Authorship Identification

Nick Fireman

April 15, 2021

Background

Background: Authorship Identification

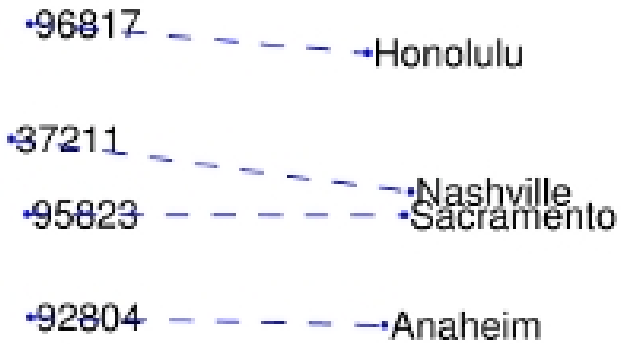
- Authorship identification can be viewed as a type of **classification**.
  - Given a fixed set of potential authors and a passage, which author wrote the passage?
- Previous approaches used language-specific statistical analysis inspired by NLP and **stylometry**:
  - Linguistic complexity
  - Part of speech frequencies
  - Function word frequencies (pronouns, prepositions, conjunctions)

Background: Word Embeddings

- Let W be a set of English words or **tokens**.
  - Tokens can include punctuation, numbers, and contractions (`n't`, `'re`).
- A **word embedding** $e : W \to \mathbb{R}^n$ is a numeric representation of words that aims to **preserve meaning**.
  - W is the **vocabulary** of the embedding.
  - n is the **dimension** of the embedding.
- Similar words are mapped to similar vectors (in Euclidean distance).
  - If $w, w' \in W$ and $|e(w') - e(w)|$ is small, then w' is similar to w.

Background: GloVe

- GloVe is an **unsupervised** algorithm which produces a word embedding from a text dataset.
- Motivated by **word co-occurrence**: two words consistently appearing near each other.
- GloVe embeddings preserve **analogical** relationships between words:



Figures: https://nlp.stanford.edu/projects/glove/

Background: RNN

- Let $e : W \to \mathbb{R}^n$ be an n-dimensional word embedding.
- The embedding e lets us map a passage to a sequence of vectors:
  - Passages are **tokenized** into a sequence $\{w_1, w_2, ..., w_t\}$ in W.
  - Then **embedded** to a sequence $\{e(w_1), e(w_2), ..., e(w_t)\}$ in $\mathbb{R}^n$.
- The RNN architecture is well-suited for understanding sequential data:

