Bias in NLP Systems

COMP-550

Nov 30, 2017

Outline

A4 reading discussions

Bias in NLP systems

Recap

A4 Reading Discussion

Any clarification questions?

How does this method relate to the work we discussed in class?

What are the strengths of the approach? Limitations?

Is it a good idea to replace parts of the model with a neural network?

- If so, which parts?

NLP in the Real World

NLP and AI systems are increasingly used to automate fact finding and decision making

- Information retrieval
- Image captioning
- Automated essay grading
- School admissions decisions
- Resume and CV filtering
- Loan and insurance approval

Want to make sure process and decisions are fair and unbiased!

Technological Fairness?

Hope:

- Use objective measures and statistical techniques to produce a fairer system, free of human biases

Reality:

- Machine learning systems can learn the biases that are inherent in the data
- ***Even worse***: the learned methods can produce results that are **more biased** than the training data!
- How can this be?

Bias in Word Embedding Models

word2vec exhibits bias!

This is okay:

man – woman ≈ king – queen

But this is NOT, and also found by word2vec!

man – woman ≈ computer programmer – homemaker

(Bolukbasi et al., 2016)

Most Gender-Biased Occupations

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Figure 1: The most extreme occupations as projected on to the *she*−*he* gender direction on g2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded.

(Bolukbasi et al., 2016)

Implications of Word Association Bias

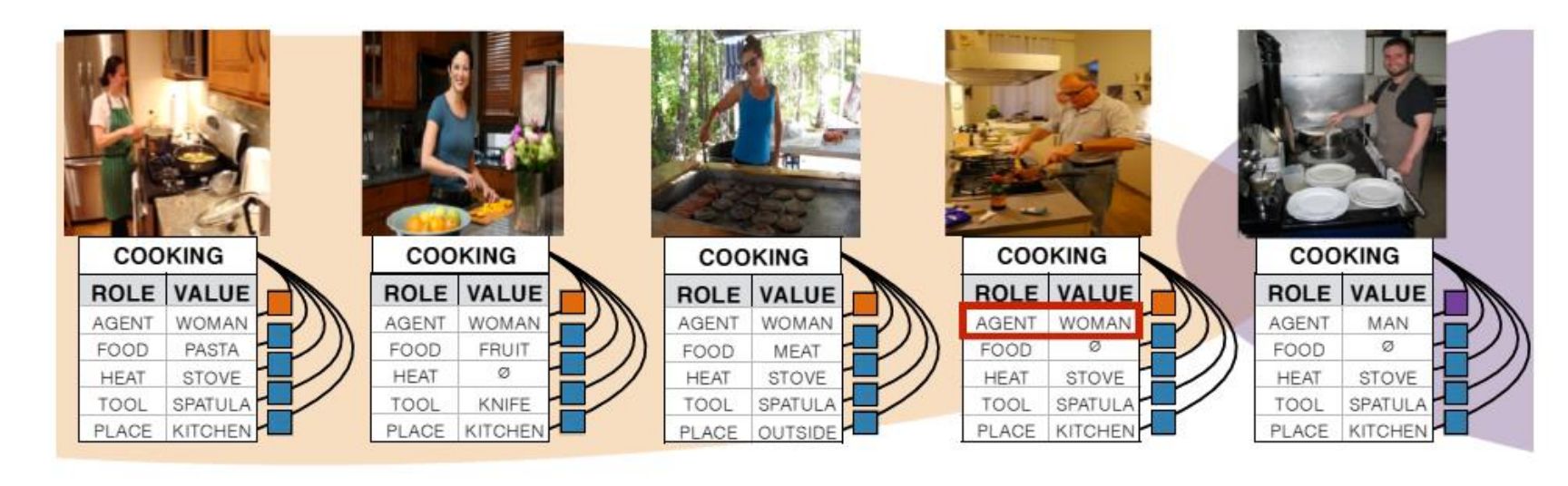
Above results due to counting of word associations

Maybe this just reflects the bias in the underlying distribution of real life – why is that so bad?

