Introduction to natural language processing for health and biological questions

Nona Naderi
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique

université PARIS-SACLAY

Vector-based semantics

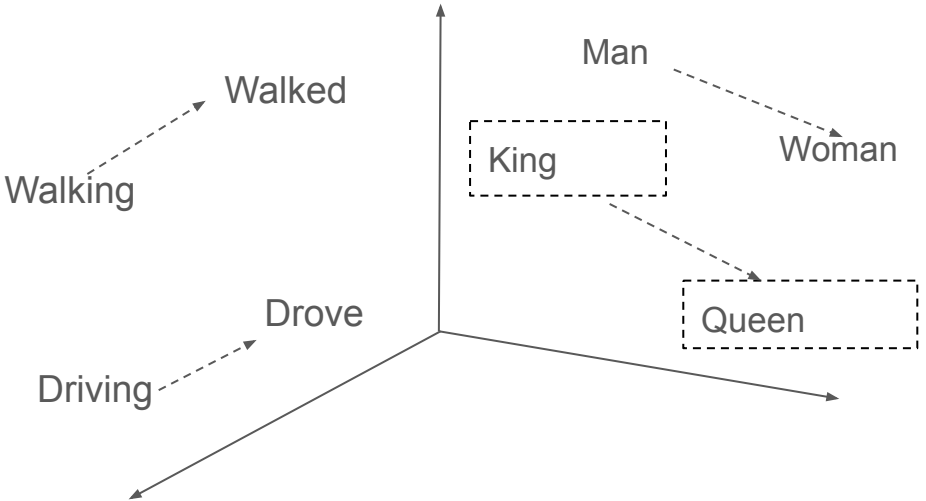
You can get a lot of information by representing a word using it context.

“Explore **bank** accounts, loans, mortgages, investing, credit cards & **banking** services”

“Measurements of **bank** erosion on rivers in Devon over a 2½ year period produced mean rates ranging from 0.08 to 1.18 metres per year and a maximum rate of 2.58 metres per year.”

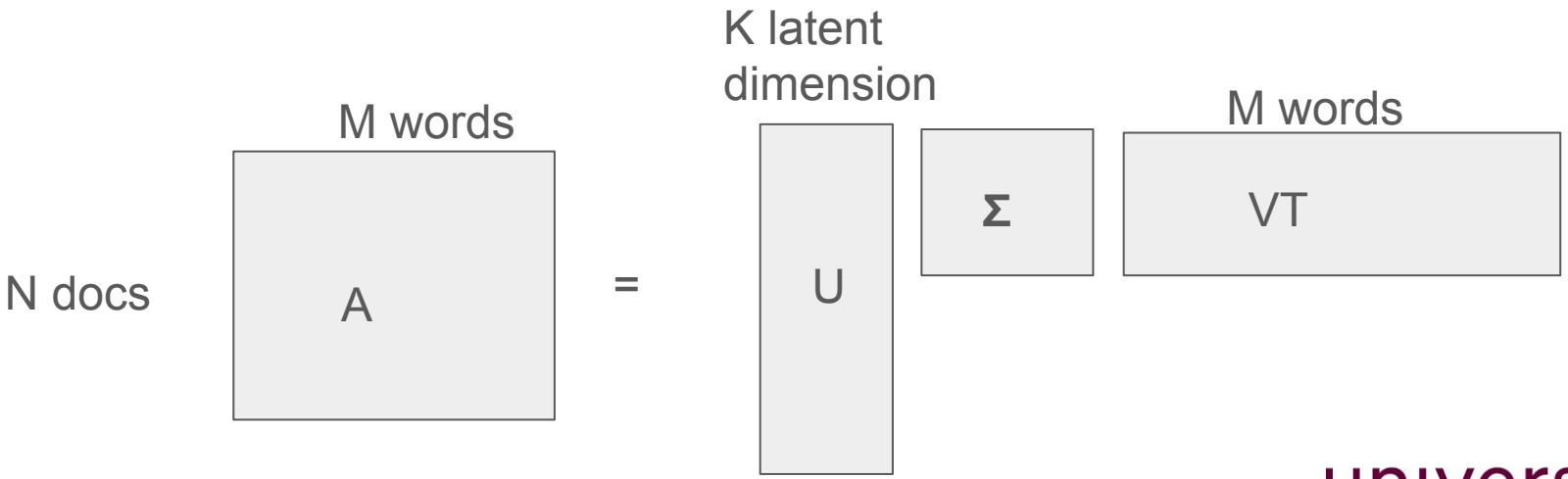
Distributed, distributional representations

With these representations, syntactic and semantic patterns are captured.



