Algorithmic Bias in Natural Language Processing

CPSC 15/PHIL 7 Ethics and Technology
Prof. Ameet Soni, Computer Science
Prof. Krista Thomason, Philosophy
Background - NLP

Natural Language Processing (NLP)
• Goal: help computers understand human language
• Artificial Intelligence + Linguistics

General tasks:
• Generation language/communicate with humans
• Analyze/understand humans
Example NLP tasks

- Translation between two languages
- Extract information from text ("Find all relevant Title VII case law")
- Sentiment analysis ("Is this review positive or negative?")
- Personal Assistant
- Search
- Summarization
- Etc.
Difficulties

• Goal: map meaning to words

• Problem: this is very hard; language is
  • Ambiguous - same word can mean different concepts
  • Rich - the same concept can be said with many words
  • “Meaning” is never observed directly

• Representation: how should computers encode words?
“John saw the woman with the telescope wrapped in paper”

- What’s wrapped in paper?
- Who has the telescope?
Word Embeddings

• Words are not discrete concepts
e.g., search “Seattle hotel” vs “Seattle motel” should be similar

• Idea: words with similar meaning occur in similar contexts
  • “hotel” and “motel” are used similarly in sentences

• Word embeddings: represented words by what they co-occur with
  • “book”, “room”, “rate” are commonly used with “hotel” and “motel”
Fast Forward

- Add theory, math, advances in computing, lots of data...
Embeddings Visualized
Learning Semantics

https://www.tensorflow.org/tutorials/representation/word2vec
Learning Relationships

Word Embedding Today

• Many approaches
• Many success stories

• Data driven - learned through provided text e.g., Wikipedia
• Dense and distributed representations
Lab Practicum Part 1: learning embeddings

• learn word embeddings
  • GloVe algorithm https://nlp.stanford.edu/projects/glove/
  • Training corpus

• Goal: validate the usefulness of word embeddings

• Read documentation and learn how to use findSimilarWords.py
Lab Practicum Part 2: Word Embedding Association Tests

- Caliskan et al., “Semantics derived automatically from language corpora contain human-like biases”
- Modeled after Implicit Association Test
- Read documentation and run weatTest.py
- Complete Lab Practicum Assignment
Gender Bias in Action: Google Translate

- o bir aşçı
- o bir mühendis
- o bir doktor
- o bir hemşire
- o bir temizlikçi
- o bir polis
- o bir asker
- o bir öğretmen