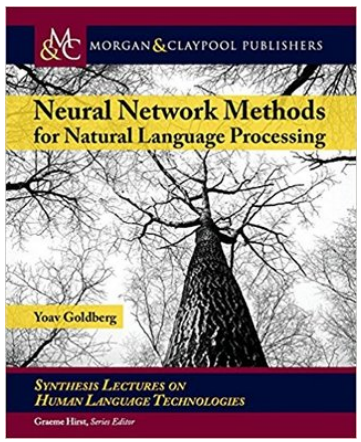
Natural Language Processing
Introduction

Felipe Bravo-Marquez

March 12, 2024

Disclaimer

- A significant part of the content presented in these slides is taken from other resources such as textbooks and publications.
- The neural network part of the course is heavily based on this book:



- Non-neural network topics, such as probabilistic language models, Naive Bayes, and HMM, are taken from Michael Collins' Columbia course [Collins, 2013].
- Also from the draft of the third edition of Dan Jurafsky and James H. Martin's book [Jurafsky and Martin, 2023].
- In addition, some slides have been adapted from online tutorials and other courses, such as Christopher Manning's Stanford course[1].

[1] ttp://web.stanford.edu/class/cs224n/

Natural Language Processing

- The amount of digitized textual data being generated every day is huge (e.g, the Web, social media, medical records, digitized books).
- So does the need for translating, analyzing, and managing this flood of words and text.
- Natural language processing (NLP) is the field of designing methods and algorithms that take as input or produce as output unstructured, **natural language data**. [Goldberg, 2017]
- Natural language processing is focused on the design and analysis of computational algorithms and representations for processing natural human language [Eisenstein, 2018]

Natural Language Processing

- Example of NLP task: Named Entity Recognition (NER):

PER						LOC	
Michael	Jordan	is	a	professor	at	Berkeley	.

Figure: Named Entity Recognition

- Example of NLP application: Chatbots like ChatGPT and Google Bard (or Gemini):

5 Challenging Properties of Humnan Language

1. **Ambiguity**: Human language is highly ambiguous.
   - For example: *I ate pizza with friends* vs. *I ate pizza with olives* vs. *I ate pizza with a fork*.
   - All these sentences have similar grammatical properties, but differ radically in what the prepositional phrase modifies (1) the proponoun “I”,2) the noun “pizza”,3) the verb “eat”).
2. **Dynamism**: Language is ever changing and evolving (e.g, Hashtags in Twitter).
3. **Discreteness**: we cannot infer the relation between two words from the letters they are made of (e.g., hamburger and pizza).
4. **Compositionality**: the meaning of a sentence goes beyond the individual meaning of their words.
5. **Sparseness**: The way in which words (discrete symbols) can be combined to form meanings is practically infinite.

Natural Language Processing and Computational Linguistics

Natural language processing (NLP) develops methods for solving practical problems involving language [Johnson, 2014].

- Automatic speech recognition.
- Machine translation.
- Information extraction from documents.

Computational linguistics (CL) studies the computational processes underlying (human) language.

- How do we understand language?
- How do we produce language?
- How do we learn language?

Similar methods and models are used in NLP and CL.

Natural Language Processing and Computational Linguistics

- Most of the meetings and journals that host natural language processing research bear the name “computational linguistics” (e.g., ACL, NACL). [Eisenstein, 2018]
- NLP and CL may be thought of as essentially synonymous.
- While there is substantial overlap, there is an important difference in focus.
- CL is essentially linguistics supported by computational methods (similar to computational biology, computational astronomy).
- In linguistics, language is the object of study.
- NLP focuses on solving well-defined tasks involving human language (e.g., translation, query answering, holding conversations).
- Fundamental linguistic insights may be crucial for accomplishing these tasks, but success is ultimately measured by whether and how well the job gets done (according to an evaluation metric) [Eisenstein, 2018].

Linguistics levels of description

The field of **linguistics** includes subfields that concern themselves with different levels or aspects of the structure of **language**, as well as subfields dedicated to studying how linguistic structure interacts with human cognition and society [Bender, 2013].

