FNLP - Week 1: Intro to NLP & Corpora

Antonio León Villares

January 2022

Contents

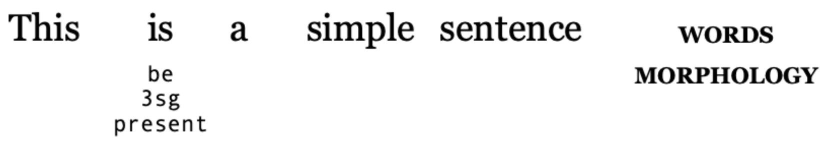
1 Introduction to NLP & Linguistics

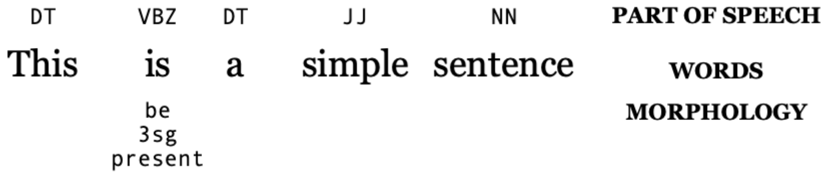
- **What is NLP?**
  - the process by which computers "learn" to process the human language
- **What is NLP used in?**
  - machine translation
  - automatic speech recognition
  - natural language generation (speech synthesis)
  - web-based question answering
- **Which technologies are at the heart of NLP?**
  - language modelling
  - POS (part of speech) tagging
  - syntactic parsing
  - semantic role labelling
- **What knowledge scope of language are important to understand language?**
  - *phonetics*: how are words **pronounced** (as a sequence of sounds)
  - *phonology*: how is each sound realised acoustically
  - *morphology*: how are words **formed**, how do they **relate** to other words in the same language
    - i.e realising that "I'm" and "I am" are the same
    - i.e realising that "doors" is the plural of "door"
  - *syntax*: how are words pieced together **coherently**
    - i.e understanding that "I like red apples" is well-formed, but "apples like red I" isn't
  - *semantics*: what type of **meaning** can we extract. Two types of semantics:
    - *lexical semantics*: the meaning of words
    - *compositional/logical semantics*: the meaning of a phrase, derived from the meaning of its constituents (i.e understanding the meaning of "end of the 18th century" as the set of temporal events that occurred during the latter years of the 1700s)
  - *pragmatic/dialogue knowledge*: how users convey intention through words
    - i.e understanding that "I won't" is more truthful and direct than "I am sorry, but I don't think I can"
  - *discourse knowledge*: how to interpret words based on previous knowledge
    - i.e if I ask "What year was Lincoln born in?" and then ask "Where was **he** born?", we want to understand that the **he** refers to the person called "Lincoln"
- **What is an utterance?**
  - an uninterrupted chain of spoken or written language
- **What is a part of speech?**
  - the building blocks used to classify the type of a word, based on how the word functions
  - in English:

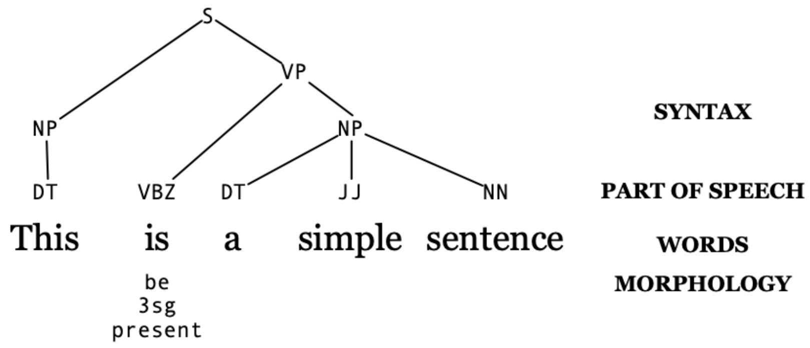
- *noun*: a word for a person, place, thing, or idea (i.e man, Edinburgh, dream)
- *pronoun*: a word used in place of a noun (i.e she, we, my)
- *verb*: a word expressing an action or being. take different **tenses** (i.e to run, brought, demanding)
- *adjective*: word used to modify or describe a noun or a pronoun (i.e beautiful, horrendous, perplexing)
- *adverb*: word used to describe or modify a verb, an adjective, or another adverb, but **never** a noun (i.e greatly, very, then)
- *preposition*: word placed before a noun or pronoun to form a phrase modifying another word in the sentence (i.e by, from, with)
- *conjunction*: word used to join words, phrases, or clauses, and indicate the relationship between the elements joined (i.e and, while, because)
- *interjection*: word used to express emotion (i.e oh, wow, ah)

