ANLP Lecture 29: Gender Bias in NLP

Sharon Goldwater

19 Nov 2019

Recap

- Some co-reference examples can't be solved by agreement, syntax, or other local features, but require semantic information ("world knowledge"?):

The [city council]$_i$ denied [the demonstrators]$_j$ a permit because…
…[they]$_i$ feared violence.
…[they]$_j$ advocated violence.

- NLP systems don't observe the world directly, but do learn from what people talk/write about.
- With enough text, this seems to work surprisingly well…
  - … but may also reproduce human biases, or even amplify or introduce new ones (depending on what we talk about and how).

Example: gender bias

The secretary read the letter to the workers. He was angry.

The secretary read the letter to the workers. She was angry.

- People have a harder time processing **anti-stereotypical** examples than **pro-stereotypical** examples.
- What about NLP systems? Is there algorithmic bias? E.g., do NLP systems
  - Produce more errors for female entities than males?
  - Perpetuate or amplify stereotypical ideas or representations?

Today's lecture

- What are some examples of gender bias in NLP and what consequences might these have?
- What is a challenge dataset and how are these used to target specific problems like gender bias?
- For one specific example (gender bias in coreference),
  - How can we systematically measure (aspects of) this bias?
  - What are some sources of the bias?
  - What can be done to develop systems that are less biased?

Biased scores in coref, language modelling

- Internal scores indicate implicit bias in coreference resolution and language modelling (Lu et al., 2019):

$1_\square$: The **doctor** ran because **he** is late. (5.08)	$1_\square$: **He** is a | **doctor**.	-9.72
$1_\bigcirc$: The **doctor** ran because **she** is late. (1.99)	$1_\bigcirc$: **She** is a | **doctor**.	-9.77
$2_\square$: The **nurse** ran because **he** is late. (−0.44)	$2_\square$: **He** is a | **nurse**.	-8.99
$2_\bigcirc$: The **nurse** ran because **she** is late. (5.34)	$2_\bigcirc$: **She** is a | **nurse**.	-8.97
(a) Coreference resolution	(b) Language modeling	$\ln \Pr[B \mid A]$

Figure 1: Examples of gender bias in coreference resolution and language modeling as measured by coreference scores (left) and conditional log-likelihood (right).

Machine translation errors

- Translating from English to Hungarian or Turkish (no gender) and back to English:

She is a janitor. He is a nurse.

↓

He's a janitor. She is a nurse.

- Translating English to Spanish (all nouns have gender).
  - Female doctor becomes male; nurse becomes female:

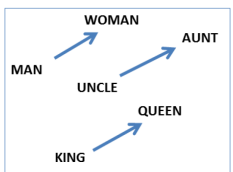
The doctor asked the nurse to help **her** in the procedure

El doctor le pidio a la enfermera que le ayudara con el procedimiento

Example 1: Google Translate, 17 Nov 2019; Example 2 from Stanovsky et al. (2019)

Word embeddings

- Famously, word embeddings can (approximately) solve analogies like `man:king :: woman:x`
  - Nearest vector to $v_{man} - v_{woman} + v_{king}$ is v_{queen}



- Almost as famously, pretrained word2vec vectors also say `man:computer programmer :: woman:homemaker` (Bolukbasi, 2016).
  - All due to word associations in the training data!

Figure: Mikolov et al. (2013)

Two kinds of implications

- Representation bias: when systems negatively impact the representation (social identity) of certain groups.
  - Implying that women should be homemakers
  - Guessing that doctors are male when translating from Hungarian.
  - Rating sentences with female noun phrases as more likely to be angry.
- Allocation bias: unfairly allocating resources to some groups.
  - Recommending to interview qualified men more often than qualified women because of irrelevant male-oriented words in their CVs that are similar to those in existing employees' CVs.

See Sun et al. (2017), citing Crawford (2017) and others.

Gender bias in coreference resolution

- Zhao et al. (2018) present work where they
  - Create a **challenge dataset** to quantify gender bias in co-reference systems.
  - Show significant gender bias in three different types of systems.
  - Identify some sources of bias and ways to de-bias systems.

Challenge dataset

- Most NLP systems are trained and tested on text sampled from natural sources (news, blogs, Twitter, etc)
- These can tell us how well systems do on average, but harder to understand specific strengths/weaknesses
- One way to investigate these: design a dataset specifically to test them.
- Typically small and used only for (dev and) test; training is still on original datasets.