1. **Phonetics**: The study of the sounds of human language.
2. **Phonology**: The study of sound systems in human languages.
3. **Morphology**: The study of the formation and internal structure of words.
4. **Syntax**: The study of the formation and internal structure of sentences.
5. **Semantics**: The study of the meaning of sentences
6. **Pragmatics**: The study of the way sentences with their semantic meanings are used for particular communicative goals.

Phonetics

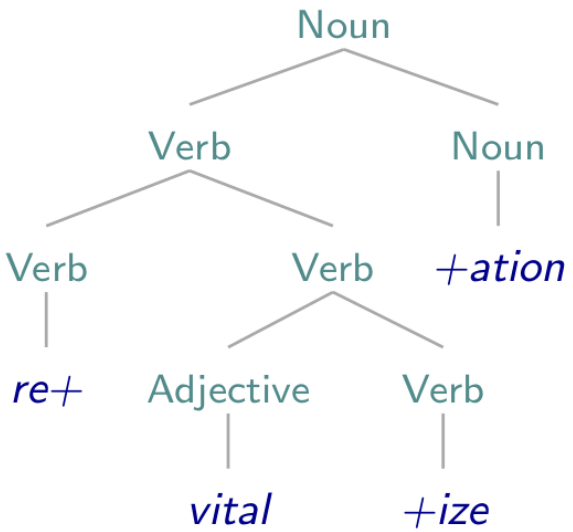
- Phonetics studies the sounds of a language [Johnson, 2014]
- It deals with the organs of sound production (e.g., mouth, tongue, throat, nose, lips, palate)
- Vowels vs consonants.
- Vowels are produced with little restriction of the airflow from the lungs out the mouth and/or the nose. [Fromkin et al., 2018]
- Consonants are produced with some restriction or closure in the vocal tract that impedes the flow of air from the lungs. [Fromkin et al., 2018]
- International Phonetic Alphabet (IPA): alphabetic system of phonetic notation.

Phonology

- Phonology: The study of how speech sounds form patterns [Fromkin et al., 2018].
- Phonemes are the basic form of a sound (e.g., the phoneme /p/)
- Example: Why **g** is silent in sign but is pronounced in the related word signature?
- Example: English speakers pronounce /t/ differently (e.g., in water)
- In Spanish /z/ is pronounced differently in Spain and Latin America.
- Phonetics vs Phonology: `http://www.phon.ox.ac.uk/jcoleman/PHONOLOGY1.htm`.

Morphology

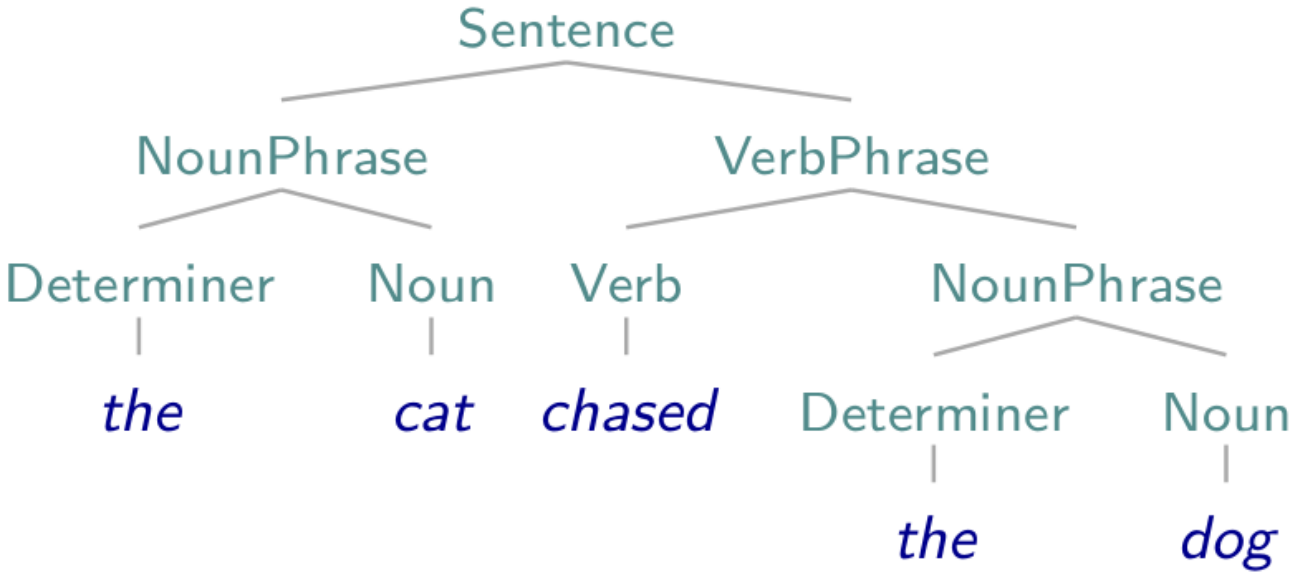
- Morphology studies the structure of words (e.g.,re+structur+ing, un+remark+able) [Johnson, 2014]
- Morpheme: The linguistic term for the most elemental unit of grammatical form [Fromkin et al., 2018]. Example morphology= morph + ology (the science of).
- Derivational morphology: process of forming a new word from an existing word, often by adding a prefix or suffix
- Derivational morphology exhibits a hierarchical structure. Example: re+vital+ize+ation



