Gendered Pronoun Resolution

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

Stanford
CS224n, Winter 2019

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.

The surgeon | explained to | the patient | how | she | would operate on | him

Data

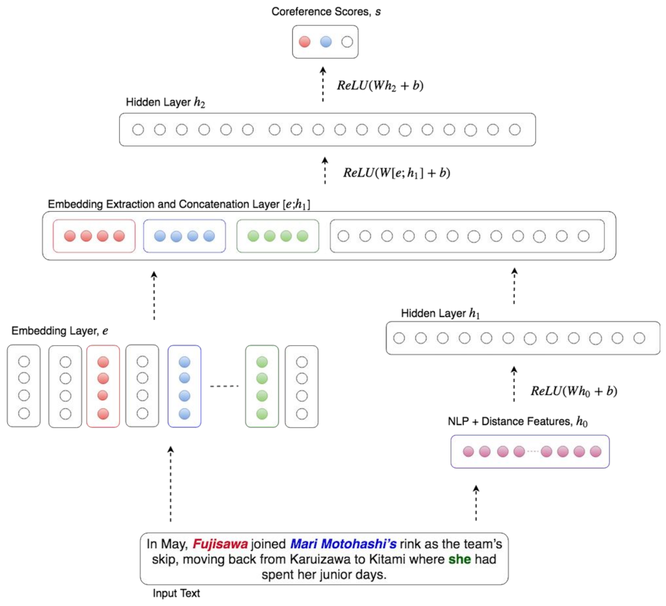
Gendered Ambiguous Pronoun (GAP) by Google AI[1]

- 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**

Architecture



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.[2]

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 CNNto generate word embeddings for key words. Learned from scratch.

LSTM: Encode span of text around candidate using bi-directional LSTM[3]

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

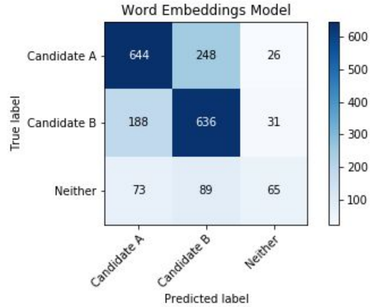
Results

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

Model	M	F	Bias	O
spaCy NeuralCoref Baseline	54.6	49.9	0.92	52.3
Distance Features Baseline	57.2	58.3	1.02	57.7
Hand-Engineered Features	69.1	67.9	0.98	68.5
GN-GloVe Word Embeddings	71.0	70.7	**0.99**	70.8
Character Embeddings	71.7	70.6	0.98	71.2
LSTM	69.2	70.6	1.02	69.9
BERT*	**74.0**	**71.7**	0.97	**72.8**

Best Overall Model:
Hand-engineered features ensembled 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 ***Eug*ne Delacroix***.

Target Pronoun: he, **Candidate A:** Philippe Burty, **Candidate B:** Eug*ne Delacroix
Ground Truth Class : A

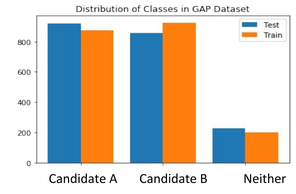
Worst Performance: "Neither" class

They then signed with their new label Enter Hama, where they will re-debut as Ela8te. 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

[1] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. arXiv preprint arXiv:1810.05201.

[2] Clark, K. and Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations. CoRR, abs/1606.01323.

[3] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In Proceedings of EMNLP, pages 188–197, Copenhagen, Denmark.