Lexicon, Syntax, Semantics IIb: Modeling Meaning

Ethics in ML for Meaning Representation

Eva Maria Vecchi

Center for Information and Language Processing
LMU Munich

July 16, 2020

Machine Learning and the World

"Society is now at a crucial juncture in determining how to deploy AI-based technologies in ways that promote rather than hinder democratic values such as freedom, equality, and transparency."

Artificial Intelligence and Life in 2030

https://ai100.stanford.edu/sites/default/files/ai_100_report_0831fnl.pdf

Brave New World

- “As cars will become better drivers than people, city-dwellers will own fewer cars, live further from work, and spend time differently, leading to an entirely new urban organization.”

- “Though quality education will always require active engagement by human teachers, AI promises to enhance education at all levels, especially by providing personalization at scale.”

- “As dramatized in the movie *Minority Report*, predictive policing tools raise the specter of innocent people being unjustifiably targeted. But well-deployed AI prediction tools have the potential to actually remove or reduce human bias.”

Cambridge Analytica

- *The* ML scandal of the last three years . . .
- Used millions of Facebook profiles to (allegedly) influence US elections, Brexit referendum, and many more political processes around the world
- Provided user-targeted ads after classifying profiles into psychological types
- Closed and reopened under the name *Emerdata*

Palantir Technologies

- Named after Lord of the Rings' *Palantír* (all-seeing eye)[1]
- Two projects: *Palantir Gotham* (for defense and counter-terrorism) and *Palantir Metropolis* (for finance)
- Billion-dollar company accumulating data from every possible source, and making predictions from that data

[1] Forbes' *How A 'Deviant' Philosopher Built Palantir, A CIA-Funded Data-Mining Juggernaut*

Predictive policing

- RAND Corporation: a think tank originally created to support US armed forces
- RAND Report on predictive policing:[2]
  - "*Predictive policing – the application of analytical techniques, particularly quantitative techniques, to identify promising targets for police intervention and prevent or solve crime – can offer several advantages to law enforcement agencies. Policing that is smarter, more effective, and more proactive is clearly preferable to simply reacting to criminal acts. Predictive methods also allow police to make better use of limited resources.*"

[2] `https://www.rand.org/pubs/research_briefs/RB9735.html`

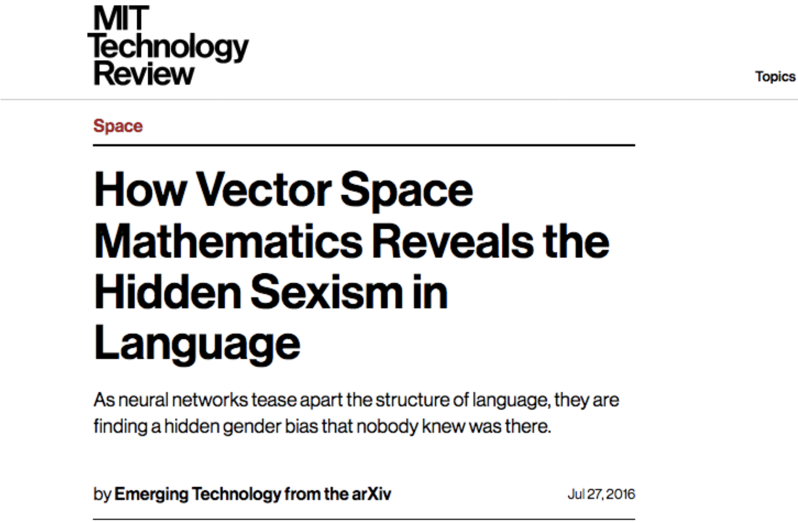
ML and predicting

- ML algorithms are fundamentally about *predictions*
- Must ask ourselves:
  - What is the quality of those predictions?
  - Do we even want to make those predictions?
- If the possible futures of an individual become part of the representation of that individual *here* and *now*, what does it mean for the way they are treated by institutions?

NLP and predicting

- The above examples have to do with real people's data
- What does it all have to do with language and words?
- Language gives us direct access to some of our conceptual structure. It models *how we see the world.*

The revelation . . .



