NLP + Bias Definitions + Bias in DH

David Lassner

(1) Representations for NLP

Multimodal Neurons in Artificial Neural Networks:
https://distill.pub/2021/multimodal-neurons/#typographic-attacks

(1) Representations for NLP: Bag of Words

- Fast and simple
- Deterministic
- Context information is lost

(2) Representations for NLP: Encoding

1. String sequence (for example a sentence)
2. Some sort of lookup for encoding
3. First Layer of the NN as Embedding: What dimensionality?

- 1-hot Character encoding
- 1-hot Word encoding
- Byte-pair encoding

(2) Representations for NLP: Pre-trained Word Embeddings

- Instead of training the whole embedding matrix for a given task, one can use pre-trained embedding matrices

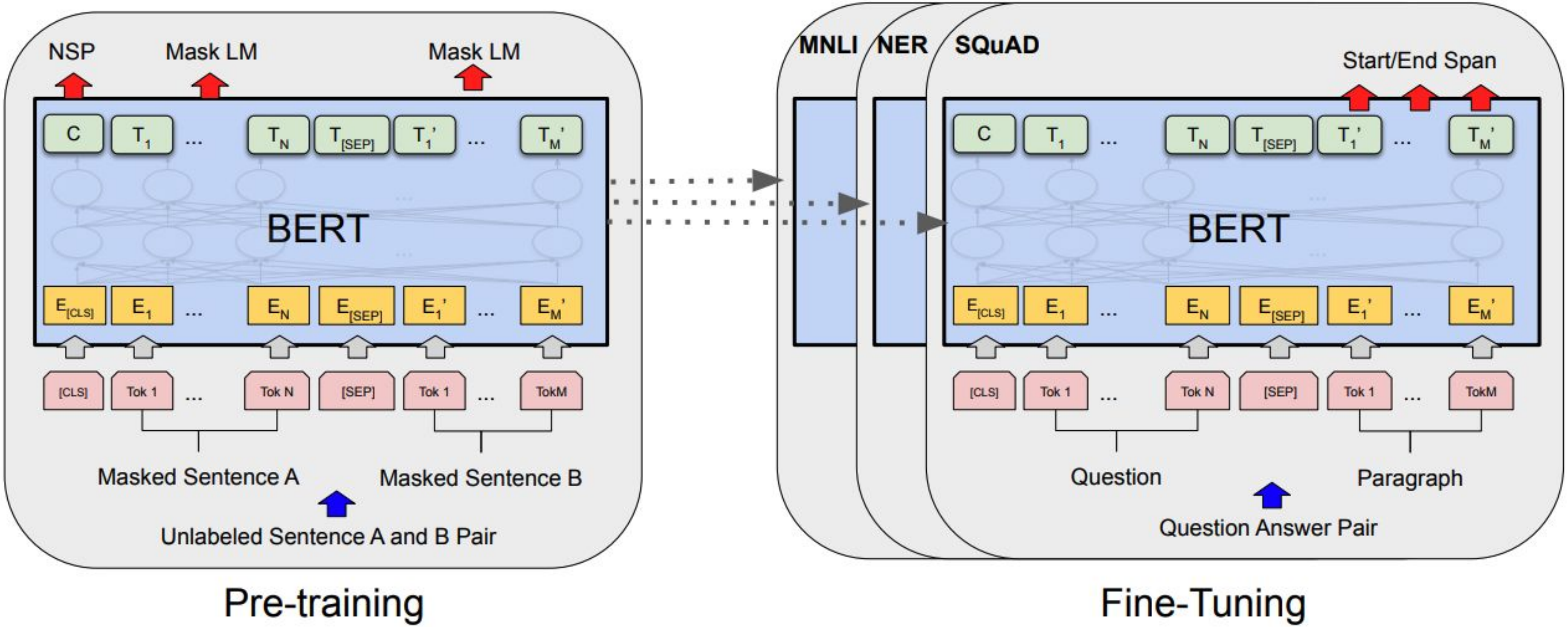
(2) Representations for NLP: Pre-trained Word Embeddings

GloVe: Pennington et al. 2014

- Cooccurrences Y in (VxV)
- Lower-dimensional approximation W in (VxD)
- Such that $W@W^t \approx Y$

(3) Representations for NLP: Contextualized Word Vectors

BERT: Devlin et al. 2019



(3) Representations for NLP: Contextualized Word Vectors

How to get back to general word representations?

Bias Definitions

The discussion has just started

The Social Impact of Natural Language Processing:
https://aclanthology.org/P16-2096.pdf

Bias Definitions 1

Measurement modeling is the process of *operationalizing theoretical constructs* and evaluating those operationalizations

Process:

1. Unobservable theoretical construct ($\mathcal{A}$)
2. Operationalization (a)
3. Measurement (â)

[Jacobs and Wallach 2019]

https://arxiv.org/abs/1912.05511

Bias Definitions 1: Measurement modeling

Example:

- $\mathcal{A}$: Socioeconomic Status
- Operationalized by a = i + p
    - i and p are again operationalized theoretical constructs (I: income, p: property)

[Jacobs and Wallach 2019]

Bias Definitions 1: Measurement modeling

Example: Topic models

[Jacobs and Wallach 2019]

Bias Definitions 1: Measurement modeling

“[..] many of the fairness-related harms that arise from computational systems emerge from the mismatch between unobservable theoretical constructs and their operationalizations.”

[Jacobs and Wallach 2019]

Bias Definitions 2

1. *Historical* ~

2. *Representation* ~

3. *Measurement* ~

(4. *Aggregation* ~)

(5. *Evaluation* ~)

6. *Deployment* ~

[Suresh and Guttag 2019]

Bias in DH

Bias in DH (from the NLP perspective)

Lessons from archives https://arxiv.org/abs/1912.10389

Bias in NLP (from the DH perspective)

NewNLP Project: https://newnlp.princeton.edu/languages/

Bias in DH (from the DH perspective)

LitBank: Born-Literary Natural Language Processing: https://people.ischool.berkeley.edu/~dbamman/pubs/pdf/Bamman_DH_Debates_CompHum.pdf

Bias in DH (from the DH perspective)

Canon

Bias in DH (from the DH perspective)

Social Characters: The Hierarchy of Gender in Contemporary English-Language Fiction: https://culturalanalytics.org/article/11055-social-characters-the-hierarchy-of-gender-in-contemporary-english-language-fiction