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 ___
    - $\text{vector('King')} - \text{vector('Man')} + \text{vector('Woman')}$ is close to $\text{vector('Queen')}$
  - Man is to doctor, as woman is to ___
    - $\text{vector('Doctor')} - \text{vector('Man')} + \text{vector('Woman')}$ is close to $\text{vector('Nurse')}$. 
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
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

<table>
<thead>
<tr>
<th>Analogy to she:he</th>
<th>Crowd-Worker Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Appropriate</td>
</tr>
<tr>
<td>midwife:doctor</td>
<td>1</td>
</tr>
<tr>
<td>sewing:carpentry</td>
<td>2</td>
</tr>
<tr>
<td>registered_nurse:physician</td>
<td>1</td>
</tr>
<tr>
<td>women:men</td>
<td>10</td>
</tr>
<tr>
<td>headscarf: turban</td>
<td>6</td>
</tr>
</tbody>
</table>
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


Thanks for listening