- The suffix usually determines the syntactic category (part-of-speech) of the derived word.

Syntax

- Syntax studies the ways words combine to form phrases and sentences [Johnson, 2014]



- Syntactic parsing helps identify **who did what to whom**, a key step in understanding a sentence.

Semantics

- Semantics studies the meaning of words, phrases and sentences [Johnson, 2014].
- Semantic roles: indicate the role played by each entity in a sentence.
- Examples of semantic roles: **agent** (the entity that performs the action), **theme** (the entity involved in the action), or **instrument** (another entity used by the agent in order to perform the action).
- Annotated sentence: **The boy** cut **the rope** with **a razor**.
- Lexical relations: relationship between different words [Yule, 2016].
- Examples of lexical relations: synonymy (conceal/hide), antonymy (shallow/deep) and hyponymy (dog/animal).

Pragmatics

- **Pragmatics**: the study of how context affects meaning in certain situations [Fromkin et al., 2018].
- Example: how the sentence “It’s cold in here” comes to be interpreted as “close the windows”.
- Example 2: Can you pass the salt?

Natural Language Processing and Machine Learning

- While we humans are great users of language, we are also very poor at formally understanding and describing the rules that govern language.
- Understanding and producing language using computers is highly challenging.
- The best known set of methods for dealing with language data rely on supervised machine learning.
- Supervised machine learning: attempt to infer usage patterns and regularities from a set of pre-annotated input and output pairs (a.k.a training dataset).

Training Dataset: CoNLL-2003 NER Data

Each line contains a token, a part-of-speech tag, a syntactic chunk tag, and a named-entity tag.

```
U.N.          NNP   I-NP   I-ORG
official      NN    I-NP   O
Ekeus         NNP   I-NP   I-PER
heads         VBZ   I-VP   O
for           IN    I-PP   O
Baghdad       NNP   I-NP   I-LOC
.             .     O      O
```

[2]Source:
https://www.clips.uantwerpen.be/conll2003/ner/

Example of NLP Task: Topic Classification

- Classify a document into one of four categories: Sports, Politics, Gossip, and Economy.
- The words in the documents provide very strong hints.
- Which words provide what hints?
- Writing up rules for this task is rather challenging.
- However, readers can easily categorize a number of documents into its topic (data annotation).
- A supervised machine learning algorithm come up with the patterns of word usage that help categorize the documents.

