CS378: Natural Language Processing
Lecture 8: Bias in Embeddings, Multilingual Embeddings

Greg Durrett

The University of Texas at Austin
Announcements

- Assignment 2 due in one week
- Bias in embeddings response due next Tuesday (submit on Canvas)
- Survey on Instapoll
Recap
Playing around with embeddings

- See Instapoll
Using Word Embeddings

- Approach 1: learn embeddings as parameters from your data
  - Often works pretty well
- Approach 2: initialize using GloVe, keep fixed
  - Faster because no need to update these parameters
- Approach 3: initialize using GloVe, fine-tune
  - Works best for some tasks
Beyond Word Embeddings
fastText: Sub-word Embeddings

- Same as SGNS, but break words down into n-grams with n = 3 to 6
  
  where:
  
  3-grams: <wh, whe, her, ere, re>
  
  4-grams: <whe, wher, here, ere>,
  
  5-grams: <wher, where, here>,
  
  6-grams: <where, where>

- Replace $w \cdot c$ in skip-gram computation with $\left( \sum_{g \in \text{ngrams}} w_g \cdot c \right)$

Bojanowski et al. (2017)
What if we want embedding representations for whole sentences?

- *Skip-thought* vectors (Kiros et al., 2015), similar to skip-gram generalized to a sentence level (more later)

- Is there a way we can compose vectors to make sentence representations? Summing?

- Will return to this in a few weeks as we move on to syntax and semantics
- How to handle different word senses? One vector for *bats*

- ELMo: train a neural language model to predict the next word given previous words in the sentence, use its internal representations as word vectors

- *Context-sensitive* word embeddings: depend on rest of the sentence

- *Huge* improvements across nearly all NLP tasks over GloVe

Peters et al. (2018)
Bias in Word Embeddings
What can go wrong with word embeddings?

- What’s wrong with learning a word’s “meaning” from its usage? Maybe some words are used in ways we don’t want to replicate?

- What data are we learning from?

- What are we going to learn from this data?
Bias Exercise

Consider learning word embeddings from a **corpus of news articles**.

1. Think about a similarity association a model might learn that you believe constitutes **bias**. For this association, list (a) what the word pair is; (b) why you think this is present in the data (e.g., give an example of how it could appear in a news story)

2. Embeddings are often used at the input layer of a neural network. Can you think of a task for which this biased association might lead to bias in the system?

Now consider learning word embeddings from a **corpus of social media data comments (think about reddit + Twitter)**.

3. Do you think you’re likely to see the bad association from above? Why or why not?

4. Come up with a new biased similarity association; list (a) what the word pair is; (b) why you think this is present in social media data
What do we mean by bias?

- Compare distance (using cosine similarity) of many occupations to the vectors for *he* and *she*

\[
\cos(u, v) = \frac{u \cdot v}{\|u\|\|v\|}
\]

- These regularities are not restricted to gendered pronouns. *receptionist* is closer to *softball* than *football*

- This work focuses on binary gender stereotypes, but it can be extended

Bolukbasi et al. (2016)
What do we mean by bias?

**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. fighter pilot  
12. boss

<table>
<thead>
<tr>
<th>Racial Analogies</th>
</tr>
</thead>
<tbody>
<tr>
<td>black → homeless</td>
</tr>
<tr>
<td>caucasian → servicemen</td>
</tr>
<tr>
<td>caucasian → hillbilly</td>
</tr>
<tr>
<td>asian → suburban</td>
</tr>
<tr>
<td>asian → laborer</td>
</tr>
<tr>
<td>black → landowner</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Religious Analogies</th>
</tr>
</thead>
<tbody>
<tr>
<td>jew → greedy</td>
</tr>
<tr>
<td>muslim → powerless</td>
</tr>
<tr>
<td>christian → familial</td>
</tr>
<tr>
<td>muslim → warzone</td>
</tr>
<tr>
<td>muslim → uneducated</td>
</tr>
<tr>
<td>christian → intellectually</td>
</tr>
</tbody>
</table>

- Nearest neighbor of \((b - a + c)\)

Bolukbasi et al. (2016)  
Manzini et al. (2019)
Debiasing

- Identify gender subspace with gendered words (avg “male” - avg “female” word)
- Project words onto this subspace
- Subtract those projections from the original word

Bolukbasi et al. (2016)
Hardness of Debiasing

- Not that effective...and the male and female words are still clustered together
- Bias pervades the word embedding space and isn’t just a local property of a few words

Gonen and Goldberg (2019)
“Toxic degeneration”: neural models that generate toxic stuff

- System trained on a big chunk of the Internet: conditioning on “SJW”, “black” gives the system a chance of recalling bad stuff from its training data

Example prompt and toxicity:

- Prompt: I'm sick of all the politically correct stuff the media are telling you: you are sick of the prejudiced white trash [Trump supporters]....
- Toxicity: Very Toxic

⚠️ Toxic generations may be triggering.
Multilingual Word Embeddings
Recall: Training Embeddings

- Input: a large corpus of text in some language (English)
- Output: embedding for each word
- What if we have multiple corpora of text in different languages?
- Learning embeddings on each language individually: these embeddings aren’t expected to have any relation
Multilingual Embeddings

- Input: corpora in many languages. Output: embeddings where similar words *in different languages* have similar embeddings

  I have an apple
  47 24 18 427

  J’ ai des oranges
  47 24 89 1981

- multiCluster: use bilingual dictionaries to form clusters of words that are translations of one another, replace corpora with cluster IDs, train “monolingual” embeddings over all these corpora

- Works okay but not all that well

Ammar et al. (2016)
Aligning existing embeddings

- What if you already have embeddings in two languages and you just want to align them?

- Given: dictionary of pairs \((x_i, z_i)\), where \(x\) are word embeddings in a source lang (English) and \(z\) are word embeddings in a target lang (French)

- Learn a matrix \(W\) to minimize the following:

\[
\min_W \sum_{i=1}^{n} \| W x_i - z_i \|^2
\]

(Looks like a loss function! Can learn with SGD on the pairs)

Mikolov et al. (2013)
Rotation learns to align these word embedding spaces! Does this cartoon match reality?

Conneau et al. (2017)
Aligning existing embeddings

Table 2: Accuracy of the word translation methods using the WMT11 datasets. The Edit Distance uses morphological structure of words to find the translation. The Word Co-occurrence technique based on counts uses similarity of contexts in which words appear, which is related to our proposed technique that uses continuous representations of words and a Translation Matrix between two languages.

<table>
<thead>
<tr>
<th>Translation</th>
<th>Edit Distance</th>
<th>Word Co-occurrence</th>
<th>Translation Matrix</th>
<th>ED + TM</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P@1</td>
<td>P@5</td>
<td>P@1</td>
<td>P@5</td>
<td>P@1</td>
</tr>
<tr>
<td>En → Sp</td>
<td>13%</td>
<td>24%</td>
<td>19%</td>
<td>30%</td>
<td>33%</td>
</tr>
<tr>
<td>Sp → En</td>
<td>18%</td>
<td>27%</td>
<td>20%</td>
<td>30%</td>
<td>35%</td>
</tr>
<tr>
<td>En → Cz</td>
<td>5%</td>
<td>9%</td>
<td>9%</td>
<td>17%</td>
<td>27%</td>
</tr>
<tr>
<td>Cz → En</td>
<td>7%</td>
<td>11%</td>
<td>11%</td>
<td>20%</td>
<td>23%</td>
</tr>
</tbody>
</table>

Mikolov et al. (2013)