

# Assessment of Gender Bias in Coreference Resolution

Ines Reinig


December 10, 2019

## Intro

*A man and his son get into a terrible car crash. The father dies, and the boy is badly injured. In the hospital, the surgeon looks at the patient and exclaims, "I can't operate on this boy, he's my son!"*

How can this be?<sup>1</sup>

---

<sup>1</sup>This riddle introduces the study of Rudinger et al. (2018) 

# Overview

Coreference resolution (quick recap)

Rudinger et al. (2018)

Webster et al. (2018)

Comparison of both gender bias studies

Discussion

## Coreference resolution (quick recap)

coreferring mention



*Victoria Chen, Chief Financial Officer of Megabucks Banking Corp since 2004, saw **her** pay jump 20%, to \$1.3 million, as the 37-year-old also became the Denver-based financial-services company's president. It has been ten years since **she** came to Megabucks from rival Lotsabucks.*<sup>2</sup>

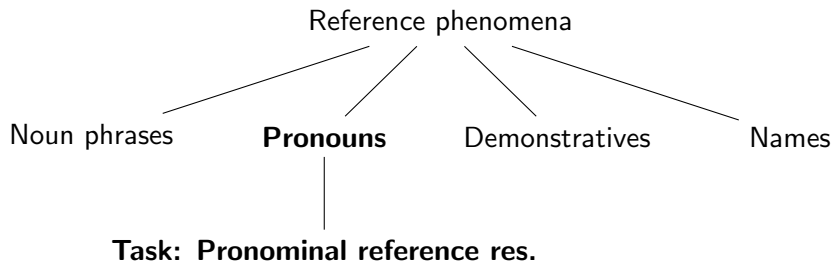
Coreference chains (examples):

- ▶ {Victoria Chen, Chief Financial Officer of Megabucks Banking Corp since 1994, **her**, the 37-year-old, the Denver-based financial-services company's president, **she**}
- ▶ {Megabucks Banking Corp, the Denver-based financial-services company, Megabucks}

---

<sup>2</sup>This example is from Jurafsky et al., ch. 21.

## Coreference resolution (quick recap)



## Resolving ambiguity in coref. res.

→ Morphosyntactic restrictions filter the set of candidate mentions:

- ▶ John has two daughters. They are still young. (person agreement)
- ▶ Mary left the bicycle in the garage after driving it around for hours. (selectional restriction)
- ▶ Mary is showing us the bicycle. It looks terrific. (gender agreement)

## Resolving ambiguity in coref. res.

Back to the riddle:

The surgeon couldn't operate on her patient: it was her son!

The Stanford CoreNLP coreference system fails to link the female pronoun "her" to the NP "The surgeon".

Coreference resolution systems can exhibit **gender bias**

## Coreference res. systems can be biased

- ▶ Rudinger et al. (2018): eval. dataset for **pronoun-occupation** pairs
- ▶ Webster et al. (2018): eval. dataset for **pronoun-named entity** pairs

Both studies have a common goal: reveal gender bias in coreference systems for pronominal reference resolution



Rudinger et al. (2018): Winogender

# Bias evaluation in Rudinger et al. 2018

3 coreference resolution systems are evaluated on “Winogender” style instances:

- ▶ Rule-based system (Lee et al. 2011)
- ▶ Statistical approach (Durrett & Klein 2013)
- ▶ Neural model (Clark & Manning 2016)

## Rule-based system (Lee et al. 2011)

- ▶ A system using hand-crafted rules related to lexical, syntactic, semantic & discourse information
- ▶ Three-step system: 1) mention detection, 2) mention processing and 3) post-processing

# Statistical approach (durrett)

- ▶ New state of the art in 2013
- ▶ No more hand-made features, instead data-driven feature templates
- ▶ Main difference to Lee et al. (2011): only shallow surface features are used

# Neural approach (Clark & Manning 2016)

- ▶ Deep neural mention-ranking model
- ▶ State-of-the-art in 2016
- ▶ Uses concatenated word embeddings in input layer

# Evaluation using Winogender instances

How fair are these three coreference systems?