Example 3: Sentiment Analysis

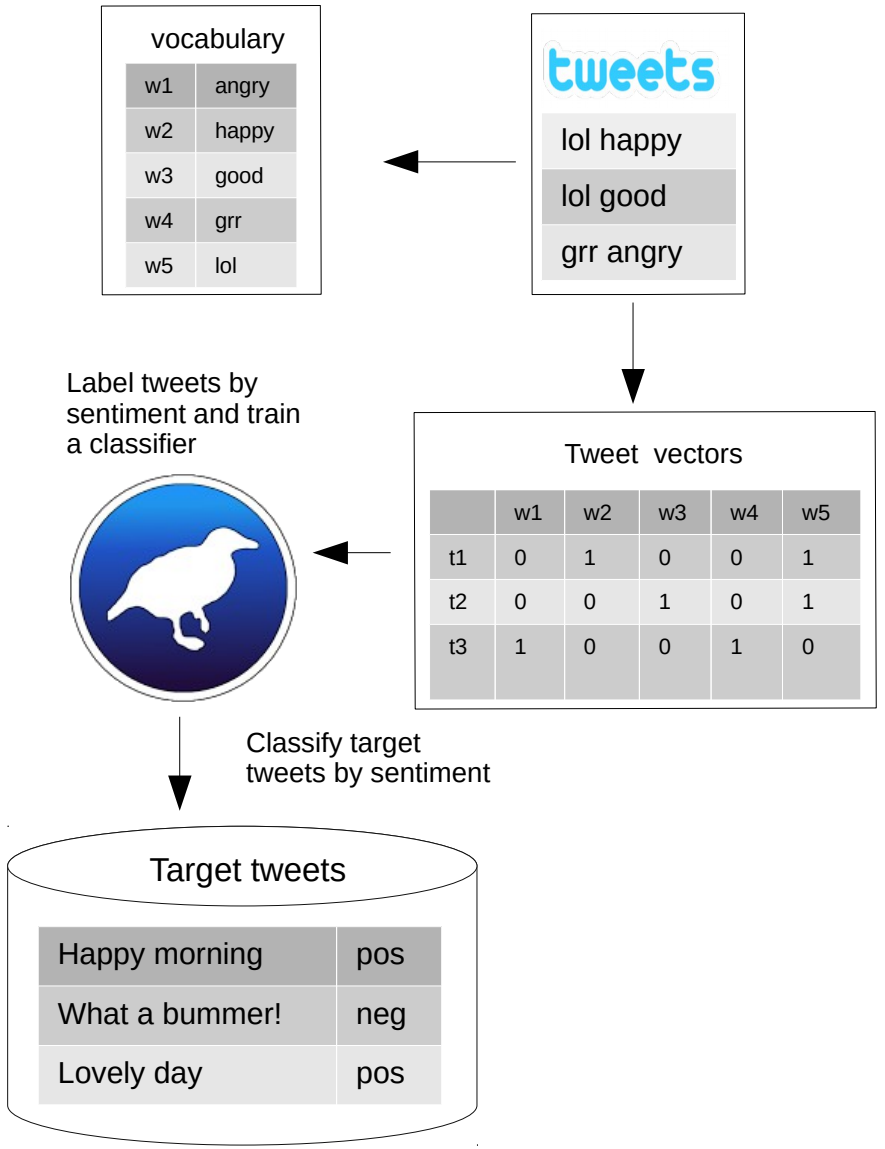
- Application of **NLP** techniques to identify and extract subjective information from textual datasets.

Main Problem: Message-level Polarity Classification (MPC)