Count vs. Predict

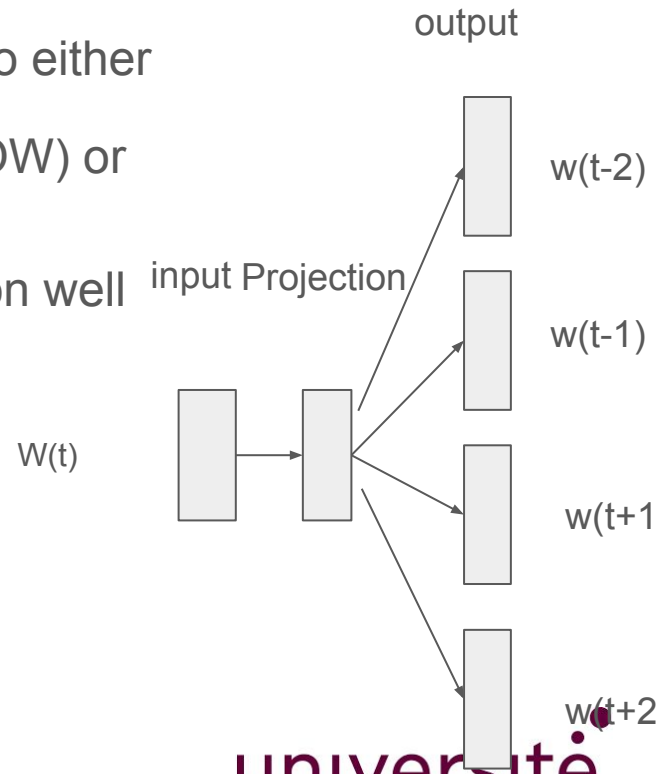
Latent Semantic Analysis (LSA): Term-document co-occurrence matrix with dimensionality reduction using Singular Value Dicomposition (SVD)



Word2vec

Word2vec CBOW/SkipGram: Train word vectors to try to either

- Predict a word given its bag-of-words context (CBOW) or
- Predict a context word from the center word
- Update word vectors until they can do this prediction well



Exercise

Use pre-trained word2vec model (google-news-300) to find the most similar terms to dog. (you can use Gensim package)

Language model

A language model is a probability distribution over words or word sequences.

What is the probability of “ I go to Paris-Saclay University”?

What is the probability of “To go I University Paris-Saclay”?

