Kissipo Learning for Deep Learning

Topic 24: Introduction to Word Embedding (20min)

Hsueh-Ting Chu

KLDL-W9-T24

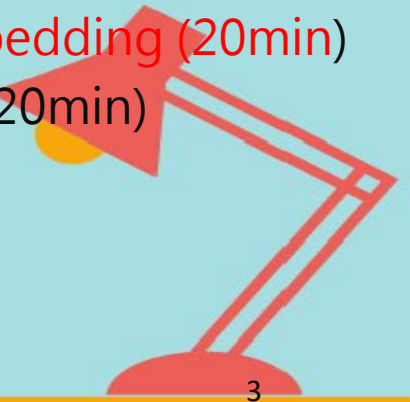
Course Schedule

- W1 - Course Introduction
- W2 - DL Programming Basics(1)
- W3 - DL Programming Basics(2)
- W4 - DL with TensorFlow
- W5 - Midterm
- W6 - DL with PyTorch
- W7 - AOI hands-on project
- W8 - RSD hands-on project
- W9 - NLP hands-on project
- W10 - Final exam

DL: Deep Learning
AOI: Automated Optical Inspection
RSD: Road Sign Detection
NLP: Natural Language Processing

Topics

- Topic 01: Introduction to Deep Learning (20min)
- Topic 02: Kissipo Learning for Deep Learning (20min)
- Topic 03: Python quick tutorial (20min)
- Topic 04: Numpy quick tutorial (15min)
- Topic 05: Pandas quick tutorial (15min)
- Topic 06: Scikit-learn quick tutorial (15min)
- Topic 07: OpenCV quick tutorial (15min)
- Topic 08: Image Processing basics (20min)
- Topic 09: Machine Learning basics (20min)
- Topic 10: Deep Learning basics (20min)
- Topic 11: TensorFlow overview (20min)
- Topic 12: CNN with TensorFlow (20min)
- Topic 13: RNN with TensorFlow (20min)
- Topic 14: PyTorch overview (20min)
- Topic 15: CNN with PyTorch (20min)
- Topic 16: RNN with Pytorch (20min)
- Topic 17: Introduction to AOI (20min)
- Topic 18: AOI simple Pipeline (A) (20min)
- Topic 19: AOI simple Pipeline (B) (20min)
- Topic 20: Introduction to Object detection (20min)
- Topic 21: YoloV5 Quick Tutorial (20min)
- Topic 22: Using YoloV5 for RSD (20min)
- Topic 23: Introduction to NLP (20min)
- Topic 24: Introduction to Word Embedding (20min)
- Topic 25: Name prediction project (20min)

