Natural Language Processing

...

≠ Neuro-Linguistic Programming

(even if ↗ is the first result on Google when you search for NLP)

About me

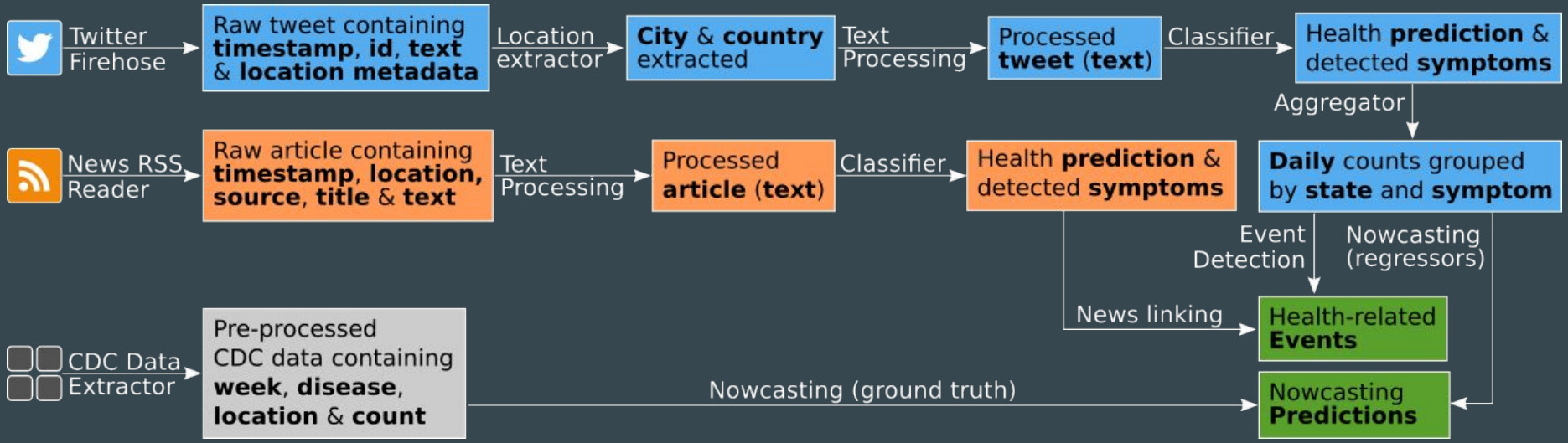
Ovidiu Șerban (https://ovidiu.roboslang.org/)

Research Associate @ Imperial College London

- Institute for Security Science and Technology (ISST)
- Data Science Institute (DSI)

Natural Language Processing, Machine Learning, Data Visualization, ...

Recent work



Șerban, O., Thapen, N., Maginnis, B., Hankin, C., & Foot, V. (2018). Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing & Management*.

My NLP process

1. **Problem definition**

   What are you trying to achieve?

2. **Data acquisition**

   How are you going to get the data out?

3. **Cleaning & parsing/tokenization**

4. **Feature representation**

   Vectors? Tokens? Paragraphs? Documents?

5. **Classification**

6. **Metrics**

Loop until you're satisfied with the result

Problem definition

Part of Speech (PoS) Tagging

- Input: Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

- Output: Profits/N soared/V at/P Boeing/N Co./N ,/, easily/ADV topping/V forecasts/N on/P Wall/N Street/N ,/, as/P their/POSS CEO/N Alan/N Mulally/N announced/V first/ADJ quarter/N results/N ./.

- N: Noun; V: Verb; P: Preposition; Adv: Adverb; Adj: Adjective ….

https://www.quora.com/What-Apache-OpenNLP-POS-Tagger-does

Tagging variations

- POS tagging problem - usually all tokens are tagged with a label
- Entity recognition - not all tokens are tagged

  Elon Musk settles with SEC over fraud charge
- Segmentation - multiple tokens become part of the same entity
- Document classification - the entire document is tagged with a label