Scenario: information retrieval; search result

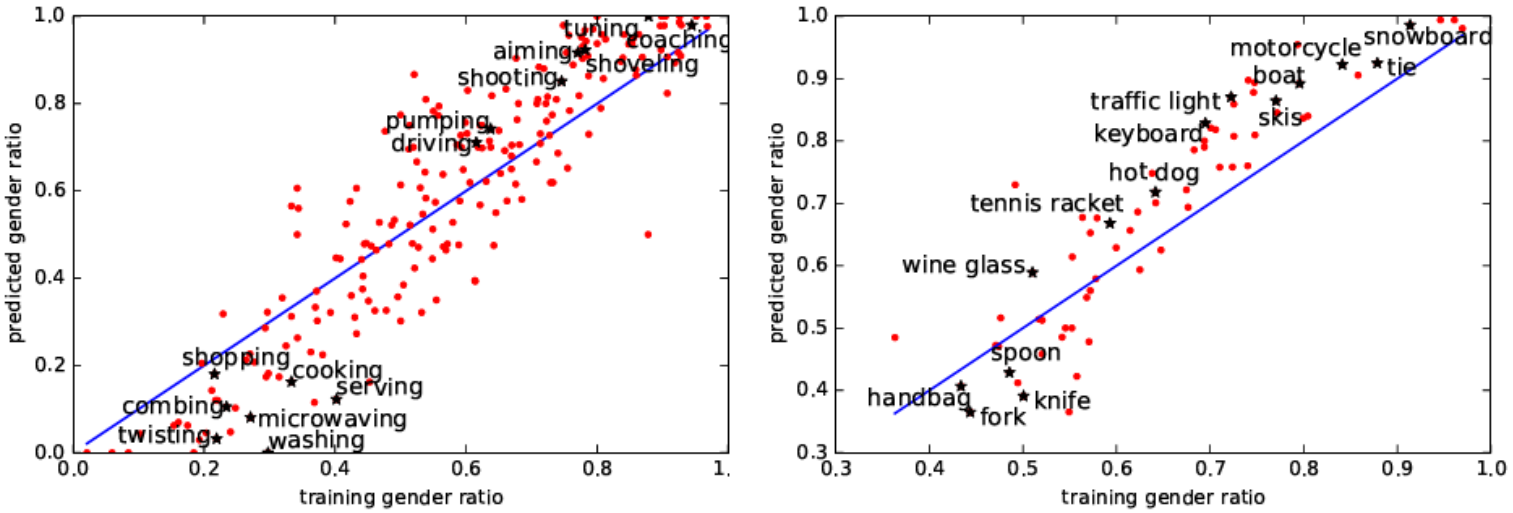
- Must produce a ranking of, say, people's home pages to show in a search query.
- e.g. "*cmu computer science phd student*"
- Given two otherwise identical webpages, an algorithm may pick a website with a man's name (e.g., *John*) over one with a woman's name (e.g., *Mary*), because the former is more distributionally similar to computer science!

Visual Semantic Role Labelling



imSitu data set (Yatskar et al., 2016)

Bias Amplification in Trained Models



(a) Bias analysis on imSitu vSRL

(b) Bias analysis on MS-COCO MLC

Figure 2: Gender bias analysis of imSitu vSRL and MS-COCO MLC. (a) gender bias of verbs toward man in the training set versus bias on a predicted development set. (b) gender bias of nouns toward man in the training set versus bias on the predicted development set. Values near zero indicate bias toward woman while values near 0.5 indicate unbiased variables. Across both dataset, there is significant bias toward males, and significant bias amplification after training on biased training data.

Result from (Zhao et al., 2017)

Why does Bias Amplification Occur?

Training data exhibits some bias

An automatic system is asked to produce a decision under uncertainty

- Ranking websites
- Labelling image as involving male or female participant

With standard loss/evaluation procedures, rational to favour more frequent class, if other information does not disambiguate

Debiasing Algorithms

General technique:

1. Identify axis or axes of bias (e.g., gender, race, religion, etc.)
2. Modify our learning or inference by adding constraints, such that the biased outcomes (as previously identified) are disfavoured

Let's consider the method of Zhao et al., (2017)

Debiasing Activity Recognition

Original inference problem:

$$\text{argmax}_{y \in Y} \, f_\theta(y, i)$$

- i.e., make the decision y (e.g., $y = \{\text{woman}, \text{meat}, \text{stove}, \ldots\}$) that maximizes the score on test instance i

Idea: for each activity v^* to debias, add a constraint:

$$b^* - \gamma \leq \frac{\sum_i y^i_{v=v^*, r \in M}}{\sum_i y^i_{v=v^*, r \in W} + \sum_i y^i_{v=v^*, r \in M}} \leq b^* + \gamma \quad (2)$$

where $b^* \equiv b^*(v^*, man)$ is the desired gender ratio of an activity v^*, γ is a user-specified margin. M and W are a set of semantic role-values representing the agent as a `man` or a `woman`, respectively.