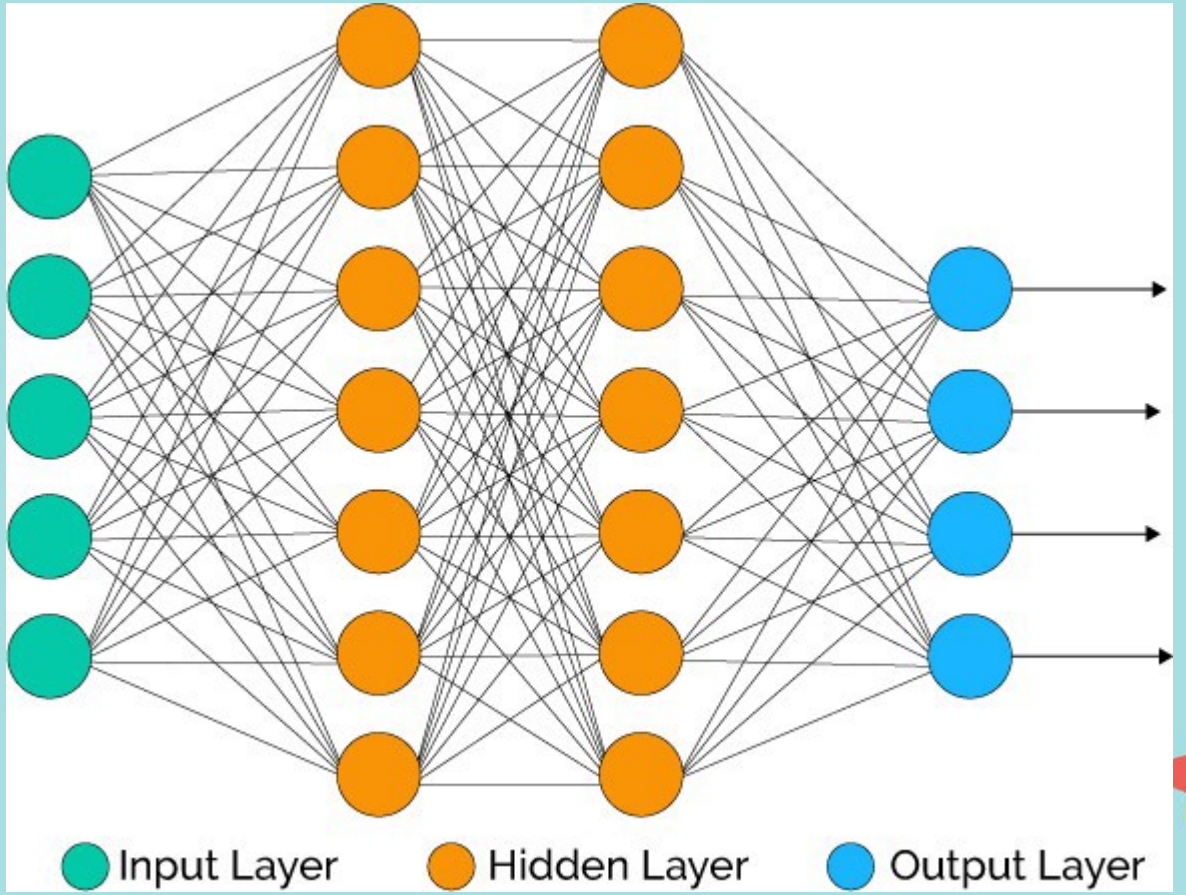


Week 9 Topics

- Topic 23: Introduction to NLP (20min)
- Topic 24: Introduction to Word Embedding (20min)
- Topic 25: Name prediction project (20min)

NLP with Deep Learning Models

Input:
Monica is a dog



Output of Natural Language Processing

- What is natural language processing
- Input is plain text.
- What is the output?
  - Authorship Attribution->Who
  - Sentiment analysis -> emotions (angry, happy, sad)
  - Text Summarization->short sentences
  - Recommender System->goods
  - Language translation->sentences in another language
  - Question Answering->answers (plain text)

Classical Models of NLP

How was NLP done before Deep Learning?

- HMMs, CRFs and other types of PGM
- Bag of words/ngrams – feature per word/ngram

the cat sat on the mat

cat	sat	on	the	mat	quickly
1	1	1	2	1	0

… |Vocabulary|

- Various ways of choosing the values: Binary, Count, TF-IDF

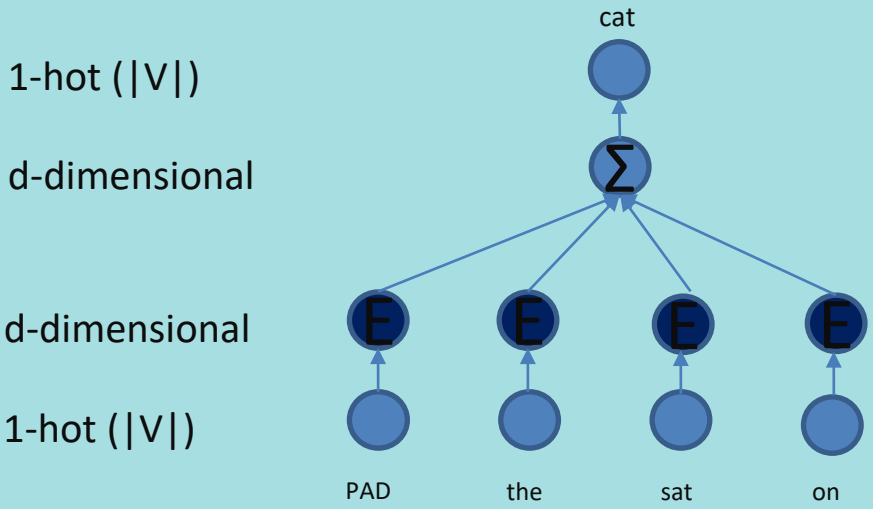
Natural Language Processing Models

- **N-grams**
- Word Embedding
- Word2Vec
- Sequence to sequence
- Attention
- BERT

