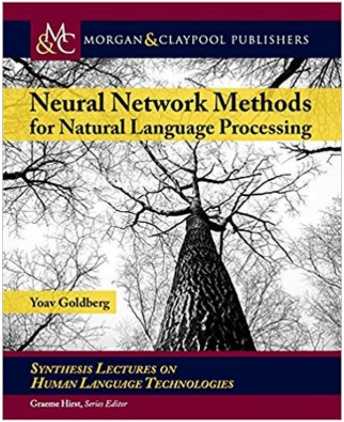
Natural Language Processing
Introduction

Felipe Bravo-Marquez

July 1, 2019

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:



Natural Language Processing

- The amount of digitalized textual data being generated every day is huge (e.g, the Web, social media, medicar records, digitalized 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**. [**?**]

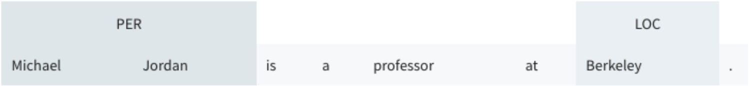


Figure: Example: Named Entity Recognition

- Human language is highly ambiguous: *I ate pizza with friends* vs. *I ate pizza with olives*.
- It is also ever changing and evolving (e.g, Hashtags in Twitter).

Natural Language Processing and Computational Linguistics

Natural language processing (NLP) develops methods for solving practical problems involving language [**?**].

- 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?
- H‘ow do we learn language?

Similiar methods and models are used in NLP and CL

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 [**?**].

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 [**?**]
- It deals with the organs of sound production (e.g., mouth, tongue, throat, nose, lips, palate)
- Example: English aspirates stop consonants in certain positions (e.g., [t^hop] vs. [stop])
- International Phonetic Alphabet (IPA): alphabetic system of phonetic notation.

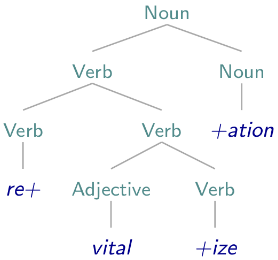
Phonology

- Phonology: The study of how speech sounds form patterns [**?**]
- 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)

Morphology

Morphology studies the structure of words [**?**]

- E.g.,re+structur+ing, un+remark+able
- 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 of the derived word

Natural Language Processing

- 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
```

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

Challenges of Language

- Three challenging properties of language: discreteness , compositionality, and sparseness.
- **Discreteness**: we cannot infer the relation between two words from the letters they are made of (e.g., hamburger and pizza).
- **Compositionality**: the meaning of a sentence goes beyond the individual meaning of their words.
- **Sparseness**: The way in which words (discrete symbols) can be combined to form meanings is practically infinite.

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

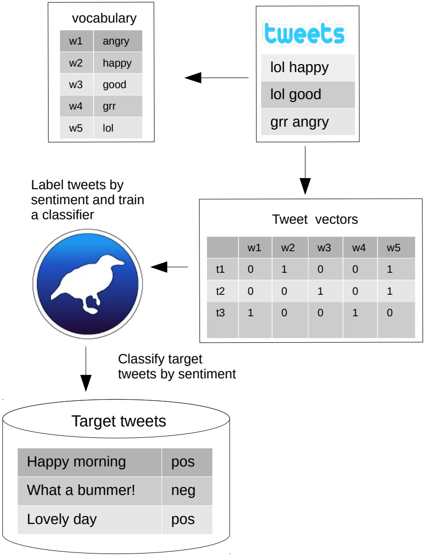
- Application of **NLP** techniques to identify and extract subjective information from textual datasets.

Main Problem: Message-level Polarity Classification (MPC)

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

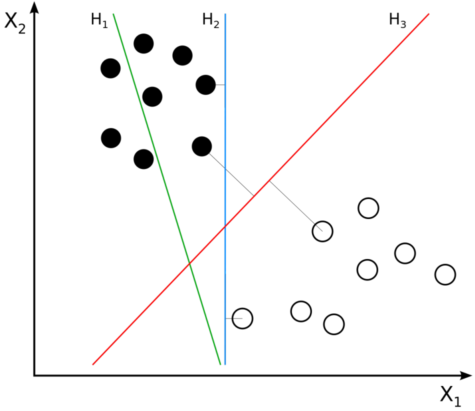
2. State-of-the-art solutions use **supervised** machine learning models trained from **manually** annotated examples [Mohammad et al., 2013].

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.

[2] Image source: Wikipedia

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 organised 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 organisers 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

- 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 representation learning to automatically learn good representations.
- Large amounts of training data and faster multicore CPU/GPU machines are key in the success of deep learning.
- **Neural networks** and **word embeddings** play a key role in modern architectures for NLP.

Roadmap

In this course we will introduce modern concepts in natural language processing based on **neural networks.** The main concepts to be convered are listed below:

1. **Word embeddings**
2. **Convolutional Neural Networks** (CNNs)
3. **Recurrent Neural Networks**: Elman, LSTMs, GRUs.

Questions?

Thanks for your Attention!

References I

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.

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.