⇒ Rudinger et al. propose a dataset specifically designed to assess **gender bias** in the systems:

- ▶ RQ: is a coref. system likely to associate a pronoun with an **occupation** based on gender (*she* vs. *he*)?
- ▶ Exclusively unambiguous pronoun resolution (human-validated)

# Winogender

<b>Occup.</b>	<b>Part.</b>	<b>Template</b>
firefighter	<u>child</u>	The OCCUPATION had to rescue the PARTICIPANT from the burning building because NOM_PRONOUN could not escape.
<u>firefighter</u>	child	The OCCUPATION had to rescue the PARTICIPANT from the burning building because NOM_PRONOUN could not just stand by and do nothing.
chemist	<u>visitor</u>	The OCCUPATION told the PARTICIPANT that NOM_PRONOUN would need to put on safety glasses before entering the laboratory.
<u>chemist</u>	visitor	The OCCUPATION told the PARTICIPANT that NOM_PRONOUN would be delighted to give a tour of the laboratory.

- ▶ 120 hand-written templates following Winograd schema
- ▶ 720 sentences (60 occupations × 2 sentence templates per occup. × 2 participants × 3 pronoun genders)

# Winogender - more examples

<b>Occup.</b>	<b>Part.</b>	<b>Template</b>
<u>counselor</u>	patient	The OCCUPATION disclosed to the PARTICIPANT that NOM_PRONOUN was professionally mandated to report certain issues.
counselor	<u>patient</u>	The PARTICIPANT disclosed to the OCCUPATION that NOM_PRONOUN had a history of substance abuse.
supervisor	<u>employee</u>	The OCCUPATION gave the PARTICIPANT feedback on POSS_PRONOUN stellar performance.
<u>supervisor</u>	employee	The PARTICIPANT gave the OCCUPATION feedback on POSS_PRONOUN managing style.
inspector	<u>homeowner</u>	The PARTICIPANT asked the OCCUPATION if the house NOM_PRONOUN had purchased was structurally sound.
<u>inspector</u>	homeowner	The PARTICIPANT asked the OCCUPATION if NOM_PRONOUN had discovered any building code violations.



# Evaluation

- ▶ In the Winogender style dataset, **correct pronoun resolution is not a function of gender**
- ▶ However all systems are not gender-neutral:
  - ▶ Male pronouns more likely to be associated with occupation than female or neutral
  - ▶ Correlation with real-word employment stats

# Evaluation

Bias is observed from 3 perspectives:

- ▶ Likelihood of resolving m vs. f pronouns as occupation
- ▶ Accuracy on “gotcha” sentences
- ▶ Correlation with real-world employment statistics (biased)

# Evaluation

68% of male-female sentence pairs are resolved differently by rule-based system

<b>System</b>	<b>male</b>	<b>female</b>	<b>neutral</b>
RULE	72	29	1
STAT	71	63	50
NEURAL	87	80	36

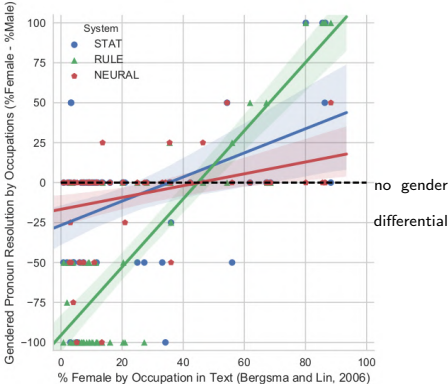
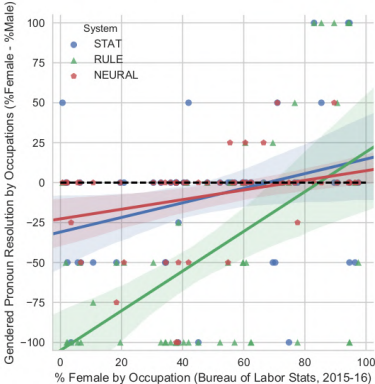
**Table:** Likelihood (%) of pronouns to be resolved as occupation

# Evaluation

<b>System</b>	<b>"Gotcha"?</b>	<b>female</b>	<b>male</b>
RULE	no	38.3	51.7
	yes	10.0	37.5
STAT	no	50.8	61.7
	yes	45.8	40.0
NEURAL	no	50.8	49.2
	yes	36.7	46.7

**Table:** Accuracy (%) by gender and difficulty

# Evaluation



y=-100 means maximum male bias, i.e. the system always resolves male pronouns to given occupation

# Evaluation & findings in Rudinger et al.

- ▶ The 3 rule-based, statistical and neural systems are biased towards male gender
- ▶ Furthermore: they **amplify biases** existing in real-world situations, e.g. occupational gender statistics

Why can we say that they amplify biases? they make **discrete choices**

Example: female managers

# Sources of bias in coref. systems

Which aspects might cause gender bias in the three systems?

