Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

Md Musa
Leibniz University of Hannover

Agenda

- Introduction
- Background
  - Word2Vec algorithm
  - Bias in the data generated from Word2Vec
  - Approach to debias the algorithm
- Experiments Overview
  - Dataset
  - Setup
  - Findings
- Pros and Cons
- Ideas for Future Work

Introduction

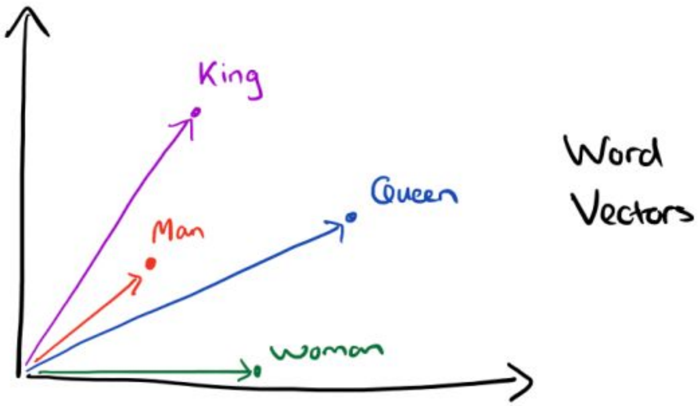
- Paper:

  Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings
- Authors:
  - Tolga Bolukbasi, Venkatesh Saligrama - Boston University
  - Kai-Wei Chang, Adam Kalai, James Zou - Microsoft Research
- Mentors:
  - Prof. Dr. Techn Wolfgang Nejdl - Leibniz University of Hannover

Word Embeddings

- Word Embeddings ?
- Word Embeddings used Word2Vec to represent text data as vector
  - Word2Vec extracts the dimension of the words
  - Finds semantic meaning and word properties of words

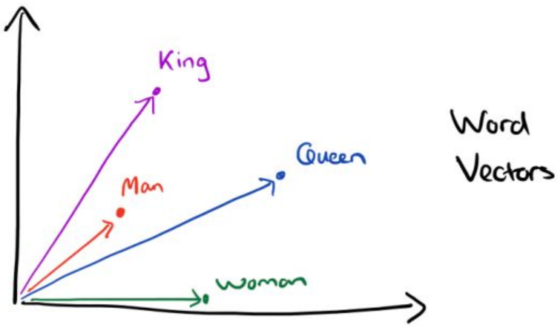
Example :



How Word2Vec Helps

- Similar words - Enhance Search Engine
- Sentiment Analysis - Few dimension can indicate whether the sentiment is good or bad
- Machine Translation - Similar words will be treated same
- Clustering - Words from same field (sports, violence, profession ...etc.) will be assigned to the same cluster

Bias in Word Embeddings



- For example :
  - Man is to King, as woman is to ___
    - **vector('King') - vector('Man') + vector('Woman')** is close to **vector('Queen')**
  - Man is to doctor, as woman is to ___
    - **vector('Doctor') - vector('Man') + vector('Woman')** is close to **vector('Nurse')**

Direct Gender Bias

- Extreme “she” occupations
  - homemaker
  - nurse
  - receptionist
- Extreme “he” occupations
  - boss
  - philosopher
  - architect

Indirect Gender Bias

Relative geometry between gender neutral words themselves

occupations related to "softball"

- pitcher
- bookkeeper
- receptionist
- registered nurse
- waitress

occupations related to "football"

- footballer
- businessman
- pundit
- maestro
- cleric

Approach to debias Word Embeddings

- **Identify gender subspace (Step 1):**
  - in their data, the authors found a single direction that largely captures gender

Approach to debias Word Embeddings

- **Hard Debias (Step 2a):**
  - Hard debiasing algorithm **removes** the **gender pair** associations for **gender neutral** words
    - neutralize: ensure that gender neutral words are zero in the gender subspace
      - “nurse” is moved to be equally male and female in the gender subspace direction
    - equalize: equalize words outside the subspace equality sets
      - Example : “babysit” equally distant to grandmother and grandfather, but closer to {grandmother, grandfather} than to {guy,gal}

Approach to debias Word Embeddings

- Soft Debias (Step 2b):
    - similar to hard debias, but it controls the trade-off between debias and original performance

Dataset

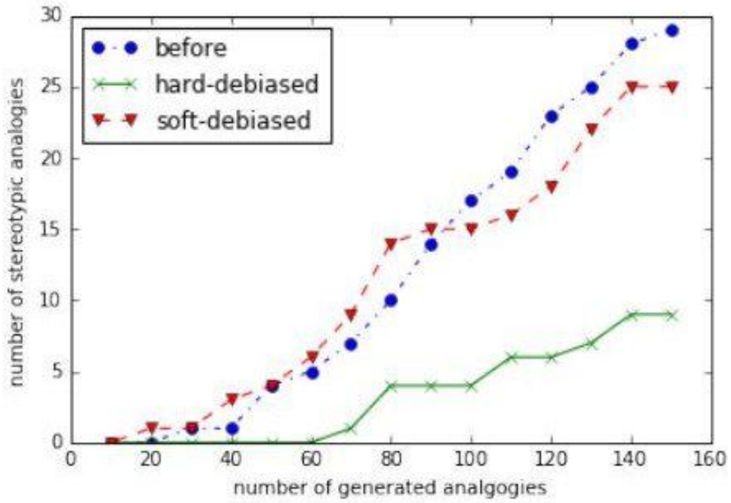
- W2vNews Dataset
  - 300-dimension English word vectors generated from Google News
  - 3 million words and phrases
- limited to 26,377 lowercase words

Setup (Questionnaires)

- Two set of questionnaires for crowd-workers
  - Questionnaire for generating gender stereotypical words
    - definitionally associated with males: dude, menswear
    - stereotypically associated with males: football, cocky
  - Questionnaire for stereotype analogies
    - stereotype analogy: doctor:man :: nurse:woman
    - appropriate analogy: king:man :: queen:woman

Number of Stereotypical Analogies

	Crowd-Worker Votes	
Analogy to she:he	Appropriate	Biased
midwife:doctor	1	10
sewing:carpentry	2	9
registered_nurse:physician	1	9
women:men	10	1
headscarf:turban	6	1



Pros & Cons

- Pros:
  - Effectively debiasing gender bias without decrease in performance
  - Lot of examples of gender stereotypes, analogies, the results of the algorithm and the crowd-worker annotations
- Cons:
  - Soft-debiasing rather unclear and results not very promising
  - Approach maybe not completely generalizable

Future Work

- Find and remove racial, ethnic and cultural bias
- Remove bias in language translation

Example :

Software Developer : Softwareentwickler. Here it gives a male translation, but it could be girl also (“Softwareentwicklerin”)

References

[1] Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." *Advances in Neural Information Processing Systems*. 2016.

[2] Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R., 2012, January. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226). ACM.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

Thanks for listening