  Elon Musk settles with SEC over fraud charge ⇝ positive | + 0.75

Translation/Language Models/Generation

- Generate a sequence of tokens (text) starting from another sequence

  (EN) *Elon Musk settles with SEC over fraud charge*
  (FR) Elon Musk règle avec SEC sur les accusations de fraude
  (RO) Elon Musk recurge la o rezoluție amiabilă cu SEC asupra acuzațiilor de fraudă

- Language Models → predicting the next token in a sequence of text

- Text generation → generate a sequence of text starting from a random token

Data acquisition

Checklist

- Source, license, etc ...
- Data quality
- Document markup & metadata
- Annotations
  - E.g. Amazon's Mechanical Turk
- Annotation confidence
  - Inter-annotator agreements
  - Weighted inter-annotator agreements (experts vs non-experts)

Cleaning, parsing and tokenization

Notes

- Cleaning, parsing and tokenization is problem dependent

  I use this NLTK parser that this tutorial recommended, but my classifier does not work properly
- Be careful not to strip too much context with your cleaning
  - Document markup may be useful
  - Punctuation may help
  - Stop word removal (which everybody does in traditional NLP) may not be the best thing
  - Stemming may hurt your performance
- Clean less on first iterations

Cleaning

- Punctuation
- Stop words
- URLs, usernames
- Hashtags
- Emoji and emoticons (https://emojipedia.org/)
  - Translation: 😀 → Grinning Face
  - **Note:** Emoji/emoticons may have different meanings on different platforms depending on the icon
- Stemming
  - fishing, fished, fisher → fish
  - argue, argued, argues, arguing → argu
- Lemma
  - go, goes, going, went, gone → go

Tokenization

- Language dependent
- Source dependent (they deal differently with tweets/ wikipedia/ news articles, etc)
- Hashtag/url tokenization
- Word expansion
  - Don’t → do not

Data representation

Notes

- Data representation is highly dependent on the classifier
- Numeric/vectorized representations
- Discrete representation (raw tokens)
- Co-occurrences (NGram frequencies)

One hot encoding

	o_ENE	o_ESE	o_East	o_NE	o_NNE	o_NNW	o_NW	o_North	o_SE	o_SSE	o_SSW	o_SW	o_South	o_Variable	o_WSW
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

TF-IDF

TF-IDF Score

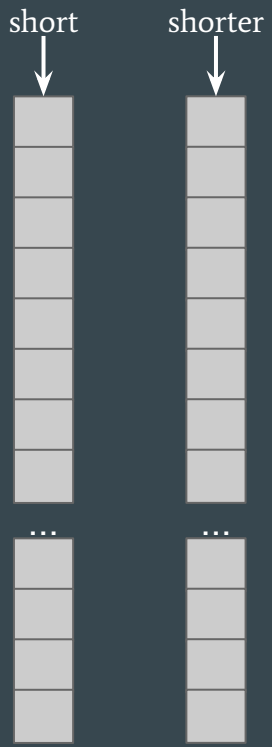
$$TF - IDF\ Score = TF_{x,y} * IDF = TF_{x,y} * \ log\frac{N}{df} \ldots\ldots (1)$$