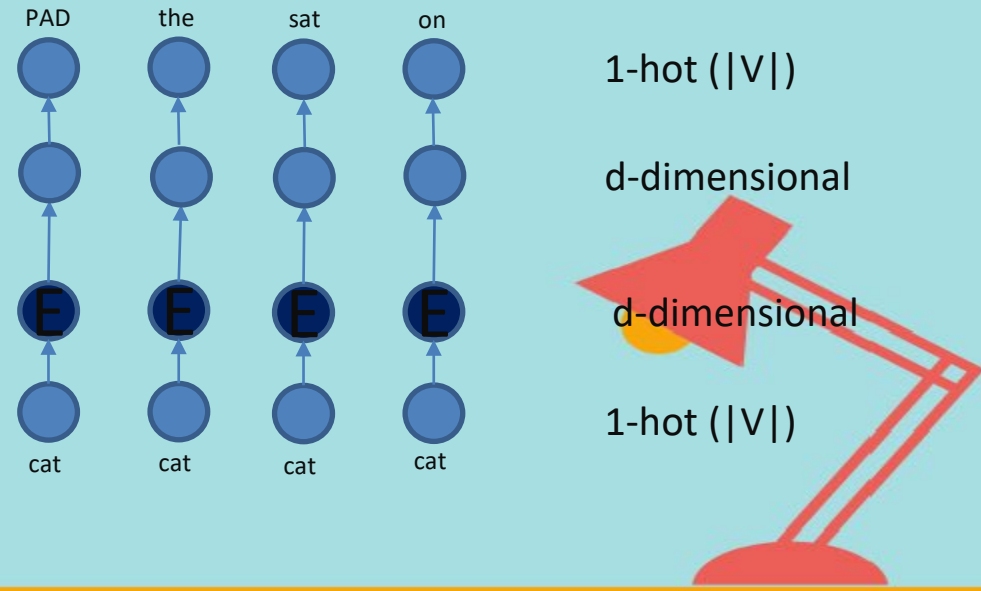
Word2Vec

- [Mikolov et al., 2013](#) (while at Google)
- Linear model (trains quickly)
- Two models for training embeddings in an *unsupervised* manner:

Continuous Bag-of-Words (CBOW)

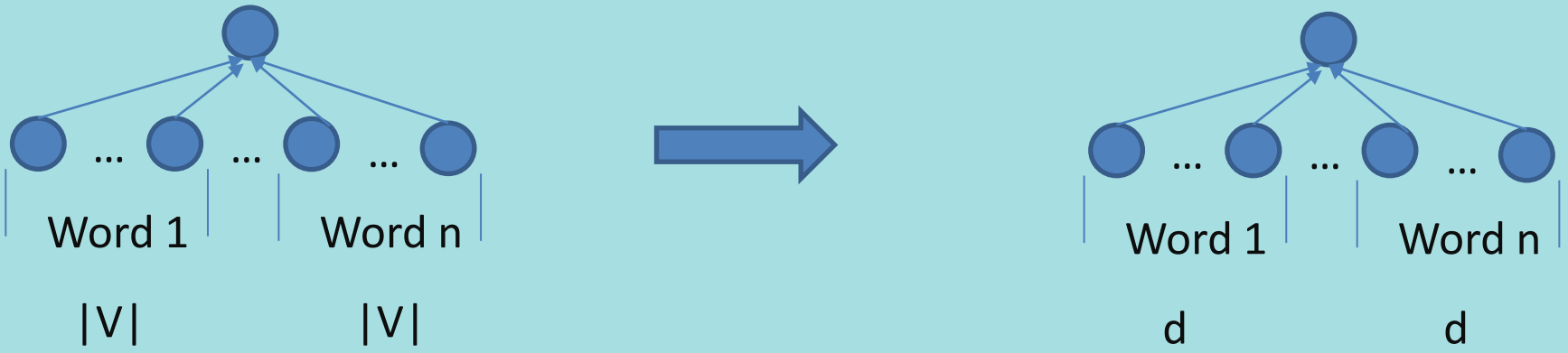


Skip-Gram



Properties of embeddings (cont.)