1. Automatically classify a sentence to classes **positive**, **negative**, or **neutral**.

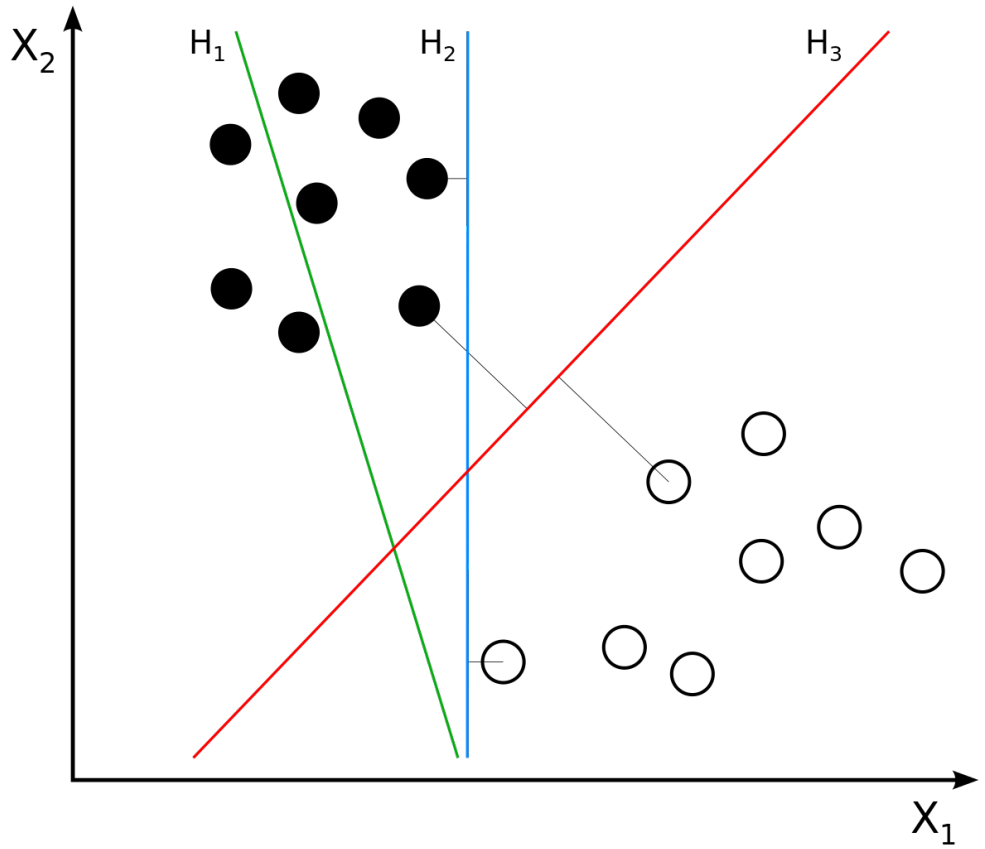
2. State-of-the-art solutions use **supervised** machine learning models trained from **manually** annotated examples [Mohammad et al., 2013].

Sentiment Classification via Supervised Learning and BoWs Vectors



Supervised Learning: Support Vector Machines (SVMs)

- Idea: Find a hyperplane that separates the classes with the maximum margin (largest separation).



- H_3 separates the classes with the maximum margin.

[3] Image source: Wikipedia

Linguistics and NLP

- Knowing about linguistic structure is important for feature design and error analysis in NLP [Bender, 2013].
- Machine learning approaches to NLP require features which can describe and generalize across particular instances of language use.
- Goal: guide the machine learning algorithm to find correlations between language use and its target set of labels.
- Knowledge about linguistic structures can inform the design of features for machine learning approaches to NLP.

Challenges in NLP

- **Annotation Costs**: manual annotation is **labour-intensive** and **time-consuming**.
- **Domain Variations**: the pattern we want to learn can vary from one corpus to another (e.g., sports, politics).
- A model trained from data annotated for one domain will **not necessarily** work on another one!
- Trained models can become outdated over time (e.g., new hashtags).

Domain Variation in Sentiment

1. For me the queue was pretty **small** and it was only a 20 minute wait I think but was so worth it!!! :D @raynwise
2. Odd spatiality in Stuttgart. Hotel room is so **small** I can barely turn around but surroundings are inhumanly vast & long under construction.

Overcoming the data annotation costs

Distant Supervision

- Automatically **label** unlabeled data (**Twitter API**) using a heuristic method.
- **Emoticon-Annotation Approach (EAA)**: tweets with positive **:)** or negative **:(** emoticons are labelled according to the polarity indicated by the emoticon [Read, 2005].
- The emoticon is **removed** from the content.
- The same approach has been extended using hashtags #anger, and emojis.
- Is not trivial to find distant supervision techniques for all kind of NLP problems.

