OpenEL: An Annotated Corpus for Entity Linking and Discourse in Open Domain Dialogue

Wen Cui\textsuperscript{1}  
Marilyn A. Walker\textsuperscript{1}  
Leanne Rolston\textsuperscript{2}  
Beth Ann Hockey\textsuperscript{2}  

\textsuperscript{1}University of California, Santa Cruz  
\textsuperscript{2}LivePerson Inc.
Outline

● Motivations
● Challenges
● The OpenEL Corpus
● Experiments
● Conclusion & Future Work
Outline

● Motivations
● Challenges
● The OpenEL Corpus
● Experiments
● Conclusion & Future Work
Motivations

● Growth of interest in open-domain dialogue systems in the past few years
● Building open-domain dialogue systems converse like humans remains extremely challenging.
Motivations

- Named Entity Linking (NEL) is the task of mapping named entity mentions to a target knowledge base.
- Coreference resolution is the KEY to NEL over dialogue.

Figure 1: An annotated conversation in OpenEL
Outline

- Motivations
- Challenges
  - The OpenEL Corpus
  - Experiments
- Conclusion & Future Work
Challenges

- Name variations
- Open-ended nature
- Ambiguous reference

Figure 2: potentially ambiguous reference resolved by dialogue context
Contributions

● Release OpenEL with both NEL and anaphora annotations
● Analyze annotation quality
● Compare existing NEL tools and establish baselines
Outline

● Motivations
● Challenges
● The OpenEL Corpus
  ● Experiments
  ● Conclusion & Future Work
OpenEL Corpus: Annotation

- Sampled existing open-domain corpus Edina (Krause et al., 2017)
- Preprocessed using DBpedia Spotlight (Mendes et al., 2011)
- 3 experienced annotators verified/corrected spans and Wikidata IDs, marked coreference.
"text": "Mandy Moore used to sing. Did they ever record together?",

"speaker-id": "A",

"entities":
[
  "annotator-1": [
    "span": [0, 10],
    "surface-form": "Mandy Moore",
    "wikidata-id": ["Q187832"],
    "is-anaphora": false,
    "span": [31, 34],
    "surface-form": "they",
    "wikidata-id": ["Q160009", "Q187832"],
    "is-anaphora": true],

  "annotator-2": [
    "span": [0, 10],
    "surface-form": "Mandy Moore",
    "wikidata-id": ["Q187832"],
    "is-anaphora": false,
    "span": [31, 34],
    "surface-form": "they",
    "wikidata-id": ["Q160009", "Q187832"],
    "is-anaphora": true]
]

"annotator-3": [
  "span": [0, 10],
  "surface-form": "Mandy Moore",
  "wikidata-id": ["Q187832"],
  "is-anaphora": false,
  "span": [31, 34],
  "surface-form": "they",
  "wikidata-id": ["Q160009", "Q187832"],
  "is-anaphora": true],

"ground-truth": [
  "span": [0, 10],
  "surface-form": "Mandy Moore",
  "wikidata-id": ["Q187832"],
  "is-anaphora": false,
  "span": [31, 34],
  "surface-form": "they",
  "wikidata-id": ["Q160009", "Q187832"],
  "is-anaphora": true]}

Figure 3: JSON formatted examples in OpenEL
First substantial entity linking corpus publicly available for open-domain dialogue

<table>
<thead>
<tr>
<th>Corpus Properties: Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>conversations</td>
</tr>
<tr>
<td>conversational turns</td>
</tr>
<tr>
<td>topics</td>
</tr>
<tr>
<td>entity mentions (incl. anaphors)</td>
</tr>
<tr>
<td>entity mentions (excl. anaphors)</td>
</tr>
<tr>
<td>anaphors</td>
</tr>
<tr>
<td>unique entities</td>
</tr>
<tr>
<td>average conversation length (turns)</td>
</tr>
<tr>
<td>average mentions per entity</td>
</tr>
<tr>
<td>average mentions per conversational turn</td>
</tr>
</tbody>
</table>

Table 1: Counts and averages representing the composition of the corpus
## OpenEL Corpus: Statistics

<table>
<thead>
<tr>
<th>Topic</th>
<th># Convs</th>
<th># Turns</th>
<th>% Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>music</td>
<td>33</td>
<td>20</td>
<td>0.60 / 0.40</td>
</tr>
<tr>
<td>pop</td>
<td>32</td>
<td>10</td>
<td>0.82 / 0.60</td>
</tr>
<tr>
<td>star wars</td>
<td>30</td>
<td>10</td>
<td>0.72 / 0.62</td>
</tr>
<tr>
<td>baseball</td>
<td>34</td>
<td>20</td>
<td>0.53 / 0.42</td>
</tr>
<tr>
<td>comedy</td>
<td>22</td>
<td>10</td>
<td>0.78 / 0.61</td>
</tr>
<tr>
<td>rap hiphop</td>
<td>5</td>
<td>10</td>
<td>0.76 / 0.68</td>
</tr>
<tr>
<td>action</td>
<td>4</td>
<td>10</td>
<td>0.72 / 0.55</td>
</tr>
<tr>
<td>basketball</td>
<td>4</td>
<td>20</td>
<td>0.68 / 0.60</td>
</tr>
<tr>
<td>horror</td>
<td>4</td>
<td>10</td>
<td>0.53 / 0.30</td>
</tr>
<tr>
<td>movies</td>
<td>3</td>
<td>20</td>
<td>0.80 / 0.55</td>
</tr>
<tr>
<td>nfl football</td>
<td>4</td>
<td>20</td>
<td>0.51 / 0.37</td>
</tr>
<tr>
<td>rock</td>
<td>4</td>
<td>10</td>
<td>0.78 / 0.63</td>
</tr>
</tbody>
</table>