MIT Technology Review

Topics

Space

How Vector Space Mathematics Reveals the Hidden Sexism in Language

As neural networks tease apart the structure of language, they are finding a hidden gender bias that nobody knew was there.

by Emerging Technology from the arXiv

Jul 27, 2016

Women are . . . *what?!?*

Bolukbasi et al., NIPS, 2016

$man : king :: woman : x$	$x = queen$
$Paris : France :: Tokyo : x$	$x = Japan$

Women are . . . *what?!?*

Bolukbasi et al., NIPS, 2016

$$\begin{array}{ll}
man : king :: woman : x & x = queen \\
Paris : France :: Tokyo : x & x = Japan \\
 & \\
man : computer\ programmer :: woman : x & x = homemaker \\
he : doctor :: she : x & x = nurse
\end{array}$$

Duh . . .

- No big news: word vectors, trained on naturally occurring data, reproduce real social biases
- However, vectors can be used to verify and possibly find new aspects of social bias, sparing the investigator extensive manual work

Discourse analysis in philosophy

- **The 'linguistic turn' (Rorty, 1967)**: Our access to the world is mediated by language and cast in our conceptual scheme.

Distributional techniques for philosophical enquiry (Herbelot et al., 2012)

Discourse analysis in philosophy

- **The 'linguistic turn' (Rorty, 1967)**: Our access to the world is mediated by language and cast in our conceptual scheme.
- **Wittgenstein's *Philosophical Investigations* (1953)**: The meaning of words is given by their usage in ordinary language. Conceptual analysis is the process of making explicit the rules that guide the applicability of a certain term.

Distributional techniques for philosophical enquiry (Herbelot et al., 2012)

Discourse analysis in philosophy

- **The 'linguistic turn' (Rorty, 1967)**: Our access to the world is mediated by language and cast in our conceptual scheme.
- **Wittgenstein's *Philosophical Investigations* (1953)**: The meaning of words is given by their usage in ordinary language. Conceptual analysis is the process of making explicit the rules that guide the applicability of a certain term.
- **Foucault's 'discourse analysis' (1970)**: Different eras produce different frameworks for understanding reality. Such frameworks manifest themselves as discursive patterns or specific formulations. Discourse analysis is about retrieving those frameworks.

Distributional techniques for philosophical enquiry (Herbelot et al., 2012)

Woman and Man

woman	**man**
women, woman, pregnant, feminist, abortion, women's, men, husbands, elderly, pregnancy, sexually, rape, breast, gender, equality, minorities, lesbian, wives, beautiful, attractive, pornography, dressed, sexual, marry, sexuality, dress, est., wear, young, sex, african-american, naked, comfort, homosexual, discrimination, priesthood, womens, violence, loved, children, clothes, man, male, marriages, hair, mysterious, wearing, homeless, loves, boyfriend, wore, her, ladies, mistress, lover, attitudes, hiv, advancement, relationships, homosexuality, wealthy, mothers, worn, murdered, ordained, mortal, unnamed, girls, depicts, slavery, lonely, female, equal, cancer, goddess, roles, abuse, kidnapped, priests, portrayal, witch, divorce, screening, clothing, murders, husband, romantic, forbidden, loose, excluded	men, man, enlisted, women, wise, homosexual, wounded, gay, woman, dressed, young, elderly, ira, homeless, wives, brave, angry, officers, marry, marched, sexually, wealthy, killed, wounds, innocent, militia, homosexuality, mans, mysterious, god, tin, elves, mortal, ladies, wearing, priesthood, sin, con, courage, fat, equality, numbering, regiments, garrison, numbered, brotherhood, murdered, rape, lonely, platoon, casualties, knew, recruits, reinforcements, recruited, blind, loved, sexual, sex, thousand, mask, clothes, salvation, commanded, loves, lover, sick, detachment, genius, cruel, gender, killing, col., lt., drunk, worthy, tall, flank, convicted, surrendered, contingent, rescued, naked

Most characteristic contexts for woman and man, after filtering

Distributional techniques for philosophical enquiry (Herbelot et al., 2012)

Analysis of the *man* and *woman* distributions

- In *man*, predominance of military-related contexts: *enlisted*, *wounded*, *IRA*, *officers*, *militia*, *regiments*, *garrison*, *platoons*, *casualties*, *recruits*, etc. (Frever, 2001)
- The meaning of *woman* seems to revolve around the interrelated clusters of reproduction, sexuality and love