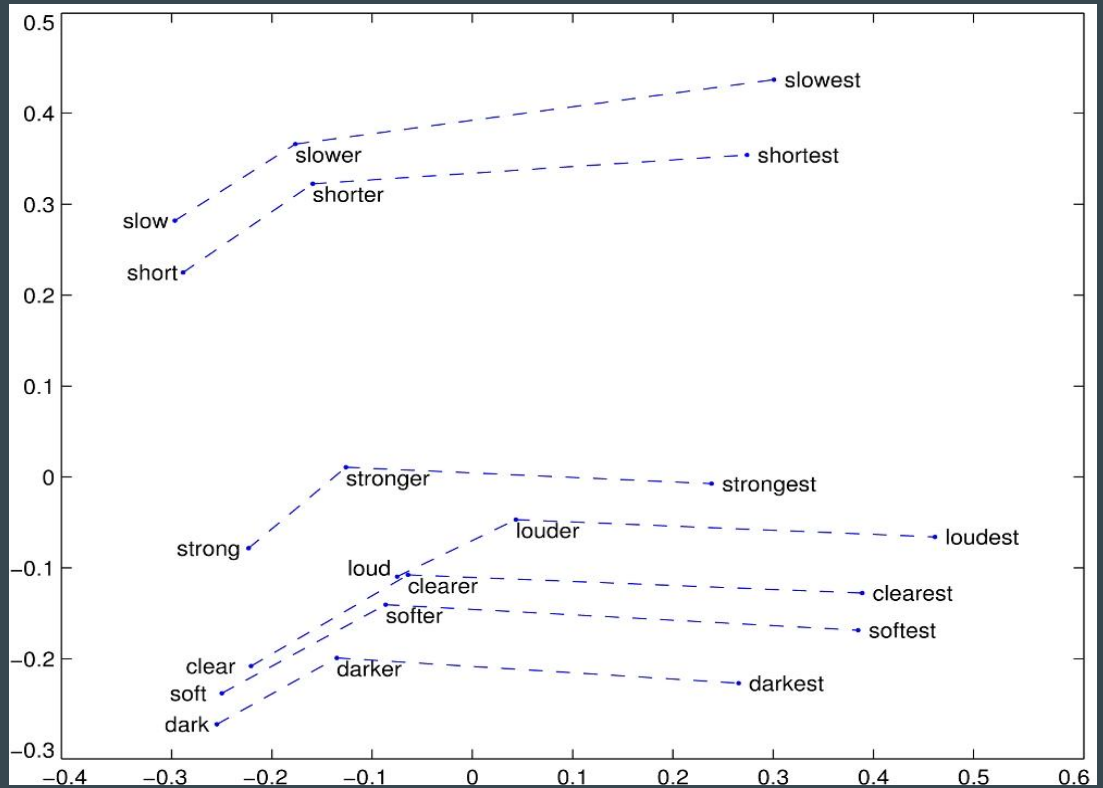
, where $TF_{x,y}$ *is the frequency of keyphrase X in the article Y,*

N is the total number of documents in the corpus.

df is the number of documents containing keyphrase X

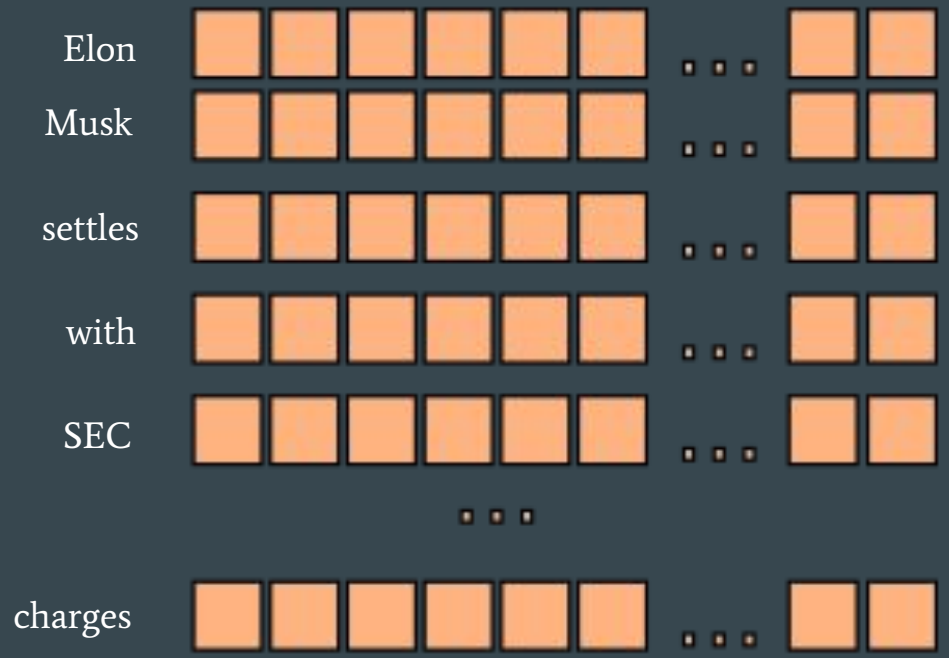
Vectorization





Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation

Document vectorization



Which vector model to choose?