- They are used as input of NLP tasks:



Total # of input parameters: n * |V| >> n * d

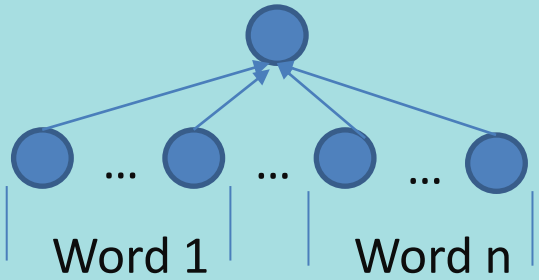
The Distributional Hypothesis (Firth, 1957)

- ‘You can tell a word by the company it keeps’

  - *The cat sat on the mat*
  - *The dog sat on the mat*
  - *The elephant sat on the mat*
  - *The quickly sat on the mat*

Issues with BOW Representation

- Sparse Input (1-hot)



p >> n (overfitting!)

- No semantic generalization
- *dog:* 1 0 0 0 0 … 0
- *cat:* 0 0 1 0 0 … 0

lots of data required, low accuracy

Properties of embeddings

- They are dense: 0.53, 0.2, -1.2, ….
- They are low dimensional: (typically 50 <= d <= 300)
- Embed domain semantics:

  `king’ - `man` + `woman` ≈ `queen`

  `paris` - `france` + `spain` ≈ `madrid`
- Generalize easily!

Authorship Attribution

- Given a set of documents and a set of authors, which authors wrote which documents?
- We need to learn the ‘style’ of each author: common words and phrases, sentence length, use of punctuation, etc. This is known as the science of Stylometry.

The Federalist Papers problem

The Federalist Papers

31 languages

Article Talk | Read Edit View history

From Wikipedia, the free encyclopedia

"The Federalist" redirects here. For the website, see The Federalist (website). For other uses, see Federalist (disambiguation).

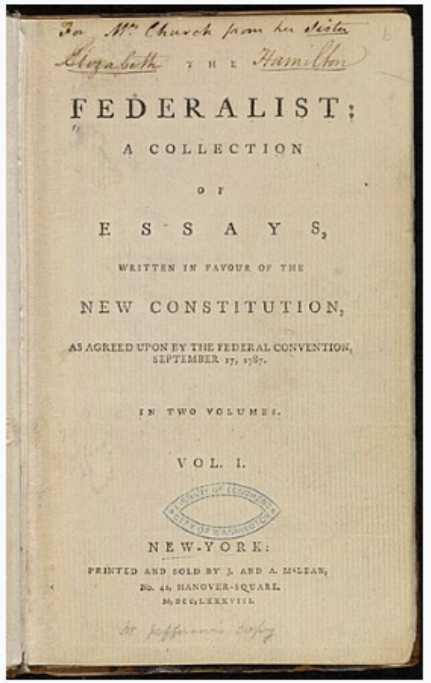
The Federalist Papers is a collection of 85 articles and essays written by Alexander Hamilton, James Madison, and John Jay under the collective pseudonym "Publius" to promote the ratification of the Constitution of the United States. The collection was commonly known as ***The Federalist*** until the name *The Federalist Papers* emerged in the 20th century.

The first 77 of these essays were published serially in the *Independent Journal*, the *New York Packet*, and *The Daily Advertiser* between October 1787 and April 1788.[1] A compilation of these 77 essays and eight others were published in two volumes as ***The Federalist: A Collection of Essays, Written in Favour of the New Constitution, as Agreed upon by the Federal Convention, September 17, 1787***, by publishing firm J. & A. McLean in March and May 1788.[2][3] The last eight papers (Nos. 78–85) were republished in the New York newspapers between June 14 and August 16, 1788.