$P(w) = P(w_1 w_2 w_3 w_4 \ldots w_k)$

$P(w_k \mid w_2 w_3 \ldots w_{k-1})$

N-grams models

Unigram model: $P(w_1)P(w_2) \ldots P(w_n)$

Bigram model: $P(w_1)P(w_2|w_1)P(w_3|w_2)...P(w_n|w_{n-1})$

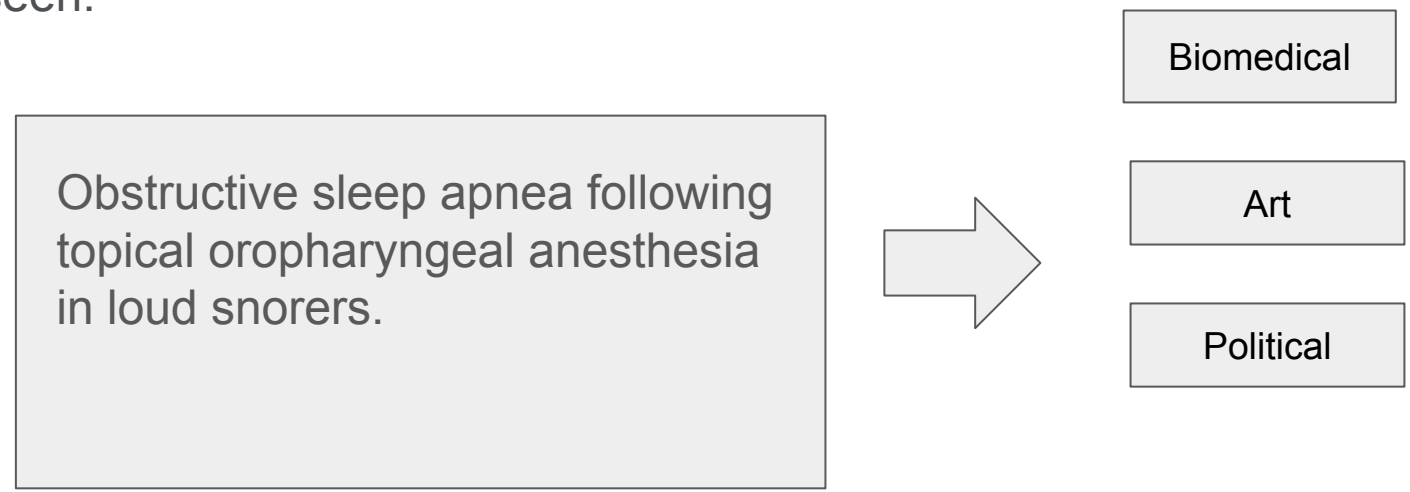
Trigram model: $P(w_1)P(w_2|w_1)P(w_3|w_2,w_1).....P(w_n|w_{n-1}w_{n-2})$

Markov chain

- A Markov model of *order 0* predicts that each letter in the alphabet occurs with a fixed probability.
- We can fit a Markov model of order 0 to a specific piece of text by counting the number of occurrences of each letter in that text, and using these counts as probabilities.
- Example: “baggage” p(a) = 2/7

Text classification

- A classifier is a function that maps inputs to a predefined sets of class labels.
- Our classifier needs to be able to classify items that our model has never seen.



Depression detection

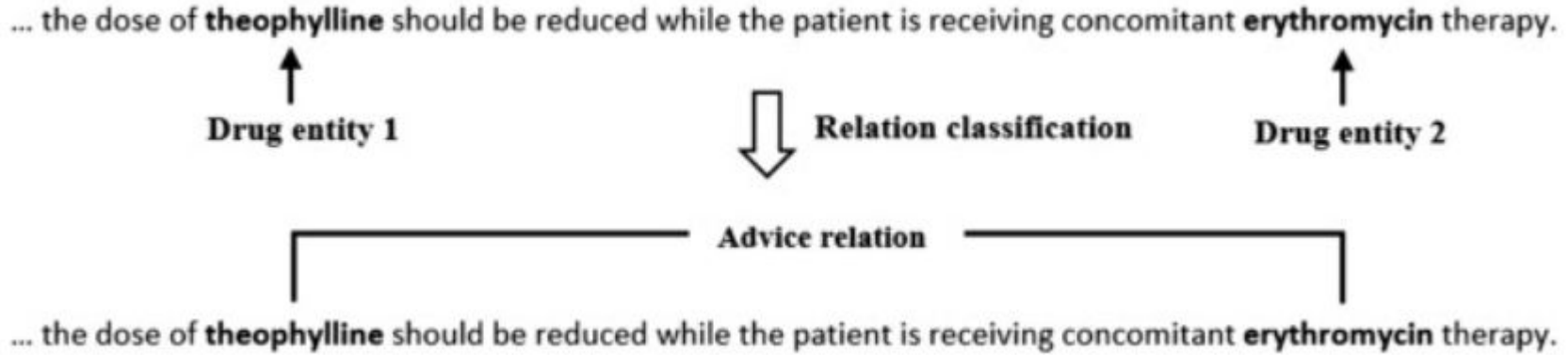
- “I feel lost and disconnected these days. Everything always goes wrong with me. It feels like I never have enough.” (Oak 2017)

What is the author’s state? Depressed or not?

