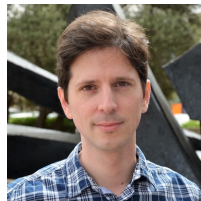
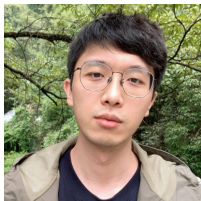




OntoGUM: Evaluating Contextualized SOTA Coreference Resolution on 12 More Genres

Yilun Zhu¹, Sameer Pradhan^{2,3}, Amir Zeldes¹



¹ *GEORGETOWN UNIVERSITY*

²   **Penn**
UNIVERSITY of PENNSYLVANIA

³ cemantix.org

The Coreference Resolution Task

*“I voted for **Mary** because **Mary** was most aligned with **my** values”, **John** said.*

Problems of Existing Coreference Datasets

| Datasets | Out-of ON domain | Scheme compatibility | Multi-genre | Multi-coreference types | Singletons |
|--|------------------|----------------------|-------------|-------------------------|------------|
| OntoNotes (Pradhan et al., 2013) | ✗ | ✓ | ✓ | ✓ | ✗ |
| WikiCoref (Ghaddar and Langlais, 2016) | ✓ | ✓ | ✗ | ✓ | ✗ |
| GAP (Webster et al., 2018) | ✓ | ✓ | ✗ | ✗ | ✗ |
| GUM (Zeldes, 2017) | ✓ | ✗ | ✓ | ✓ | ✓ |
| ARRAU (Poesio et al., 2018) | ✓ | ✗ | ✓ | ✓ | ✓ |
| PreCo (Chen et al., 2018) | ✓ | ✗ | ✓ | ✓ | ✓ |

Problems of Existing Out-of-domain Evaluation

- **No study** has investigated if **contextualized embeddings** encounter the same **overfitting problem** identified by Moosavi and Strube (2017)
- Previous work may **underestimate the performance degradation** on **WikiCoref**
 - embeddings were also trained on Wikipedia themselves (Moosavi and Strube, 2018)
 - -> higher coreference scores on Wikipedia texts

OntoGUM Dataset

- Conversion from GUM using **gold standard syntax trees**
- Statistics
 - 168 documents with 12 genres, ~150K tokens
 - 19,378 mentions, 4,471 clusters
- Genres
 - Text: News / Fiction / Bio / Academic / Forum / Travel / How-to / Textbook
 - Speech: Interview / Political / Vlog / Conversation



<https://github.com/yilunzhu/ontogum>

Dataset Conversion

- OntoNotes \subseteq GUM
 - **Don't need** human annotation to recognize additional mentions in the conversion process
- **Annotation layers** used in the conversion
 - Coreference layer
 - Gold syntax trees

| | |
|----------|--|
| GUM: | <i>Kim visited Seoul ... The visit ...</i> |
| OntoGUM: | <i>Kim visited Seoul ... The visit ...</i> |

- Gold speaker information (fiction, reddit and spoken data)
- Annotation agreement
 - Agreement study on 3 docs (2,500 tokens, 371 mentions), **8/371** errors
 - Span detection: **~0.96** CoNLL coreference score: **~0.92**

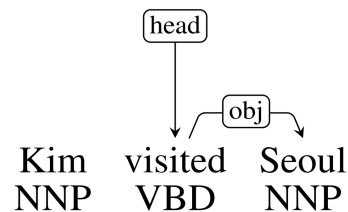



Figure1: Gold Syntax in GUM

Experiments & Results 1/3

| Models | OntoNotes | OntoGUM |
|---|-----------|---------|
| dcoref (Manning et al. 2014, CoreNLP) | 57.8 | |
| e2e + SpanBERT (Joshi et al., 2019, SOTA) | 79.6 | |

Experiments & Results 1/3

| Models | OntoNotes | OntoGUM |
|---|-----------|---------|
| dcoref (Manning et al. 2014, CoreNLP) | 57.8 | 39.7 |
| e2e + SpanBERT (Joshi et al., 2019, SOTA) | 79.6 | 64.6 |



