Motivation
- Rich punctuations play important roles in many NLP tasks.
- ASR systems provide plain word streams.
- Small-scale models are hard to guarantee performance and generalization ability.
- We need data and model with larger scale to adapt to various genres of text.

Data Comparison
- Data used in previous research
  - PTB, CTB
  - Manual speech transcripts
- Features used in previous research
  - Syntax information
  - Prosodic cues
- What we used
  - abundant diversity
  - large scale
  - formal/informal

Task Formulation
- Punctuation Symbols ➔ Labels

<table>
<thead>
<tr>
<th>Original Symbols</th>
<th>Unified</th>
<th>Punctuation</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>. . . . . :</td>
<td>,</td>
<td>Comma</td>
<td>CO</td>
</tr>
<tr>
<td>. . . . . -</td>
<td>-</td>
<td>Period</td>
<td>PE</td>
</tr>
<tr>
<td>! ! ! ! !</td>
<td>!</td>
<td>Exclamation</td>
<td>EX</td>
</tr>
<tr>
<td>Space, None</td>
<td>None</td>
<td>None</td>
<td>EM</td>
</tr>
</tbody>
</table>

- Sequence Labeling
  - You know what? 水稻、小麦、玉米，都是中国的主要农作物。
  - 预处理
    - 你 知 道 吗 水稻 小麦 玉米 都 是 中国 的 主 要 农作物

Evaluation Metrics
- Precision (denoted by P), Recall (denoted by R) and F1-score (the harmonic mean of precision and recall, denoted by F)
- Only evaluate for CO, PE, EX, QU, and SE

Methods
- CRF-based Model

<table>
<thead>
<tr>
<th>CRF Feature-Templates</th>
</tr>
</thead>
</table>

| CRF Feature-Templates |

- Rule Settings
  - Conjunction: insert comma before a conjunction.
  - Parenthesis: insert comma or period to the front and rear of parenthesis.
  - Interrogative sentence: insert question mark after the tone word if interrogative indicator is detected.
  - Exclamatory sentence: insert exclamation mark after the tone word if exclamatory indicator is detected.

- LSTM Model and Multiview-LSTM Model
  - Words: \( x_t = \{w_0, w_1, \ldots, w_{t-1}, w_{t+1}, \ldots, w_T\} \)
  - POS-tag / Chunking-tag:
    - \( \nu_t = \{h_0, h_1, \ldots, h_{t-1}, h_{t+1}, h_T\} \)
    - \( \nu_t = \{p_0, p_1, p_{t+1}, p_{t+2}, p_{t+3}, p_T\} \)
  - Multiview: concatenate \( x_t \) and \( \nu_t / \nu_T \)

Experiments
- CRF-based Model vs. LSTM Model (left, F1-score of each label)
  - LSTM significantly outperforms CRF-based model.
  - LSTM model knows better about when to stop.
  - In Chinese, the selection of punctuations relies more on long-term context.
    - It’s hard to distinguish comma and period.
  - Scale vs. Performance (right, F1-score of sentence boundary)
    - Increasing data scale helps improve model performance
    - Low model complexity may limit the potential of large data

Conclusion
- Long-term dependency matters!
- The more data, the higher performance! (even though there are some noise)
- Syntax information can boost the model when using small corpora.
- Need more design of both architecture and algorithm.