New Optimization Problem

$$\max_{\{y^i\}\in\{Y^i\}} \quad \sum_i f_\theta(y^i, i), \qquad \text{s.t.} \quad A\sum_i y^i - b \leq 0, \tag{3}$$

- where $\{Y^i\}$ represents the space of all possible label assignments to all test instances
- constraints are taken from equation (2) for each activity

This is expensive to solve exactly; use an approximate method based on Lagrange multipliers

Performance

Method	Viol.	Amp. bias	Perf. (%)
vSRL: Development Set			
CRF	154	0.050	24.07
CRF + RBA	107	0.024	23.97
vSRL: Test Set			
CRF	149	0.042	24.14
CRF + RBA	102	0.025	24.01
MLC: Development Set			
CRF	40	0.032	45.27
CRF + RBA	24	0.022	45.19
MLC: Test Set			
CRF	38	0.040	45.40
CRF + RBA	16	0.021	45.38

Table 2: Number of violated constraints, mean amplified bias, and test performance before and after calibration using RBA. The test performances of vSRL and MLC are measured by top-1 semantic role accuracy and top-1 mean average precision, respectively.

Reduced bias amplification without much loss in classification performance!

Summary of Current Work

Bias is a problem in NLP systems

Naïve methods can exacerbate problem

Possible to reduce effect of biases without sacrificing task performance

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NIPS 2016.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. EMNLP 2017.

Recap of Course

What have we done in COMP-550?

Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Domains of natural language

- Acoustic signals, phonemes, words, syntax, semantics, …
- Speech vs. text
- **Natural language understanding** (or **comprehension**) vs. **natural language generation** (or **production**)

Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Goals

- Language technology applications
- Scientific understanding of how language works

Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Methodology and techniques

- Gathering data: language resources
- Evaluation
- Statistical methods and machine learning
- Rule-based methods

Current Trends and Challenges

Speculations about the future of NLP

Better Use of More Data

Large amounts of data now available

- Unlabelled
- Noisy
- May not be directly relevant to your specific problem

How do we make better use of it?

- Unsupervised or lightly supervised methods
- Prediction models that can make use of data to learn what features are important (neural networks)
- Incorporate linguistic insights with large-scale data processing

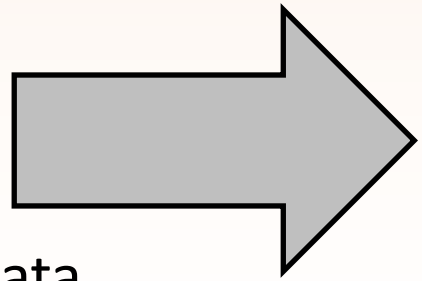
Using More Sources of Knowledge

Old set up:

Annotated data set → (Feature extraction + Simple supervised learning) → Model predictions

Better model?

Background text
General knowledge bases
Domain-specific constraints
Directly relevant annotated data



Model predictions

Away From Discreteness

Discreteness is sometimes convenient assumption, but also a problem

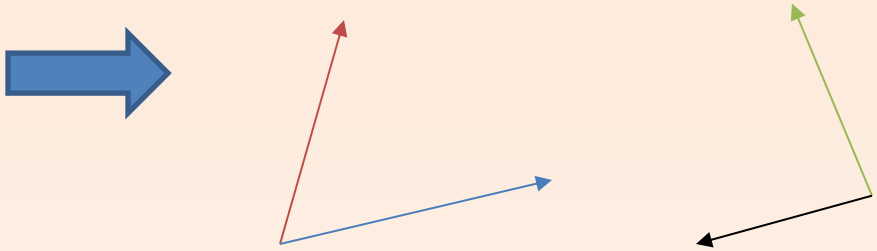
- Words, phrases, sentences and labels for them
- Symbolic representations of semantics
- Motivated a lot of work in regularization and smoothing

Representation learning

- Learn continuous-valued representations using co-occurrence statistics, or some other objective function
- e.g., vector-space semantics

Continuous-Valued Representations

cat, *linguistics*, NP, VP



Advantages:

- Implicitly deal with smoothness, soft boundaries
- Incorporate many sources of information in training vectors

Challenges:

- What should a good continuous representation look like?
- Evaluation is often still in terms of a discrete set of labels

Broadening Horizons

We are getting better at solving specific problems on specific benchmark data sets.

- e.g., On WSJ corpus, POS tagging performance of >97% matches human-level performance.

Much more difficult and interesting:

- Working across multiple kinds of text and data sets
- Integrating disparate theories, domains, and tasks

Connections to Other Fields

Cognitive science and psycholinguistics

- e.g., model L1 and L2 acquisition; other human behaviour based on computational models

Human computer interaction and information visualization

- That's nice that you have a tagger/parser/summarizer/ASR system/NLG module. Now, what do you do with it?
- **Multi-modal** systems and visualizations

That's It!

Good luck on your projects and finals!