The authors of *The Federalist* intended to influence the voters to ratify the Constitution. In Federalist No. 1, they explicitly set that debate in broad political terms:

> It has been frequently remarked, that it seems to have been reserved to the people of this country, by their conduct and example, to decide the important question, whether societies of men are really capable or not, of establishing good government from reflection and choice, or whether they are forever destined to depend, for their political constitutions, on accident and force.[4]

The Federalist Papers



https://en.wikipedia.org/wiki/The_Federalist_Papers

The Federalist Papers

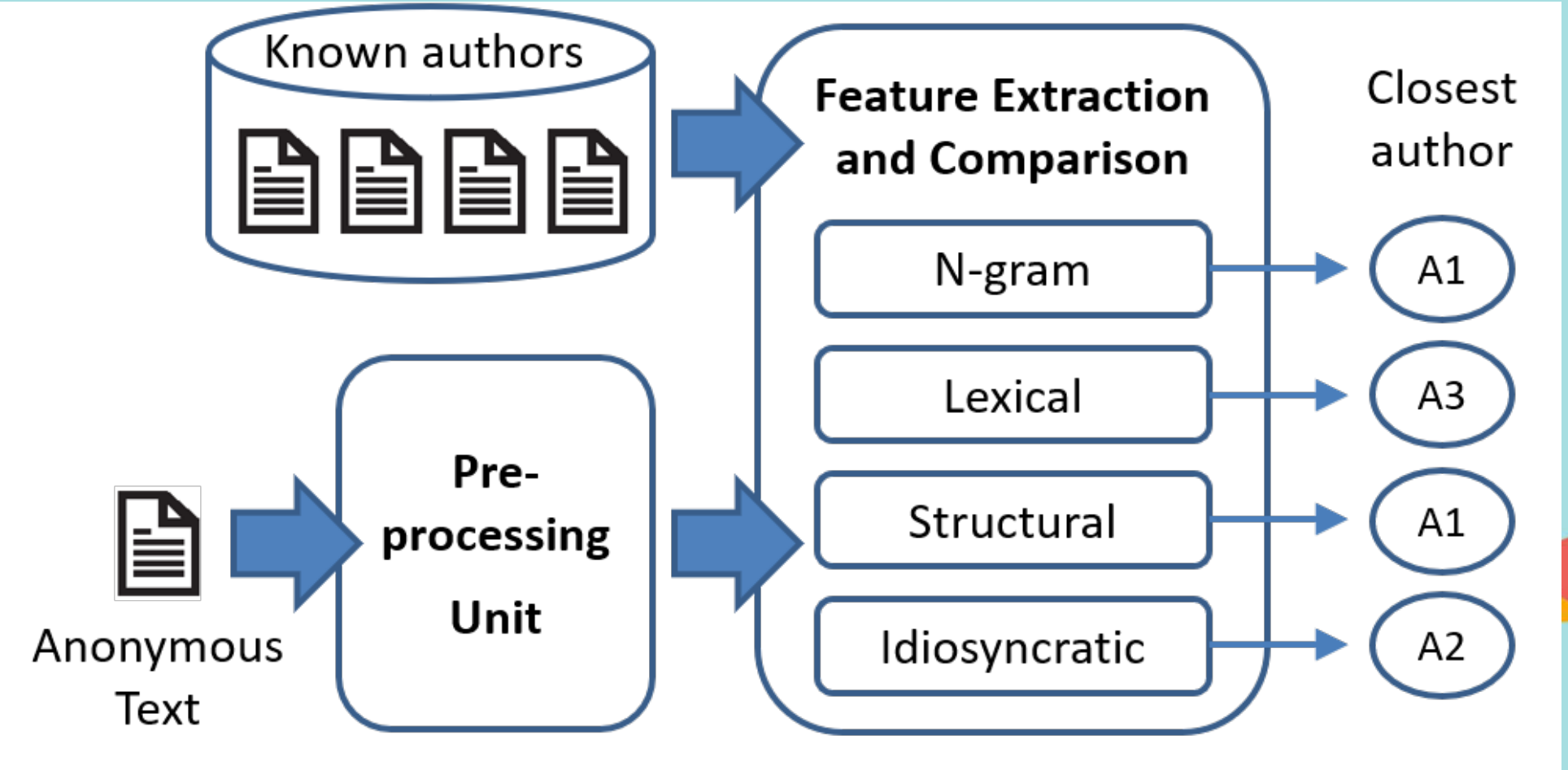
- Set of 85 papers written between 1787 and 1788.
- The authorship of 12 of these papers is disputed between Hamilton and Madison.