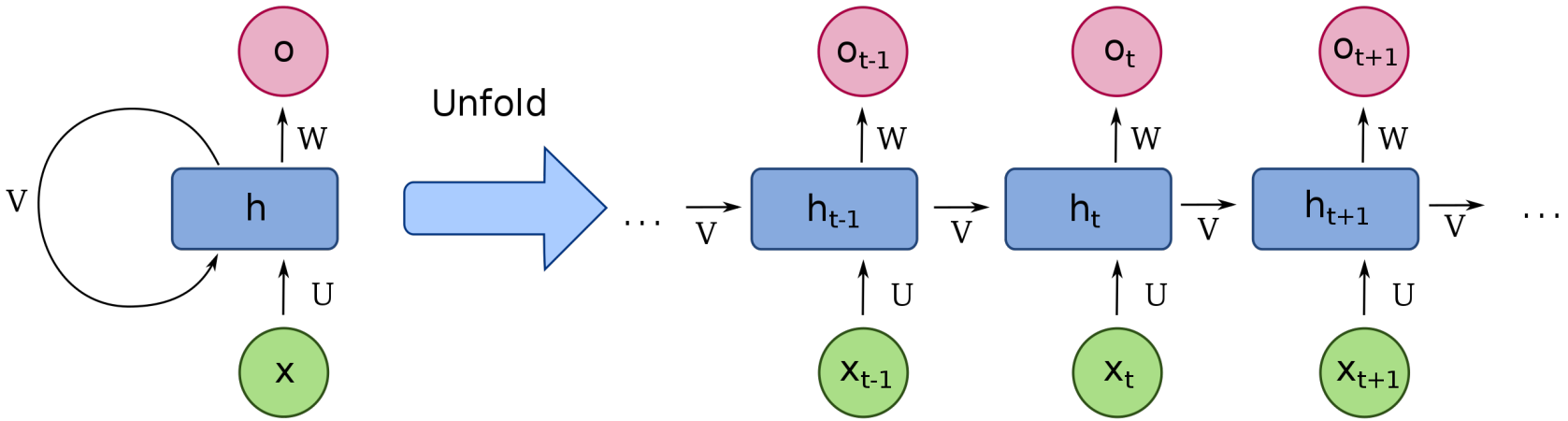


Figure: Wikipedia

Background: Architecture

- This project attempts to classify **short passages** written by the two authors Charles Dickens and Mark Twain.
- Our high-level approach to author identification:
  - Extract passages from a literary dataset for training and testing.
  - Use a word embedding to map each passage to a sequence of vectors.
  - Use an RNN to classify the sequential data.

Datasets

Datasets: Project Gutenberg

- Project Gutenberg hosts public domain English literature.
- A 2015 research work[1] collected and **cleaned** 3,036 books on PG into a single dataset.
  - Metadata, licenses, and transcribers' notes were removed from the original documents.
- 47 books by Mark Twain (3.3 million tokens).
- 61 books by Charles Dickens (7.1 million tokens).

[1] Shibamouli Lahiri. "Complexity of Word Collocation Networks: A Preliminary Structural Analysis"

Datasets: Pretrained GLoVE embeddings

- The researchers who introduced GloVe also provide **pretrained** word embeddings produced by GloVe.
- Embedding 1 (`6B.50d`) trained on 6 billion tokens from Wikipedia.
  - Dimension 50.
  - 400K-word vocabulary.
- Embedding 2 (`42B.300d`) trained on 42 billion tokens from Common Crawl.
  - Dimension 300.
  - 1.9M-word vocabulary.

Experiments

Experiments: Challenges

- Dickens (1812-1870) and Twain (1835-1910) were contemporaries, meaning they may have a similar writing style.
- 19th-century writing may not be accurately captured by pretrained GloVe models trained on 21st-century language.
- 12,361 tokens parsed did not lie in the GloVe vocabulary:

```
cripplewayboo
davenportseseses
methoozellers
schnorkel
summonsizzing
witchingest
```

Experiments: Hyperparameter Tuning

- A round of 18 experiments was held for **hyperparameter tuning**.
  - 2 pretrained GloVe embeddings: `6B.50d`, `42B.300d`.
  - 3 RNN hidden sizes: 25, 50, 100.
  - 3 learning rates: 0.0005, 0.001, 0.005.
- For each experiment:
  - 40,000 training passages and 10,000 testing passages are sampled from the works of each author.
  - Each passage is 30 tokens.
  - The model is trained for 100 epochs, with testing every 10 epochs.

