The CRECIL Corpus: a New Data Set for Extraction of Relations between Characters in Chinese Multi-party Dialogues

Yuru Jiang*, Yang Xu*, Yuhang Zhan*, Weikai He*, Yilin Wang*, Zixuan Xi*, Meiyun Wang*, Xinyu Li*, Yu Li*, and Yanchao Yu†
* jiangyuru@bistu.edu.cn † y.yu@napier.ac.uk

Introduction
We describe a new freely available Chinese multi-party dialogue corpus for automatic extraction of dialogue-based character relationships:
- extracted from the original TV scripts of a Chinese sitcom called “I Love My Family”.
- contains complex family-based human daily spoken conversations in Chinese
- introduced human annotation scheme for both global Character relationship map and character reference relationship – between 140 entities.

We also carried out a data exploration experiment by deploying a BERT-based model to extract character relationships on the CRECIL corpus and another existing relation extraction corpus (DialogRE[1]).
- extracting character relationships is more challenging in CRECIL than in DialogRE.

Referential Relationship

Dialogue-based Character Relationship Triples Generator

Comparison with DialogRE

Table 1: Comparison between different categories (%)

<table>
<thead>
<tr>
<th>Category</th>
<th>EN-DialogRE</th>
<th>CN-DialogRE</th>
<th>CRECIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>59.4</td>
<td>63.7</td>
<td>56.8</td>
</tr>
<tr>
<td>Test</td>
<td>57.9</td>
<td>63.2</td>
<td>54.4</td>
</tr>
</tbody>
</table>

Table 2: Comparison between the CRECIL corpus and the DialogRE corpus (%)

<table>
<thead>
<tr>
<th>Category</th>
<th>CRECIL</th>
<th>DialogRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average length</td>
<td>23.8</td>
<td>225.8</td>
</tr>
<tr>
<td>Average # turns</td>
<td>12.9</td>
<td>3.3</td>
</tr>
<tr>
<td>Average # of speakers</td>
<td>3.3</td>
<td>21.8</td>
</tr>
<tr>
<td>Average # of relations</td>
<td>4.5</td>
<td>1.2</td>
</tr>
</tbody>
</table>

Figure 1: Global Relationship Diagram

Figure 2: Schematic diagram of referential relationship labelling in the CRECIL corpus

Figure 3: Example of Dialogue-based Character Relationship Triples Generator in the CRECIL corpus

Conclusion & Future Work
- We presented a novel human-annotated dialogue-based relation extraction data set (CRECIL) for multi-party conversations in Chinese.
- We have introduced the Chinese-oriented character relationship categories and labelling rules for annotating the corpus.
- The results demonstrate that extracting character relationships is more challenging in CRECIL than in DialogRE.

Current Work: Explore the character relationship characteristics of Chinese multi-party dialogues and build a better-performance character relationship extraction model.

Acknowledgements
This work is supported by the National Natural Science Foundation of China (Grant No.61602044).

References