- Both systems encounter **a substantial degradation** on **OntoGUM**

Experiments & Results 2/3

- **Genre disparity** does not guarantee low performance (e.g., **vlog**), and errors occur readily even in **overlapping genres** (e.g., **news**)
- Performance is correlated with the proportions of pronouns

| | PRON (R) | Other (R) | Total | CoNLL | Span |
|-----------|------------|------------|-------|-------|------|
| <i>vl</i> | 600 (.66) | 309 (.34) | 909 | 1 | 1 |
| <i>it</i> | 1223 (.45) | 1485 (.55) | 2708 | 2 | 6 |
| <i>cn</i> | 729 (.61) | 323 (.39) | 1052 | 3 | 2 |
| <i>sp</i> | 245 (.40) | 364 (.60) | 609 | 4 | 4 |
| <i>bi</i> | 796 (.34) | 1529 (.66) | 2325 | 5 | 3 |
| <i>fc</i> | 1700 (.61) | 1091 (.39) | 2791 | 6 | 5 |
| <i>ac</i> | 262 (.21) | 997 (.79) | 1259 | 7 | 10 |
| <i>vy</i> | 300 (.22) | 1053 (.78) | 1353 | 8 | 7 |
| <i>rd</i> | 1337 (.55) | 1077 (.45) | 2414 | 9 | 8 |
| <i>nw</i> | 340 (.19) | 1483 (.81) | 1823 | 10 | 9 |
| <i>wh</i> | 1001 (.47) | 1129 (.53) | 2130 | 11 | 11 |
| <i>tx</i> | 165 (.34) | 315 (.66) | 480 | 12 | 12 |

Table 1: Genre-breakdown Statistics of OntoGUM

Experiments & Results 3/3

- Genre disparity does not guarantee low performance (e.g., vlog), and errors occur readily even in overlapping genres (e.g., news)
- Performance is correlated with the **proportions of pronouns or gold speaker information**

| | PRON (R) | Other (R) | Total | CoNLL | Span |
|-----------|------------|------------|-------|-------|------|
| <i>vl</i> | 600 (.66) | 309 (.34) | 909 | 1 | 1 |
| <i>it</i> | 1223 (.45) | 1485 (.55) | 2708 | 2 | 6 |
| <i>cn</i> | 729 (.61) | 323 (.39) | 1052 | 3 | 2 |
| <i>sp</i> | 245 (.40) | 364 (.60) | 609 | 4 | 4 |
| <i>bi</i> | 796 (.34) | 1529 (.66) | 2325 | 5 | 3 |
| <i>fc</i> | 1700 (.61) | 1091 (.39) | 2791 | 6 | 5 |
| <i>ac</i> | 262 (.21) | 997 (.79) | 1259 | 7 | 10 |
| <i>vy</i> | 300 (.22) | 1053 (.78) | 1353 | 8 | 7 |
| <i>rd</i> | 1337 (.55) | 1077 (.45) | 2414 | 9 | 8 |
| <i>nw</i> | 340 (.19) | 1483 (.81) | 1823 | 10 | 9 |
| <i>wh</i> | 1001 (.47) | 1129 (.53) | 2130 | 11 | 11 |
| <i>tx</i> | 165 (.34) | 315 (.66) | 480 | 12 | 12 |

Table 1: Genre-breakdown Statistics of OntoGUM

Conclusion

- We release the **largest open, gold, coreference dataset with new genres** following **the OntoNotes scheme**
- Results showed **a lack of generalizability of existing systems**, especially in **genres low in pronouns and lacking speaker information**
- A **genre-by-genre** analysis reveals relative **strengths and weaknesses** of current approaches

Conclusion

- We release the largest open, gold, coreference dataset with new genres following the OntoNotes scheme
- Results showed a lack of generalizability of existing systems, especially in genres low in pronouns and lacking speaker information
- A genre-by-genre analysis reveals relative strengths and weaknesses of current approaches

We hope people can use **OntoGUM as an out-of-domain benchmark for systems developed using **OntoNotes!****