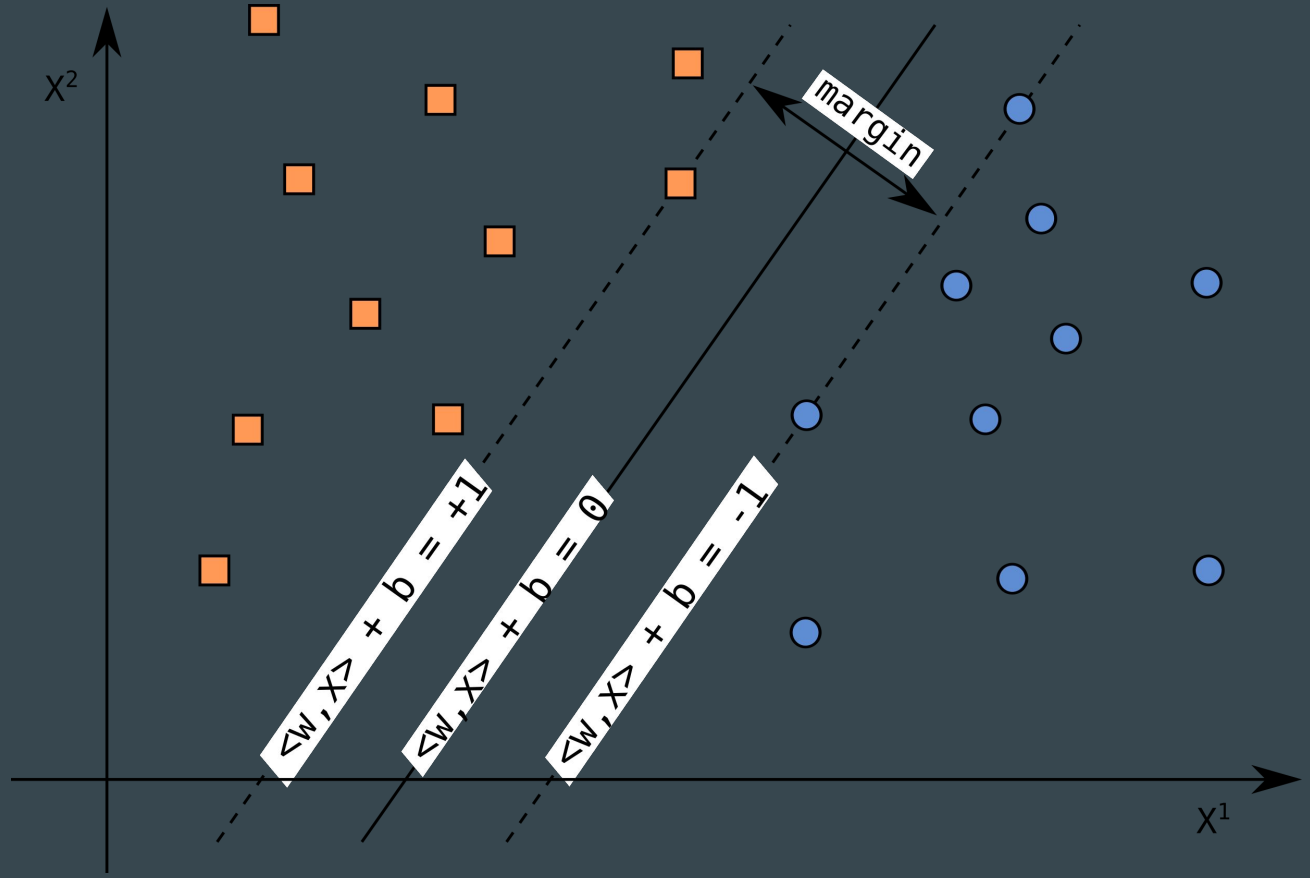
Flavor	Dataset	Vector size	Rare Words			SimLex 999			WordSim 353		
			score	TR	oov	score	TR	oov	score	TR	oov
FastText	Twitter	300	48.92	**36.26**	14%	36.76	**36.76**	0%	69.95	**69.95**	0%
GloVe	Twitter	300	45.82	2.56	83%	14.64	13.86	1%	57.26	46.55	8%
Word2Vec	Twitter	300	38.70	31.60	14%	31.40	31.40	0%	61.44	61.44	0%
FastText	OpenSubtitles	300	51.73	**38.89**	13%	41.85	**41.85**	0%	71.97	**71.97**	0%
GloVe	OpenSubtitles	300	33.41	20.53	49%	16.40	16.57	0%	53.58	53.47	0%
Word2Vec	OpenSubtitles	300	35.44	28.57	13%	34.88	34.88	0%	60.27	60.27	0%