•Alexander Hamilton (51 articles: Nos. 1, 6–9, 11–13, 15–17, 21–36, 59–61, and 65–85)
•James Madison (17 articles: Nos. 10, 14, 18–20, 37–48)
•John Jay (5 articles: Nos. 2–5 and 64).
•Disputed (12 articles: Nos. 49-58, 62,63) Hamilton or Madison

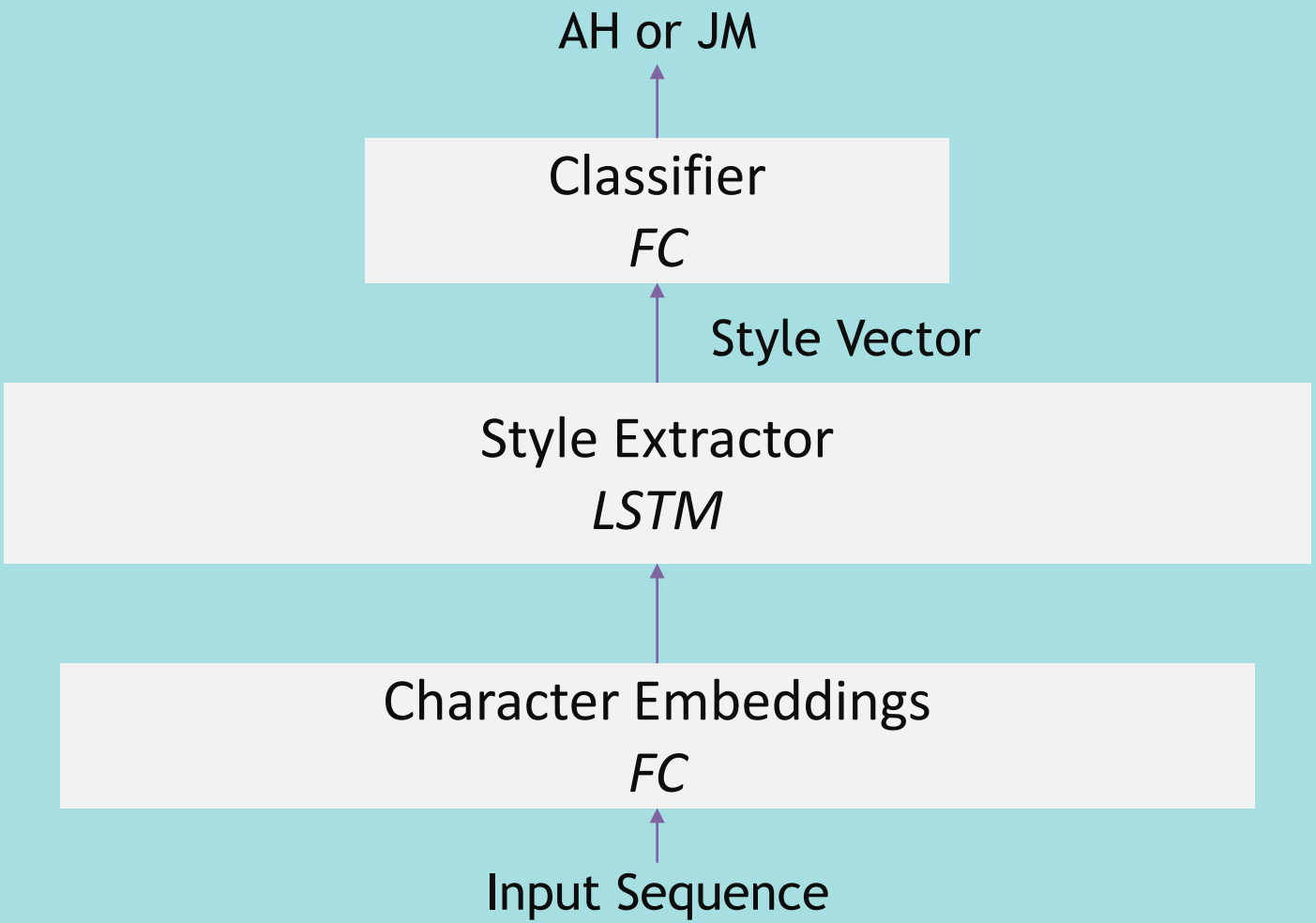
Assumptions

- Single author per document – either Hamilton or Madison
- Author style is consistent across multiple documents

Authorship Attribution Algorithm



Network Architecture



Thanks!

Q&A