What about the states of the following messages’ authors? (Uddin et al., 2022)

1. I've been so busy lately, and not felt any good. My doctor did not seem to bother. Schoolwork can be quite enjoyable, but now I have lost motivation to do anything; spend my days thinking about what I need to do, but it's challenging to get started. I'm not happy anyway, so I often think it would be better not to live. I kind of have no feelings. What's happening to me?
2. I feel mentally exhausted and struggle to get through everyday life. I cry every day. Getting up in the morning feels like a struggle. I know I have to get through it, but it troubles me a lot, both mentally and physically. If I have to do something after work, I must sleep to be able to manage. Motivation is gone. Work is no longer enjoyable. Some days I sleep a lot, other days, nothing. What shall I do?

Biomedical relation classification



(Zhang et al., 2019)

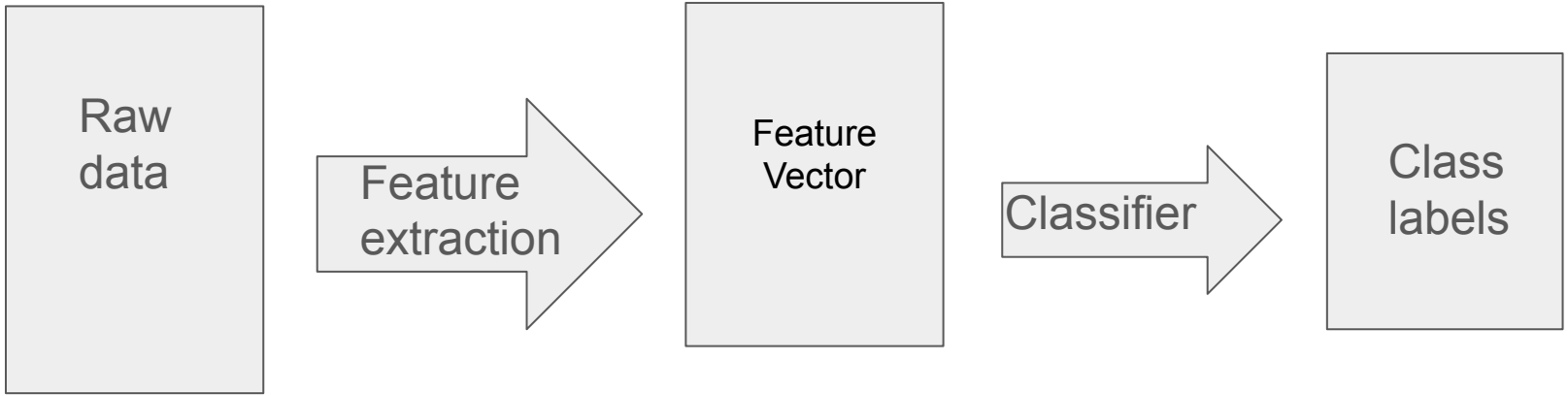
Text classification examples

- Gender identification
- Subject classification
- Dementia detection
- …

Text classification definition

- **Input:**
  - a document, a sentence, …
  - A fixed set of classes
- **Output:** A predicted class

Text classification



Approaches to classification

- Rule-based (accurate but expensive)
- Supervised machine learning

Bag of words feature representation

- **SARS-CoV-2** has rapidly spread across the globe and **infected** hundreds of millions of people worldwide. As our experience with this virus continues to grow, our understanding of both short-term and long-term complications of **infection** with **SARS-CoV-2** continues to grow as well.