Document vectorization

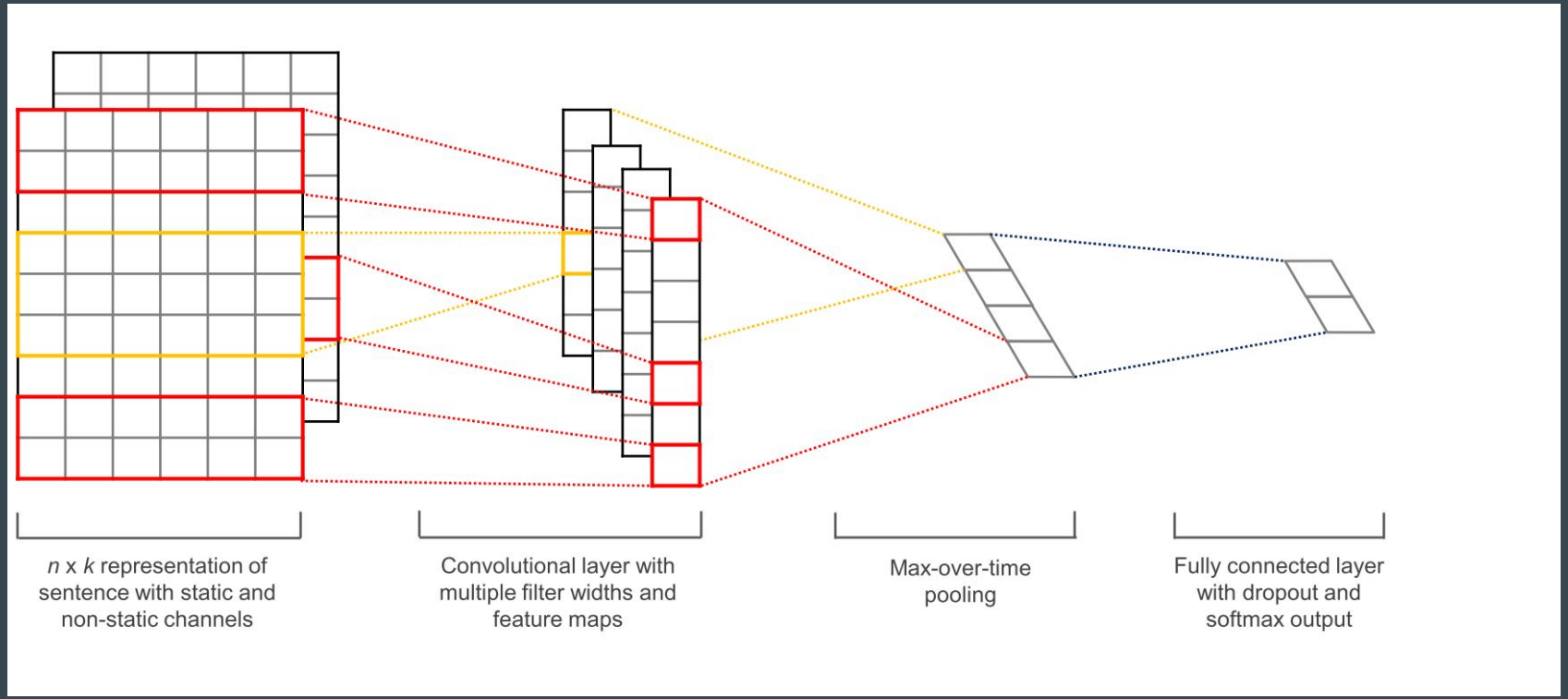
- Similar to token vectorization → one vector representation for the whole document (latent representation)

