Gendered Pronoun Resolution

Mustafa Abdool and Sarah Egler
{moose878, segler}@stanford.edu

Objective

Coreference resolution is the task of identifying all tokens in text which refer to the same entity. Current state-of-the-art systems show high gender bias in male vs. female pronoun resolution performance due to imbalanced datasets.

Data

Gendered Ambiguous Pronoun (GAP) by Google AI:
- Gender balanced
- 4000 train, 4000 test, 908 validation examples

In May, Fujisawa joined Mari Motohashi’s rink as the team’s skip, moving back from Karuizawa to Kitami where she had spent her junior days.

Target Pronoun: she, Candidate A: Fujisawa, Candidate B: Mari Motohashi

Ground Truth Class (one of: A, B, NEITHER): A

Methods

Distance Features (baseline): Distance (in characters and words) from candidate to pronoun, total length of text, offset of candidate.

Syntactic Structure Features: Count of pronouns in text, count of pronouns between candidate and target, number of times candidate appears in the text, syntactic dependency of candidate and pronoun, relative sentence position.

Word embeddings (Gender-Neutral-GloVe / GloVe): target pronoun, head word of candidate, head word of target pronoun

Character level embeddings: Use character embeddings and a character based CNN to generate word embeddings for key words. Learned from scratch.

LSTM: Encode span of text around candidate using bi-directional LSTM

BERT: Fine-tune BERT output layer after extracting embedding of target pronoun and candidate.

Results

F1 Scores by gender (M/F) and overall (O)

<table>
<thead>
<tr>
<th>Model</th>
<th>M</th>
<th>F</th>
<th>Bias</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>spaCy NeuralCoref Baseline</td>
<td>54.2</td>
<td>49.9</td>
<td>0.92</td>
<td>52.3</td>
</tr>
<tr>
<td>Distance Features Baseline</td>
<td>57.2</td>
<td>58.3</td>
<td>1.02</td>
<td>77.1</td>
</tr>
<tr>
<td>Hand-Engineered Features</td>
<td>69.1</td>
<td>67.9</td>
<td>0.98</td>
<td>68.5</td>
</tr>
<tr>
<td>GN-GloVe Word Embeddings</td>
<td>71.0</td>
<td>70.7</td>
<td>0.99</td>
<td>70.8</td>
</tr>
<tr>
<td>Character Embeddings</td>
<td>71.7</td>
<td>70.6</td>
<td>0.98</td>
<td>71.2</td>
</tr>
<tr>
<td>LSTM</td>
<td>69.2</td>
<td>70.6</td>
<td>1.02</td>
<td>69.9</td>
</tr>
<tr>
<td>BERT*</td>
<td>74.0</td>
<td>71.7</td>
<td>0.97</td>
<td>72.8</td>
</tr>
</tbody>
</table>

Best Overall Model:
- Hand-engineered features ensemble with fine-tuning pre-trained BERT

Least Gender biased:
- Word Embeddings +
- Hand-Engineered Features

Analysis

Best Performance: Clear cues for syntactic parallelism captured in our features

Philippe Burty (6 February 1830 – 3 June 1890) was a French art critic. He contributed to the popularization of Japonism and the revival of etching, supported the Impressionists, and published the letters of Eugène Delacroix.

Target Pronoun: he, Candidate A: Philippe Burty, Candidate B: Eugène Delacroix

Ground Truth Class: A

Worst Performance: “Neither” class

They then signed with their new label Enter Hama, where they will re-debut as Elate. On 26 June, Rania performed in a China event with 3 DR Music trainees: Jian, Jieun (a former member of LPG) and Crystal. On 15 August 2016, Alex announced that she was the group’s new leader.

Target Pronoun: she, Candidate A: Rania, Candidate B: Crystal

Ground Truth Class: Neither

Gender Bias
Training on a gender-balanced dataset mitigates gender bias.

GN-GloVe does not improve bias over GloVe in the context of this dataset.

Future Work

Retrain state-of-the-art models on the larger OntoNotes replacing all pronouns with gender neutral “ze”

Overcome limitations of small dataset size and gain more context with Wikipedia text with link provided as part of each example.

References