- ▶ rule-based: semantic sieves using online knowledge bases/encyclopedia
- ▶ statistical: data
- ▶ neural: data & pre-trained word embeddings

Webster et al. (2018): GAP



# Webster et al: GAP

## Contributions:

- ▶ Build dataset “GAP” to evaluate bias of coref. systems
- ▶ Evaluate 4 off-the-shelf resolvers on GAP
- ▶ Propose several baselines for coref. res. on GAP

# Evaluated off-the-shelf resolvers

- ▶ One rule-based architecture (Lee et al. 2013)
- ▶ Three neural resolvers:
  - ▶ Clark & Manning (2015)
  - ▶ Wiseman et al. (2016)
  - ▶ Lee et al. (2017)

# Evaluation using GAP

How fair are these four coreference systems?

Webster et al. build a dataset GAP specifically designed to assess gender bias in the system:

- ▶ Goal similar to Rudinger et al.
- ▶ But GAP differs from Winogender in several ways

## GAP: overview

NE 1	NE 2	Sentence
Jose de Venecia Jr (FALSE)	Abalos (FALSE)	[...] Jose de Venecia III, son of House Speaker <u>Jose de Venecia Jr</u> , alleged that <u>Abalos</u> offered <b>him</b> US\$10 million to withdraw his proposal on the NBN project.
Sophie (FALSE)	Jeni (TRUE)	[...] The remaining trio head back to the cottage [...], but the leprechaun tricks <u>Sophie</u> and Ben into striking <u>Jeni</u> with their axes, killing <b>her</b> .
Malave (TRUE)	Greg Joiner (FALSE)	[...] <u>Malave</u> took a fight in Boston, Mass. against <u>Greg Joiner</u> , winning by a knockout in the 3rd round. Then <b>he</b> faced former World Lightweight Champion Ken Buchanan [...].

- ▶ Domain: Wikipedia (ideal?)
- ▶ Large dataset: 8,9k instances (human-annotated,  $\kappa = 0.74$ )

# GAP: extraction

Constraints are applied during extraction:

- ▶ Only 3 possible patterns
- ▶ Extracted sentences are sub-sampled to ensure broad coverage of domains, balanced m:f ratio & balanced pattern ratio

Goal: build a balanced dataset & limit success of naïve coreference systems

# Evaluation

- ▶ Bias measure: ratio of f to m F1-scores
- ▶ Simple and interpretable measure of bias

$$B = \frac{F1_f}{F1_m}$$

- ▶ B **close to 1**: little to no bias (ideal)
- ▶ B close to 0: masculine bias
- ▶ B above 1: feminine bias

# Evaluation

- ▶ In GAP as well, correct pronoun resolution is not a function of gender
- ▶ But all evaluated resolvers **favor better resolution of masculine pronouns**

Model	M	F	<b>B</b>	O
Lee et al. 2013	55.4	45.5	<b>0.82</b>	50.5
Clark & Manning	58.5	51.3	0.88	55.0
Wiseman et al.	68.4	59.9	0.88	64.2
Lee et al. 2017	67.2	62.2	<b>0.92</b>	64.7

**Table:** Performance & bias of off-the-shelf-resolvers on GAP dev set

Performance on OntoNotes test set is comparable, however Lee et al. (2017) has highest bias (0.75)

## Evaluation & findings in Webster et al. (2018)

- ▶ F1-scores overall on GAP & OntoNotes not very high (low recall because of conservativeness)
- ▶ Gender bias present in all systems on OntoNotes and GAP, even though GAP is gender-balanced