Classifiers

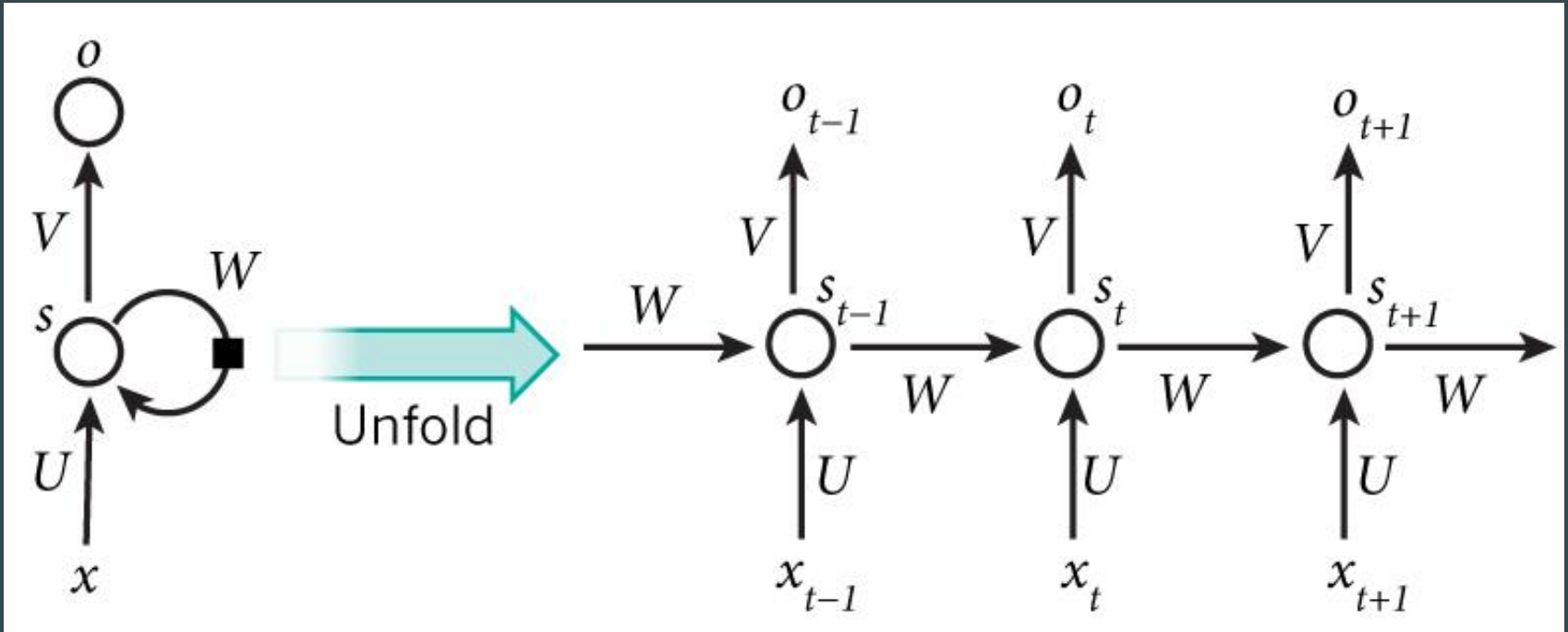
Linear SVM



CNN


n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

LSTM



Metrics

Metrics

- Precision = tp / (tp + fp) Recall = tp / (tp + fn)
  - Precision: probability that a randomly selected retrieved sample is relevant
  - Recall: probability that a randomly selected relevant sample is retrieved in a search.
- Accuracy = (tp + tn) / total
- F1 = 2*P*R / (P + R)
- Error rate = 1 - Accuracy
- Domain dependent:
  - BLEU (bilingual evaluation understudy) → translation quality score

Error analysis

- Vector representation issues:
  - Out of Vocabulary (OOV) problems
  - Synonyms
  - Rare words
- Lack of context - too much cleaning, stemming issues
- Negation
- Lack of data for certain classes

Questions