Crowdsourcing

- Rely on services like **Amazon Mechanical Turk** or **Crowdflower** to ask the **crowds** to annotate data.
- This can be expensive.
- It is hard to guarantee quality.

Sentiment Classification of Tweets

- In 2013, The Semantic Evaluation (SemEval) workshop organized the "Sentiment Analysis in Twitter task" [Nakov et al., 2013].
- The task was divided into two sub-tasks: the expression level and the message level.
- Expression-level: focused on determining the sentiment polarity of a message according to a marked entity within its content.
- Message-level: the polarity has to be determined according to the overall message.
- The organizers released training and testing datasets for both tasks. [Nakov et al., 2013]

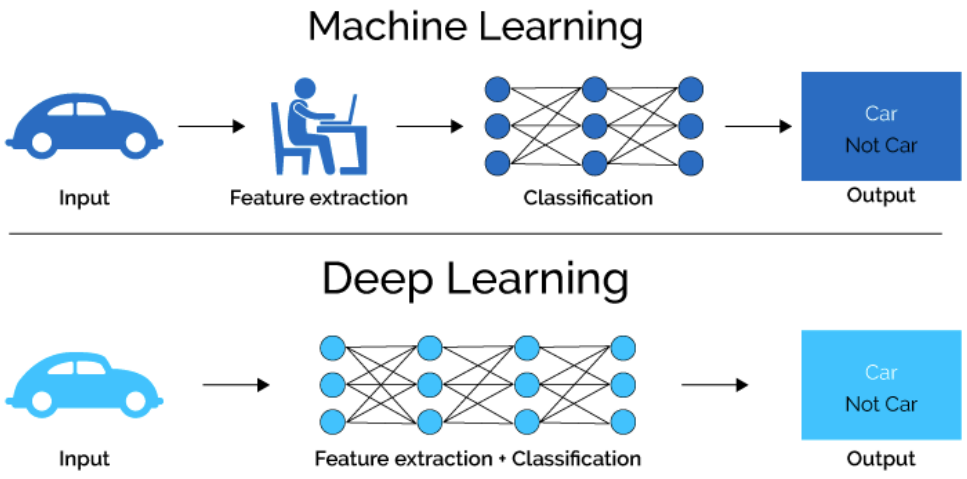
The NRC System

- The team that achieved the highest performance in both tasks among 44 teams was the *NRC-Canada* team [Mohammad et al., 2013].
- The team proposed a supervised approach using a linear SVM classifier with the following hand-crafted features for representing tweets:
  1. Word *n*-grams.
  2. Character *n*-grams.
  3. Part-of-speech tags.
  4. Word clusters trained with the Brown clustering method [Brown et al., 1992].
  5. The number of elongated words (words with one character repeated more than two times).
  6. The number of words with all characters in uppercase.
  7. The presence of positive or negative emoticons.
  8. The number of individual negations.
  9. The number of contiguous sequences of dots, question marks and exclamation marks.
  10. Features derived from polarity lexicons [Mohammad et al., 2013]. Two of these lexicons were generated using the PMI method from tweets annotated with hashtags and emoticons.

Feature Engineering and Deep Learning

- Up until 2014 most state-of-the-art NLP systems were based on feature engineering + shallow machine learning models (e.g., SVMs, HMMs).
- Designing the features of a winning NLP system requires a lot of domain-specific knowledge.
- The NRC system was built before deep learning became popular in NLP.
- Deep Learning systems on the other hand rely on neural networks to automatically learn good representations.

Feature Engineering and Deep Learning



- Deep Learning yields state-of-the-art results in most NLP tasks.
- Large amounts of training data and faster multicore GPU machines are key in the success of deep learning.
- **Neural networks** and **word embeddings** play a key role in modern NLP models.