## Webster et al. coreference baselines

Model	M	F	<b>B</b>	O
Lee et al. 2017	67.2	62.2	0.92	64.7
Random	43.6	39.3	0.90	41.5
Token Dist.	50.1	42.4	0.85	46.4
Parallelism	<b>67.1</b>	<b>63.1</b>	<b>0.94</b>	<b>65.2</b>
Parallelism+URL	<b>71.1</b>	<b>66.9</b>	<b>0.94</b>	<b>69.0</b>
Transf.-Single	58.6	51.2	0.87	55.0
Transf.-Multi	59.3	52.9	0.89	56.2

Table: Performance & bias of several models & baselines on GAP dev set

# Overview of GAP & Winogender

Same application, different approaches:

	GAP	Winogender
Domain	Wikipedia	Occupations
Reference	NE	Nominal mention
Annotation	5 choices	binary choice
Evaluation	F1-scores ratio	3 measures

# Comparison of GAP & Winogender

## Advantages of GAP:

- ▶ Closer to real-world data (vs. artificially created)
- ▶ Large
- ▶ Broad domain coverage

## Advantages of Winogender:

- ▶ Strict Winograd-like schema  $\Rightarrow$  allows precise observations
- ▶ Occupational domain allows comparison to statistics

# Conclusion

Positive in both studies:

- ▶ Careful construction of dataset
- ▶ Variety of evaluated resolvers
- ▶ Evaluation: extensive in Rudinger et al. and simple & interpretable in Webster et al.

# Discussion

## Questions:




- ▶ Rudinger et al.: Winogender schema show the **presence** of gender bias in coref. systems. But can they prove its **absence**?

# Discussion





## Questions:

- ▶ Rudinger et al.: Winogender schema show the **presence** of gender bias in coref. systems. But can they prove its **absence**?
- ▶ What would you pay attention to when trying to build a coref system that is as gender neutral as possible?

# References





-  Bergsma, Shane, and Dekang Lin. "Bootstrapping path-based pronoun resolution." Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics. 2006.
-  Clark, Kevin, and Christopher D. Manning. "Entity-centric coreference resolution with model stacking." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015.
-  Clark, Kevin, and Christopher D. Manning. "Deep reinforcement learning for mention-ranking coreference models." Empirical Methods on Natural Language Processing (EMNLP). 2016.

# References




-  Durrett, Greg, and Dan Klein. "Easy victories and uphill battles in coreference resolution." Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013.
-  Jurafsky, Dan, and James H. Martin. "Computational Discourse." Speech and language processing. An introduction to natural language processing, computational linguistics, and speech recognition. 2nd ed, Prentice Hall, Pearson Education International. 2009.
-  Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the fifteenth conference on computational natural language learning: Shared task. Association for Computational Linguistics. 2011.
-  Lee, Heeyoung, et al. "Deterministic coreference resolution based on entity-centric, precision-ranked rules." Computational Linguistics 39.4 (2013): 885-916.




# References

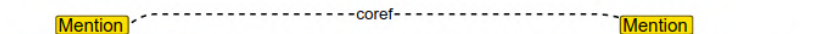
-  Lee, Kenton, et al. "End-to-end neural coreference resolution." Proceedings of EMNLP (2017): 188-197.
-  Levesque, Hector, Ernest Davis, and Leora Morgenstern. "The winograd schema challenge." Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning. 2012.
-  Pradhan, Sameer, et al. "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes." Joint Conference on EMNLP and CoNLL-Shared Task. Association for Computational Linguistics. 2012.
-  Rudinger, Rachel, et al. "Gender bias in coreference resolution." Proceedings of NAACL. 2018.

# References

-  Webster, Kellie and Recasens, Marta and Axelrod, Vera and Baldrige, Jason. "Mind the GAP: A balanced corpus of gendered ambiguous pronouns." Transactions of the Association for Computational Linguistics 6 (2018): 605-617.
-  Wiseman, Sam, Alexander M. Rush, and Stuart M. Shieber. "Learning global features for coreference resolution." Proceedings of NAACL-HLT (2016): 994-1004.
-  Zhao, Jieyu, et al. "Gender bias in coreference resolution: Evaluation and debiasing methods." Proceedings of NAACL. 2018.

## [extra slide] More Stanford CoreNLP examples

  
The chemist told the visitor that he would be delighted to give a tour of the laboratory .  
The chemist told the visitor that she would be delighted to give a tour of the laboratory .

  
The homeowner called the plumber to get an estimate for his services .  
The homeowner called the plumber to get an estimate for her services .