Task Description

- Task 1a: Identify the cited text spans in the reference paper according to the given citation.
- Task 1b: Identify the facet of the cited text span.
- Task 2: Generate a structured summary of the reference paper with the length not exceeding 250 words.

Framework of cited span identification

Methods Detail

For identification of cited text spans, we view it as a combination of information retrieval task and binary-classification task.

- Information retrieval task:
  - BM25 Model
  - Vector Space Model with TF-IDF
  - Language Model with Word Embedding

- Binary-classification task:
  - Random Forest Model
    - Jaccard similarity
    - BM25 similarity
    - Vectorized TF-IdF Similarity
    - AAN word embedding alignment
    - Average distance of AAN word embedding
    - Section similarity

- Two-layer Voting System
  - the combination of the both kinds of models with experimental threshold

For facet identification, we train random forest models for facet Method, Aim, Implication and Result respectively. The features are extracted from the Bag-of-Words model and positional information.

Results

<table>
<thead>
<tr>
<th>System id</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1127</td>
<td>0.1703</td>
<td>0.1357</td>
</tr>
<tr>
<td>2</td>
<td>0.1127</td>
<td>0.1703</td>
<td>0.1357</td>
</tr>
<tr>
<td>3</td>
<td>0.1116</td>
<td>0.2052</td>
<td>0.1446</td>
</tr>
<tr>
<td>4</td>
<td>0.1309</td>
<td>0.1703</td>
<td>0.1480</td>
</tr>
</tbody>
</table>

Conclusion

- Semantic-based citation identification is the main stream of the former exploration and the popular deep learning methods do not achieve satisfactory results because of the limitation of the scale of the dataset.

- the voting method is an effective strategy to improve the performance of the systems, because it synthesizes the advantages of each meta-model and reflects a powerful performance of forecasting.

- the distribution of facets to the reference sentences according to the citations is imbalanced and the summary merely extracted from the cited spans may not comprehensive and complete. Hence, how to combine citation information and other useful information for summary generation could be a consideration when doing scientific summarization.

References