Ling/CSE 472: Introduction to Computational Linguistics

5/30/17

Ethics & NLP 2

Overview

- Debiasing word embeddings
- Reading questions
- Course learning goals: Reflections
- Course evaluations

Understanding research papers

- What are they doing?
- Why are they doing it?
- How are they evaluating it?
- How did they actually do it?

Bolukbasi et al 2016: Debiasing Word Embeddings

- What are they doing?
  - (1) Measuring bias in trained word embeddings
  - (2) Removing bias from trained word embeddings

What are word embeddings?

- A representation of lexical semantics: “You know a word by the company it keeps” (Firth, J. R. 1957:11)

- Vector-space, continuous representations of words

- In vogue right now, but not really new: previous approaches included Latent Semantic Analysis and Latent Dirchlet Analysis.

- Word embeddings are trained with neural nets and are much more efficient to train than LSA/LDA (Mikolov et al 2013)

- Basic idea (per Mikolov et al): better language models through representations that capture degrees of similarity between words

What are word embeddings?

- N-dimensional vectors where each dimension represents co-occurrence with some other word in the vocabulary

- Train an NN to do some other task (e.g. predict words in a window)

- Extract the learned representations of the words -> use in other tasks

Bolukbasi et al 2016: Debiasing Word Embeddings

- What are they doing?
  - (1) Measuring bias in trained word embeddings
  - (2) Removing bias from trained word embeddings

Measuring bias in word embeddings: What & why?

- Words are similar if their vectors are close to each other in the high-dimensional vector space

- What is captured about words by that similarity?

- What do the authors mean by bias here?

- Why measure bias?

Measuring bias in word embeddings: Evaluation

- Two tasks for crowd workers:

- (1) of occupation words judged by the system to be highly ‘she’ correlated or highly ‘he’ correlated, ask: stereotypically female, stereotypically male, neutral?

- (2) of analogies proposed by the system, ask: gender stereotype or gender appropriate?

Measuring bias in word embeddings: How

- Find list of occupation words
- Rank according to how close the vector for each is to the vectors for *she* and *he*
- Generate analogies for *she : he :: x : y* for all pairs (x,y) with a high enough similarity between them

Removing bias from word embeddings

- What?
- Why?
- How evaluated?
- How?

Reading questions

- How were the word vectors identified? The paper says that it was trained on a corpus, but what information from the corpus was used? Did it simply group words that tended to occur close to each other?

- If word embedding is based on co-occurrence, then what if some two words co-occur in a corpus, but what is being discussed semantically is how they are opposites/not related.

Reading questions

- If the NLP tool is trained on data-set, isn't this gender bias merely reflecting that the language is biased?

- Wouldn't it be more constructive if we can recognize the gender disparity in our model, but still use them to build our models as one of many parameters, than simply neutralizing those elements to see the model behave as desired? The reason why a Machine Learning Model based on the facts can amplify biases seems more about the people that use the model in an undesirable way.

Reading questions

- Also, this whole issue strikes me as something of a band-aid fix for deeper issues; I'm fairly certain that most people's first reaction to the "male:doctor::female:X" analogy would also be "nurse" (even if they don't say that aloud)--this issue isn't limited to machines. In the first place, it's practically a leading question: since there's no intrinsic relationship between "doctor" and "male" other than the fact that we tend to think that doctors are usually male, the best answer we can come up with is "nurse," since we also tend to think that nurses are usually female, and "nurse" is clearly related to "doctor." I'm not sure what the "correct" completion of this analogy would be (I don't think "physician" is quite right either); the best answer might just be not to respond.

- Basically, these biases must also exist in the sources the data were collected from; shouldn't we focus on those instead? Also, who's to decide which biases are good or bad?

Reading questions

- It seems like this algorithm requires one to identify a subspace of gender biased words, and this is done in part through a hand picked / moderated list. Through the technique used in this paper, could one identify other biased subspaces?

- Is it possible to select which words individually should be hard debiased vs. soft debiased? If so, how would one go about making the distinction and how could we be sure the reasoning was comprehensive?

Reading questions

- I'm also interested in how much the de-biasing impacts the performance of one of these word embeddings (outside of evaluating whether or not it produces biased analogies). I feel like getting rid of useful information is going to lead to worse performance.

Course learning goals

- What have you learned about:
    - Be familiar with computational linguistic tools and resources, and how they are applied in research in both computational linguistics and other subfields
    - Have a rough sense of the state of the art in this subfield
    - Be able to conceptualize problems from the perspective of computational linguistics

Overview

- Debiasing word embeddings
- Reading questions
- Course learning goals: Reflections
- Course evaluations