - The unseen intent vectors can be generated from CDSSM by feeding the tri-letter vectors of the new intent as input without model re-training.

- **Intention Classification Performance**

<table>
<thead>
<tr>
<th>Intent</th>
<th>MAP (%)</th>
<th>Orig.</th>
<th>Exp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seen</td>
<td>58.3</td>
<td>58.9</td>
<td>61.6</td>
</tr>
<tr>
<td>Unseen</td>
<td>66.8</td>
<td>66.6</td>
<td>66.0</td>
</tr>
<tr>
<td>K=1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>K=3</td>
<td>9.1</td>
<td>9.1</td>
<td>9.1</td>
</tr>
<tr>
<td>K=5</td>
<td>31.0</td>
<td>33.5</td>
<td>34.5</td>
</tr>
<tr>
<td>K=10</td>
<td>66.7</td>
<td>66.7</td>
<td>67.7</td>
</tr>
<tr>
<td>K=20</td>
<td>36.0</td>
<td>36.6</td>
<td>36.6</td>
</tr>
</tbody>
</table>

- The expanded models consider new intents without training samples, and produces similar but slightly worse than original models for seen intents due to higher uncertainty from more intent candidates.
- For unseen intents, expanded models are able to capture the correct intents and achieve higher than 30% of MAP when K \( \geq 3 \), which indicates the encouraging performance when considering more than 100 intents.

- **Actionable Item Detection**
  - This task investigates actionable item detection in meetings (human-human genre), where the intelligent assistant dynamically provides the participants access to information (e.g. scheduling a meeting, taking notes) without interrupting the meetings.
  - A CDSSM is applied to learn the latent semantics for human actions and utterances from human-machine and human-human interactions.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Mismatch</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive</td>
<td>52.8</td>
<td>64.5</td>
</tr>
<tr>
<td>Generative</td>
<td>55.0</td>
<td>64.8</td>
</tr>
<tr>
<td>Bidirectional</td>
<td>59.1</td>
<td>68.9</td>
</tr>
</tbody>
</table>

- A CDSSM is applied to learn the latent semantics for human actions and utterances from human-machine and human-human interactions.
- The improvement of bidirectional estimation suggests that the predictive and generative model can compensate each other, and then provide more robust estimated scores for the goal of actionable item detection.

**Conclusion**
- The experiments show that the learned embeddings
  - capture the semantics borrowed from other domains and can be used to flexibly expand the intents through high-level representations.
  - carry the crucial high-level semantics and can be applied to different genres for easy adaptation and extension.