Experiments: 6B.50d GloVe

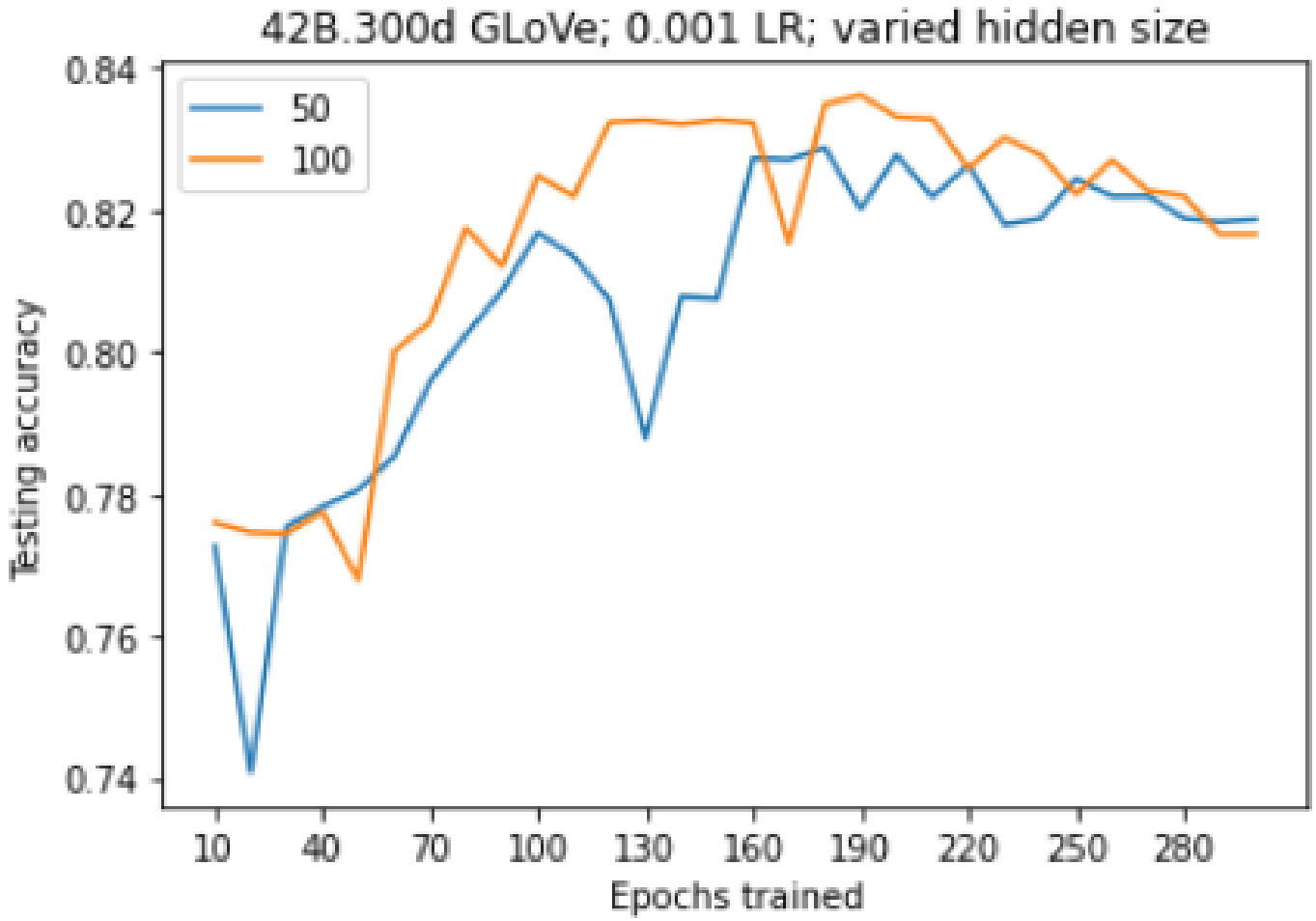
GLoVE	Hidden Size	LR	Peak Test Acc.	Final Acc.
6B.50d	25	0.0005	0.7597 (90 epochs)	0.7586
		0.001	0.7854 (100 epochs)	0.7854
		0.005	0.7972 (100 epochs)	0.7972
	50	0.0005	0.7593 (90 epochs)	0.7440
		0.001	0.7809 (100 epochs)	0.7809
		0.005	0.7956 (100 epochs)	0.7956
	100	0.0005	0.7631 (100 epochs)	0.7631
		0.001	0.7897 (100 epochs)	0.7897
		0.005	0.7947 (100 epochs)	0.7947

Experiments: 6B.300d GloVe

GLoVE	Hidden Size	LR	Peak Test Acc.	Final Acc.
42B.300d	25	0.0005	0.7966 (90 epochs)	0.7813
		0.001	0.8036 (80 epochs)	0.7766
		0.005	0.8075 (60 epochs)	0.7840
	50	0.0005	0.7990 (100 epochs)	0.7990
		0.001	**0.8194 (90 epochs)**	**0.7929**
		0.005	0.7967 (60 epochs)	0.7810
	100	0.0005	0.7959 (90 epochs)	0.7876
		0.001	**0.8083 (90 epochs)**	**0.8067**
		0.005	0.8129 (70 epochs)	0.7949

Experiments: Final Results

- The best two models were trained for 300 epochs.
- Accuracy decreased after 200 epochs, indicating overfitting.



Thanks for your attention!