Table 2: Number of conversations per topic, number of turns per conversation, proportion of turns with an entity (incl anaphors / excl anaphors).
### OpenEL Corpus: Agreement

#### Table 3: Pairwise annotator agreement: Cohen Kappa and F-scores

<table>
<thead>
<tr>
<th></th>
<th>Cohen Kappa</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Span &amp; Wikidata ID</strong></td>
<td>0.79</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>0.82</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>0.82</td>
<td>0.68</td>
</tr>
<tr>
<td><strong>(average)</strong></td>
<td><strong>0.81</strong></td>
<td><strong>0.67</strong></td>
</tr>
<tr>
<td><strong>Span</strong></td>
<td>0.83</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>0.86</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>0.84</td>
<td>0.75</td>
</tr>
<tr>
<td><strong>(average)</strong></td>
<td><strong>0.84</strong></td>
<td><strong>0.74</strong></td>
</tr>
<tr>
<td><strong>Wikidata ID</strong></td>
<td>0.85</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>0.86</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>0.89</td>
<td>0.83</td>
</tr>
<tr>
<td><strong>(average)</strong></td>
<td><strong>0.87</strong></td>
<td><strong>0.79</strong></td>
</tr>
</tbody>
</table>
OpenEL Corpus: Disagreement

<table>
<thead>
<tr>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Star Wars topic)</td>
</tr>
<tr>
<td>A1: Do you like the star wars movies Q22092344?</td>
</tr>
<tr>
<td>A2: Do you like the star wars movies Q462?</td>
</tr>
<tr>
<td>A3: Do you like the star wars Q462 movies?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(Baseball topic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1: Wouldn’t think of missing it. I never miss that Q213417 match up.</td>
</tr>
<tr>
<td>A2: Wouldn’t think of missing it. I never miss that match up Q213417.</td>
</tr>
<tr>
<td>A3: Wouldn’t think of missing it. I never miss that match up.</td>
</tr>
</tbody>
</table>
Outline

● Motivations
● Challenges
● The OpenEL Corpus
● Experiments

● Conclusion & Future Work
A: Now if Belt will just start hitting ...
B: Dude, if he can keep getting on base by taking walks, I don’t care if he never gets another home run.
A: A walk’s as good as a single. I’ll take it.
B: You hate to see all that power go to waste, though.
B: Looks like Posey’s all recovered from his concussion.

Experiments

- Setup: **UTTERANCE**, **DIALOGUE**, **DISCOURSE**
A: Now if Belt will just start hitting ...
B: Dude, if he can keep getting on base by taking walks, I don’t care if he never gets another home run.
A: A walk’s as good as a single. I’ll take it.
B: You hate to see all that power go to waste, though.
B: Looks like Posey’s all recovered from his concussion.
A: Now if Belt will just start hitting ...
B: Dude, if Belt can keep getting on base by taking walks, I don’t care if Belt never gets another home run.
A: A walk’s as good as a single. I’ll take it.
B: You hate to see all that power go to waste, though.
B: Looks like Posey’s all recovered from his concussion.
Experiments

- **Baseline systems**
  - DBpedia Spotlight (Mendes et al., 2011)
  - WAT (Piccinno et al., 2014): MD (SVM) + ED (PageRank) + Pruning (SVM) trained on Wikipedia articles
  - REL (Hulst et al., 2020): NER (Flair) + ED (MLP) trained on AIDA
  - Flair (Akbik et al., 2019) + BLINK (Li et al., 2020): BLINK is a BERT based ED system
Figure 4: NEL evaluation result (strict match). *indicates statistically significant compared to UTTERANCE setting with $p<0.05$
Posey’s all recovered from his concussion.
A: Now if Belt will just start hitting . . .
B: Dude, if he can keep getting on base by taking walks, I don’t care if he never gets another home run.
A: A walk’s as good as a single. I’ll take it.
B: You hate to see all that power go to waste, though.
B: Looks like Posey’s all recovered from his concussion.

A: Now if Belt will just start hitting . . .
B: Dude, if Belt can keep getting on base by taking walks, I don’t care if Belt never gets another home run.
A: A walk’s as good as a single. I’ll take it.
B: You hate to see all that power go to waste, though.
B: Looks like Posey’s all recovered from his concussion.

(Posey, Posey County Q6307475)
(Posey, James Posey, Q717793)
(Posey, Buster Posey, Q971912)
Outline

● Motivations
● Challenges
● The OpenEL Corpus
● Experiments
● Conclusion & Future Work
Conclusion & Future Work

- Presented OpenEL corpus with very high IAA
- Demonstrated the gap between baselines and human performance
- Future work
  - Full discourse representation
  - Off-the-shelf coreference
Thank you!

For code and data:

https://github.com/wenzi3241/OpenEL_corpus