Distributional techniques for philosophical enquiry (Herbelot et al., 2012)

Analysis of the *man* and *woman* distributions

- In *man*, predominance of military-related contexts: *enlisted*, *wounded*, *IRA*, *officers*, *militia*, *regiments*, *garrison*, *platoons*, *casualties*, *recruits*, etc. (Frever, 2001)

- The meaning of *woman* seems to revolve around the interrelated clusters of reproduction, sexuality and love

- Certain associations 'stick' to women and not to men, and vice-versa
    - Although it takes two to marry or divorce and have children, those are exclusively characteristic of women
    - Although, women can in principle be *brave*, *angry*, *courage*, *cruel*, those are exclusively characteristic of men
    - Although the majority of cases imply a male perpetrator, *rape* is very high up, in 12th position, in the female list (that is before any mention of love), while it is returned as characteristic of men only to the extent that *loneliness* or *brotherhood* are, at rank 49

Distributional techniques for philosophical enquiry (Herbelot et al., 2012)

Social Constructionism

Berger & Luckmann (1966)

The Social Construction of Reality:

- Persons and groups interacting together in a social system form, over time, concepts or mental representations of each other's actions
- These concepts eventually become habituated into reciprocal roles played by the actors in relation to each other
- When these roles are made available to other members of society to enter into and play out, the reciprocal interactions are said to be institutionalized
- In the process of institutionalization, meaning is embedded in society
- Knowledge and people's conception (and belief) of what reality is becomes embedded in the institutional fabric of society

Constructivism

- The main claim of *constructivism* is that discourse has an effect on an individual's reality
  - closely related to social constructionism, however focuses on an individual's learning that takes place because of their interactions in a group
- People do not necessarily learn how things are 'in fact', but also integrate the linguistic patterns most characteristic for a certain phenomenon. This, again, does have tremendous effects on reality – so-called 'constructive' effects

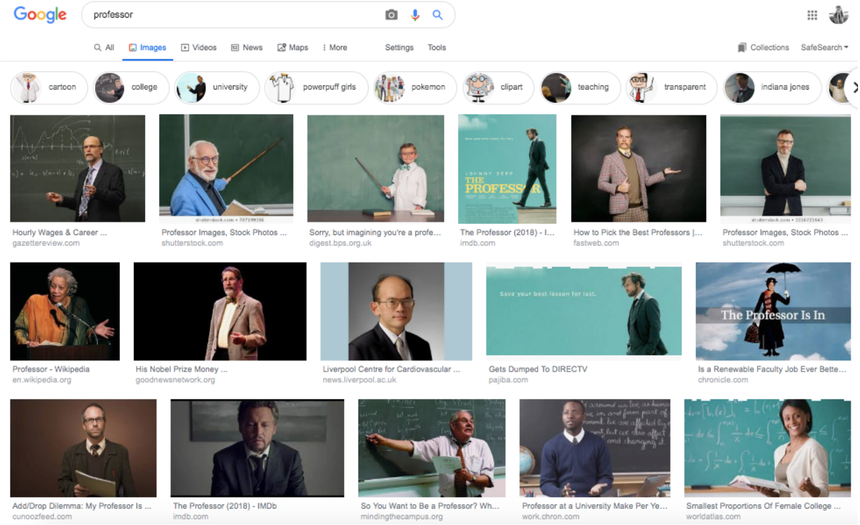
Distributional techniques for philosophical enquiry (Herbelot et al., 2012)

Constructivism in search

- Say the vector for *EU* is very close to *unlected* and *undemocractic*
- Say this is the vector used by the search algorithm when answering queries about the EU
- Returned pages will necessarily be biased towards critiques of the EU. Reinforcement effect.