Deep Learning and Linguistic Concepts

- If deep learning models can learn representations automatically, are linguistic concepts still useful (e.g., syntax, morphology)?
- Some proponents of deep-learning argue that such inferred, manually designed, linguistic properties are not needed, and that the neural network will learn these intermediate representations (or equivalent, or better ones) on its own [Goldberg, 2016].
- The jury is still out on this.
- Goldberg believes many of these linguistic concepts can indeed be inferred by the network on its own if given enough data.
- However, for many other cases we do not have enough training data available for the task we care about, and in these cases providing the network with the more explicit general concepts can be very valuable.

History

NLP progress can be divided into three main waves: 1) rationalism, 2) empiricism, and 3) deep learning [Deng and Liu, 2018].

1950 - 1990 Rationalism: approaches endeavored to design hand-crafted rules to incorporate knowledge and reasoning mechanisms into intelligent NLP systems (e.g, ELIZA for simulating a Rogerian psychotherapist, MARGIE for structuring real-world information into concept ontologies).

1991 - 2009 Empiricism: characterized by the exploitation of data corpora and of (shallow) machine learning and statistical models (e.g., Naive Bayes, HMMs, IBM translation models).

2010 - Deep Learning: feature engineering (considered as a bottleneck) is replaced with representation learning and/or deep neural networks (e.g., `https://www.deepl.com/translator`). A very influential paper in this revolution: [Collobert et al., 2011].

[3] Dates are approximated.

A fourth wave??

- Large Language Models (LLMs) like ChatGPT, GPT4, Llama and Bard are deep neural networks trained on large corpora (hundreds of billions of tokens) and a large parameter space (billions) to predict the next word from a fixed-size context.
- One of the most striking features of these models is their ability for few-shot, one-shot, and zero-shot learning, often referred to as “in-context learning”.
- This implies their capacity to acquire new tasks with minimal human-annotated data, simply by providing the appropriate instruction or prompt.
- Thus, despite being rooted in the deep learning paradigm, they introduce a disruptive approach to NLP.

Roadmap

In this course we will introduce modern concepts in natural language processing based on **statistical models** (second wave) and **neural networks** (third wave). The main concepts to be covered are listed below:

1. Text classification.
2. Linear Models.
3. Naive Bayes.
4. Hidden Markov Models.
5. Neural Networks.
6. Word embeddings.
7. Convolutional Neural Networks (CNNs)
8. Recurrent Neural Networks: Elman, LSTMs, GRUs.
9. Attention.
10. Sequence-to-Sequence Models.
11. Transformer
12. Large Language Models.

Questions?

Thanks for your Attention!

References I


Bender, E. M. (2013).
Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax.
Synthesis lectures on human language technologies, 6(3):1–184.


Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992).
Class-based n-gram models of natural language.
Computational linguistics, 18(4):467–479.


Collins, M. (2013).
Statistical nlp lecture notes.
Lecture Notes.


Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011).
Natural language processing (almost) from scratch.
Journal of machine learning research, 12(Aug):2493–2537.


Deng, L. and Liu, Y. (2018).
Deep Learning in Natural Language Processing.
Springer.


Eisenstein, J. (2018).
Natural language processing.
Technical report, Georgia Tech.

References II


Fromkin, V., Rodman, R., and Hyams, N. (2018).
An introduction to language.
Cengage Learning.


Goldberg, Y. (2016).
A primer on neural network models for natural language processing.
J. Artif. Intell. Res.(JAIR), 57:345–420.


Goldberg, Y. (2017).
Neural network methods for natural language processing.
Synthesis Lectures on Human Language Technologies, 10(1):1–309.


Johnson, M. (2014).
Introduction to computational linguistics and natural language processing (slides).

2014 Machine Learning Summer School.


Jurafsky, D. and Martin, J. H. (2023).
Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.
Prentice Hall, Upper Saddle River, NJ, USA, 3rd (draft) edition.

References III

Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013).
Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets.
Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013).

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013).
Semeval-2013 task 2: Sentiment analysis in twitter.
In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.

Read, J. (2005).
Using emoticons to reduce dependency in machine learning techniques for sentiment classification.
In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yule, G. (2016).
The study of language.
Cambridge university press.