SARS-CoV-2	2
infected	1
infection	1
...	**...**

Evaluation

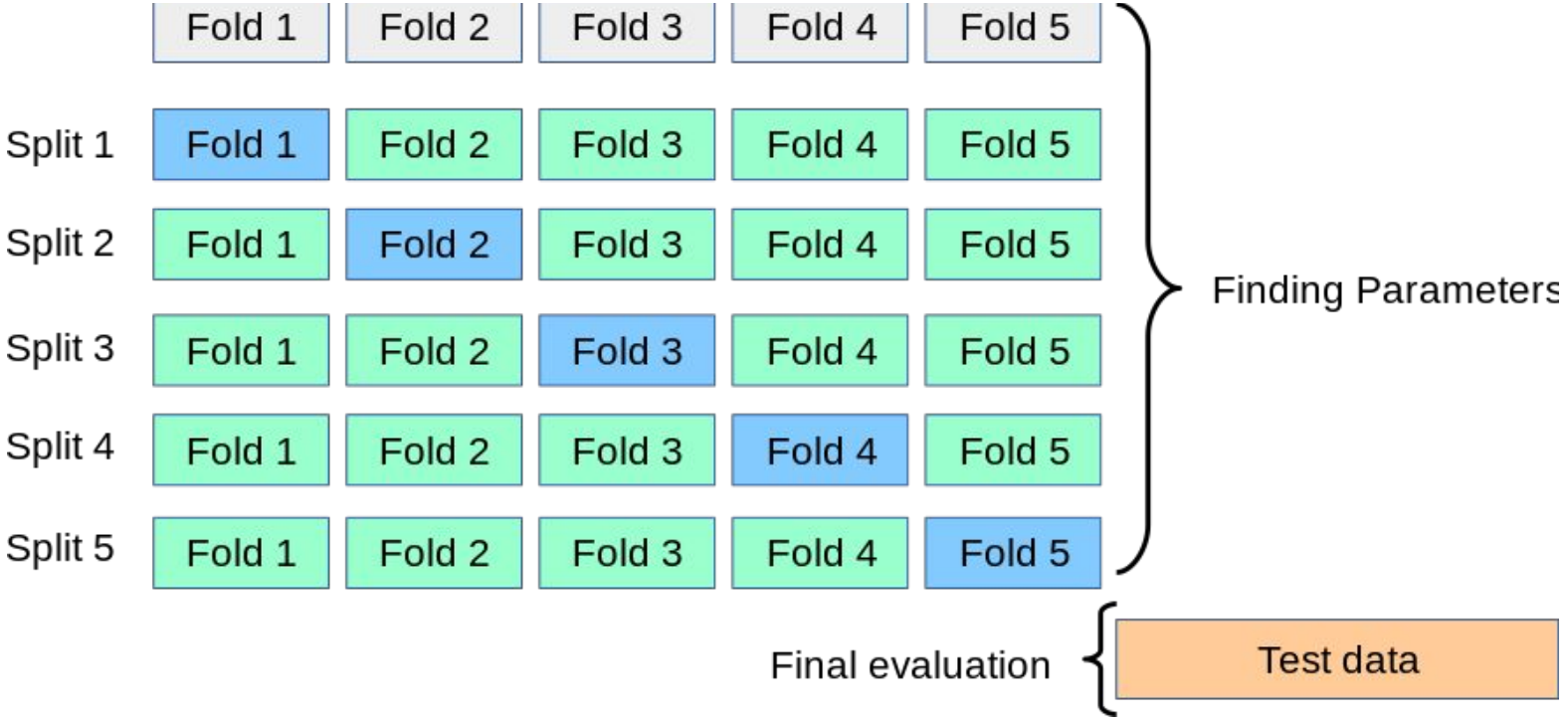
Precision: What percentage of items that were assigned label X do actually have label X in the test data?
Recall: What percentage of items that have label X in the test data were assigned label X by the system?
F-Measure: harmonic mean of precision and recall
$F = (2 \cdot P \cdot R) / (P + R)$

Confusion matrix c

Docs in Test set	Predicted c1	Predicted c2	Predicted c3
True c1	50	3	0
True c2	2	35	0
True c3	1	0	10

Data splits

Cross-validation



Scikitlearn

Assignment 2

Get the corpus from:
https://github.com/sebischair/Medical-Abstracts-TC-Corpus

Two models:

1. Baseline, tf-idf unigrams
2. Tf-idf unigrams, bigrams and trigrams

- Use scispacy for both
- Present precision, recall, f-measure for each class and for overall
- Present confusion matrices for test set for both models

References

- Oak, S. (2017, March). Depression detection and analysis. In *2017 AAAI spring symposium series*.
- Uddin, M. Z., Dysthe, K. K., Følstad, A., & Brandtzaeg, P. B. (2022). Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applications*, *34*(1), 721-744.
- Zhang, Y., Lin, H., Yang, Z., Wang, J., Sun, Y., Xu, B., & Zhao, Z. (2019). Neural network-based approaches for biomedical relation classification: a review. *Journal of biomedical informatics*, *99*, 103294.