The WinoBias dataset

- Based on Winograd schema idea; tests gender bias using pairs of pro-/anti-stereotypical sentences:

Pro: [The physician]$_i$ hired [the secretary]$_j$ because [he]$_i$ was overwhelmed with clients.
Anti: [The physician]$_i$ hired [the secretary]$_j$ because [she]$_i$ was overwhelmed with clients.

Pro: [The physician]$_i$ hired [the secretary]$_j$ because [she]$_j$ was highly recommended.
Anti: [The physician]$_i$ hired [the secretary]$_j$ because [he]$_j$ was highly recommended.

- Compute the difference in average accuracy between pro-stereotypical and anti-stereotypical sentences.

The WinoBias dataset

- Also includes "Type 2" sentence pairs, such as:

Pro: [The physician]$_i$ called [the secretary]$_j$ and told [her]$_j$ to cancel the appointment.
Anti: [The physician]$_i$ called [the secretary]$_j$ and told [him]$_j$ to cancel the appointment.

- What's different about these? Would you expect them to show more or less bias than Type 1 pairs (below)? Why?

Pro: [The physician]$_i$ hired [the secretary]$_j$ because [he]$_i$ was overwhelmed with clients.
Anti: [The physician]$_i$ hired [the secretary]$_j$ because [she]$_i$ was overwhelmed with clients.

Pro: [The physician]$_i$ hired [the secretary]$_j$ because [she]$_j$ was highly recommended.
Anti: [The physician]$_i$ hired [the secretary]$_j$ because [he]$_j$ was highly recommended.

The WinoBias dataset

- In Type 2, the pronoun can syntactically **only** refer to one of the entities (otherwise would need reflexive).

Pro: [The physician]$_i$ called [the secretary]$_j$ and told [her]$_j$ to cancel the appointment.

Anti: [The physician]$_i$ called [the secretary]$_j$ and told [him]$_j$ to cancel the appointment.

- In Type 1, both possibilities are syntactically allowed; only the semantics constrains the resolution.

Pro: [The physician]$_i$ hired [the secretary]$_j$ because [he]$_i$ was overwhelmed with clients.

Anti: [The physician]$_i$ hired [the secretary]$_j$ because [she]$_i$ was overwhelmed with clients.

- So, if systems learn/use syntactic info as well as semantics, then Type 2 should be easier and less susceptible to bias.

Constructing the pairs

- Used US Labor statistics to choose 40 occupations ranging from male-dominated to female-dominated.
  - (might not be so in other countries!)
- Constructed 3160 sentences according to templates:
  - **Type 1:** [entity1] [interacts with] [entity2] [conjunction] [pronoun] [circumstances]
  - **Type 2:** [entity1] [interacts with] [entity2] and then [interacts with] [pronoun] [circumstances]

Testing coreference systems

- Three systems are tested on WinoBias:
  - Rule-based (Stanford Deterministic Coreference System, 2010)
  - Feature-based Log-linear (Berkeley Coreference Resolution System, 2013)
  - Neural (UW End-to-end Neural Coreference Resolution System, 2017)
- Rule-based doesn't train; others are trained on OntoNotes 5.0 corpus.

Out-of-the-box results

- Yes, systems are biased… (numbers are F1 scores)

Method	T1-pro	T1-anti	T1-Diff	T2-pro	T2-anti	T2-Diff
Neural	76.0	49.4	26.6	88.7	82.0	13.5
Feature	66.7	56.0	10.6	73.0	65.2	15.7
Rule	76.7	37.5	39.2	50.5	39.9	21.3

- All systems do much better on Pro than Anti (large Diff).
- For Neural and Rule, Diff is much bigger for Type 1 (T1) than Type 2 (T2), as expected.
- For Feature, Diff is larger for T2: unexpected, and paper does not comment on possible reasons!

Likely reasons

- Biases in immediate training data: Like many corpora, OntoNotes itself is biased.
  - 80% of mentions headed by gendered pronoun are male.
  - Male gendered mentions are >2x as likely to contain a job title as female mentions.
  - OntoNotes contains various genres; same trends hold for all of them.
- Biases in other resources used:
  - For example, the pre-trained word embeddings used by some of the systems.

Augmenting data by gender-swapping

To address the bias in OntoNotes, Zhao et al. create additional training data by gender-swapping the original data, as follows.