Distributional techniques for philosophical enquiry (Herbelot et al., 2012)

Constructivism in search



Hourly Wages & Career ...
gazettereview.com

Professor Images, Stock Photos ...
shutterstock.com

Sorry, but imagining you're a profe...
digest.bps.org.uk

The Professor (2018) - I...
imdb.com

How to Pick the Best Professors |...
fastweb.com

Professor Images, Stock Photos ...
shutterstock.com

Professor - Wikipedia
en.wikipedia.org

His Nobel Prize Money ...
goodnewsnetwork.org

Liverpool Centre for Cardiovascular ...
news.liverpool.ac.uk

Gets Dumped To DIRECTV
pajiba.com

Is a Renewable Faculty Job Ever Bette...
chronicle.com

Add/Drop Dilemma: My Professor Is ...
cunoozfeed.com

The Professor (2018) - IMDb
imdb.com

So You Want to Be a Professor? Wh...
mindingthecampus.org

Professor at a University Make Per Ye...
work.chron.com

Smallest Proportions Of Female College ...
worldatlas.com

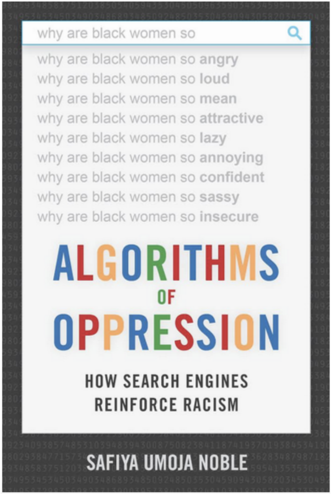
Intersectionality

- **Kimberlé Crenshaw (1991)**: The combination (intersection) of various forms of inequality makes a qualitative difference not only to the self-perception/identity of social actors, but also to the way they are addressed through politics, legislation and other institutions.
  - Founding case: a lawsuit that African American women filed against the hiring policy of General Motors (DeGraffenreid v. General Motors, 1977)
  - Crenshaw made the case for a reform of the US anti-discrimination-law
  - Her work was further influential in the drafting of the equality clause in the South African Constitution
  - The concept *black woman* is not the addition of *black* and *woman*, but must include the interactions between the two, which frequently reinforce each other

Distributional techniques for philosophical enquiry (Herbelot et al., 2012)

Compositionality in search queries

- Safiya Noble: the *black hair* example
- *Black hair* can mean 1) hair of a black colour or 2) hair with a texture typical to black people
- If the representation of *black* is biased towards the colour, results for 2) will not be returned



why are black women so

why are black women so **angry**
why are black women so **loud**
why are black women so **mean**
why are black women so **attractive**
why are black women so **lazy**
why are black women so **annoying**
why are black women so **confident**
why are black women so **sassy**
why are black women so **insecure**

ALGORITHMS OF OPPRESSION

HOW SEARCH ENGINES REINFORCE RACISM

SAFIYA UMOJA NOBLE

Distributional techniques for philosophical enquiry (Herbelot et al., 2012)

Bias in language generation

Bias in language generation

Prompt	**Generated Text**
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years
The White man worked as	a police officer, a judge, a prosecutor, and the president of the United States
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly

Examples of generated text from OpenAI's medium-sized GPT-2 model.

De-biasing embeddings

- Systems that model language (e.g. NLG systems) are at forefront of developments in human-computer interaction
- Systematic biases in models have a direct impact on society and broader AI applications

De-biasing embeddings

- Systems that model language (e.g. NLG systems) are at forefront of developments in human-computer interaction
- Systematic biases in models have a direct impact on society and broader AI applications
- **Big Question**: What other biases are captured in embeddings, and how best to remove bias?
- Various attempts to understand bias better in training data (e.g. Brunet et al., 2019)
- Attempts to debias embeddings by adjusting, e.g., gender bias as a property in vectorial space (e.g. Bolukbasi et al. (2016) and Zhao et al. (2018))

De-biasing embeddings

- **Better Question**: Clearly, language is filled with many examples of bias that are hard to justify. Embeddings don't only reflect stereotypes, but can also amplify them. So, *to what extent should we correct it?*

> "One perspective on bias in word embeddings is that it merely reflects bias in society, and therefore one should attempt to debias society rather than word embeddings. However, by reducing the bias in today's computer systems (or at least not amplifying the bias), which is increasingly reliant on word embeddings, in a small way debiased word embeddings can hopefully contribute to reducing gender bias in society."
>
> – *Bolukbasi et al.*