1. Anonymize named entities

French President **Emmanuel Macron** appeared today ... Mr. **Macron** has been criticized for his ... He announced his ...

French President **E1 E2** appeared today ... Mr. **E2** has been criticized for his ... He announced his ...

Augmenting data by gender-swapping

2. Create a dictionary of gendered terms and their gender-swapped versions, e.g.

she ↔ he, her ↔ him, Mrs. ↔ Mr., mother ↔ father

3. Replace gendered terms with their gender-swapped versions:

French President E1 E2 appeared today ... **Mr.** E2 has been criticized for **his** ... He announced **his** ...

French President E1 E2 appeared today ... **Mrs.** E2 has been criticized for **his** ... **She** announced **her** ...

Additional methods

- Reduce gender bias in pre-trained word embeddings using methods from Bolukbasi et al. (2016)
- Gender balance frequencies in other word lists obtained from external resources.

Final results

- After applying all de-biasing methods:

Method	T1-pro	T1-anti	T1-Diff	T2-pro	T2-anti	T2-Diff
Neural	63.9	62.8	1.1	81.3	83.4	-2.1
Feature	62.3	60.4	1.9	71.1	68.6	2.5

- The Diffs are all much smaller (in fact, mostly no longer statistically significant).

Interim discussion

- Is what I've said so far enough to conclude that
  - Co-ref systems are likely to produce incorrect results in anti-stereotypical sentences?
  - The proposed de-biasing methods remove gender bias from co-ref systems?
  - We should use these methods for new systems?
- If not, what further evidence would help answer these questions? What other questions should we be asking?

Results on OntoNotes

- Are co-ref systems likely to produce incorrect results in anti-stereotypical sentences?
  - On challenge data set, yes!
  - What about on more typical text?
- To see, evaluate on anonymized OntoNotes dev set, original vs gender-swapped:

Model	Original	Swapped
Neural	66.4	65.9
Feature	61.3	60.3

  - Suggests that on easy (in-domain) cases, gender bias isn't likely to cause many errors (good news!)
  - But real out-of-domain cases are probably somewhere between OntoNotes and WinoBias, and WinoBias shows that hard cases do cause errors!

Did we remove all bias from the systems?

- Maybe. But we can't conclude gender bias is gone (or negligible) even if Diff is close to zero.
  - WinoBias only tests a particular type of sentence.
  - Gender bias might affect other types of sentences that weren't measured.
- An example of a one-way test (similar to statistical hypothesis tests and many other scientific experiments):
  - Can provide evidence that bias does exist
  - Lack of evidence does not mean no bias exists

Should we always use these methods?

- We probably want to know whether doing so negatively impacts results on more typical cases, and if so how much.
- In this case, only slightly.
  Results on OntoNotes dev:

Model	Out-of-box	De-biased
Neural	67.7	66.3
Feature	61.7	61.0

- A bigger impact would cause a bigger ethical dilemma, and motivate developing a better method.
- We also don't know whether/how well this method applies to other datasets/languages/etc.

Where does it leave us?

- So, still leaves open questions, but a good start towards measuring and reducing gender bias in coreference systems.
- Algorithmic bias (gender and otherwise) is a growing area of research in NLP.
- For other tasks, see recent review of work on mitigating gender bias in NLP by Sun et al. (2019).

Questions for review

- What are some examples of gender bias in NLP and what consequences might these have?
- What is a challenge dataset and how are these used to target specific problems like gender bias?
- For one specific example (gender bias in coreference),
  - How can we systematically measure (aspects of) this bias?
  - What are some sources of the bias?
  - What can be done to develop systems that are less biased?

References

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. 'Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings'. In *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 4349–4357.

Kate Crawford. 2017. The Trouble With Bias. Keynote at Neural Information Processing Systems (NIPS'17).

Lu, Kaiji, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. 'Gender Bias in Neural Natural Language Processing'. arXiv:1807.11714

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. 'Linguistic Regularities in Continuous Space Word Representations'. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751. Atlanta, Georgia: Association for Computational Linguistics.

Stanovsky, Gabriel, Noah A. Smith, and Luke Zettlemoyer. 2019. 'Evaluating Gender Bias in Machine Translation'. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1679–84.

Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. 'Mitigating Gender Bias in Natural Language Processing: Literature Review'. In , 1630–40.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. 'Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods'. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20.