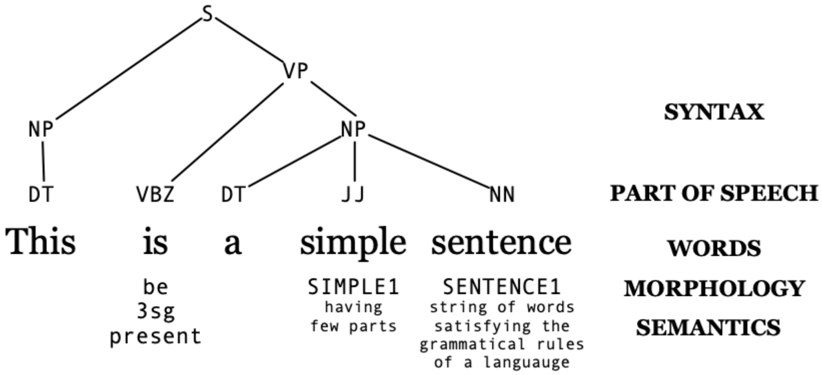
With all of the above, there are many layers to understanding simple phrases, like:

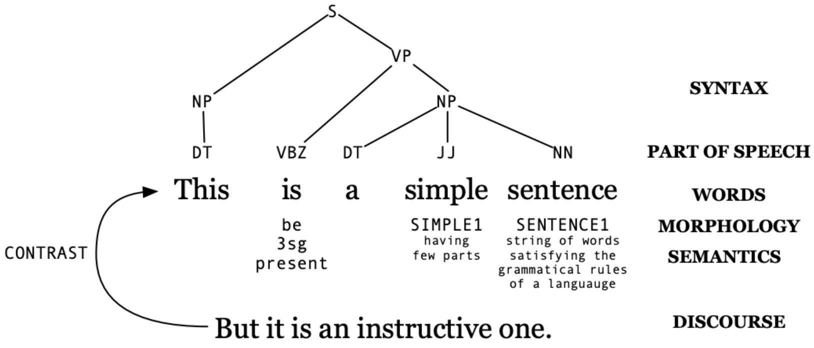
"This is a simple sentence"











2 The Challenges of NLP

There are many challenges to NLP, beyond being able to understand all the layers to a sentence.

2.1 Variability

- **Why is variability a challenge in NLP?**
  - there are many ways of saying the same thing
    - “He drew the house” vs “He made a sketch of the house”
    - “Some kids popped by” vs “A few children visited”

2.2 Ambiguity

- **Why is ambiguity a challenge in NLP?**
  - language, to be efficient at communicating ideas, is naturally **ambiguous**
  - this means that the same phrase can be interpreted in many different ways (in fact, combinatorially many - Catalan numbers)
    - “I made her duck”
      - “I cooked an aquatic bird for her”
      - “After throwing something at her, she bent her knees to avoid the object”
      - “I cooked an aquatic bird that is hers”
- **What types of ambiguities can we consider?**

  - *homophones*: words which sound the same
    - blue vs blew
  - *word senses*: words which can be used in many different ways
    - bank (a side of the river, or a financial institution)
  - *part of speech*: words which can be used as different parts of speech
    - duck (noun or verb)
  - *syntactic structure*: ambiguity derived from the sentence structure
    - "I saw her with a telescope"
    - 1) Using a telescope, I saw her
    - 2) I saw her using/carrying a telescope
  - *quantifier scope*: ambiguities derived from using different scopes in a sentence
    - "Every child loves some movie"
    - 1) Each child has at least one move which they love
    - 2) There is one movie which is universally loved by children
  - *reference*: when a pronoun in a sentence could refer to 2 different nouns. Solving this ambiguity is known as **co-reference resolution**
    - "John dropped the goblet onto the glass table and it broke"
    - did the table break? Or was it the goblet?
  - *discourse*:
    - "The meeting is cancelled. Nicholas is not coming."
    - 1) Since Nicholas is not coming, the meeting was cancelled
    - 2) Since the meeting is cancelled, Nicholas is not coming.

2.3 Data Sparsity: Zipf's Law

- **Why is sparsity a challenge in NLP?**
  - data is **sparse** in general
  - there are many elements which exist, but a model won't see in training
  - need to account for these missing words in some way for a model to be any good
- **Can't we just pick some data which has enough of everything?**
  - *Zipf's Law* says that:

    "the rank-frequency distribution of data is inversely related"

  - that is, if we have data, and we rank it based on frequency, the following relation holds:

$$f \times r = k$$

    - f: the frequency of a label
    - r: the rank of the label
    - k: a constant
  - in particular, Zipf's Law applies to text data, so independently of the amount of data, we will always have highly **infrequent** or **0-frequency** words

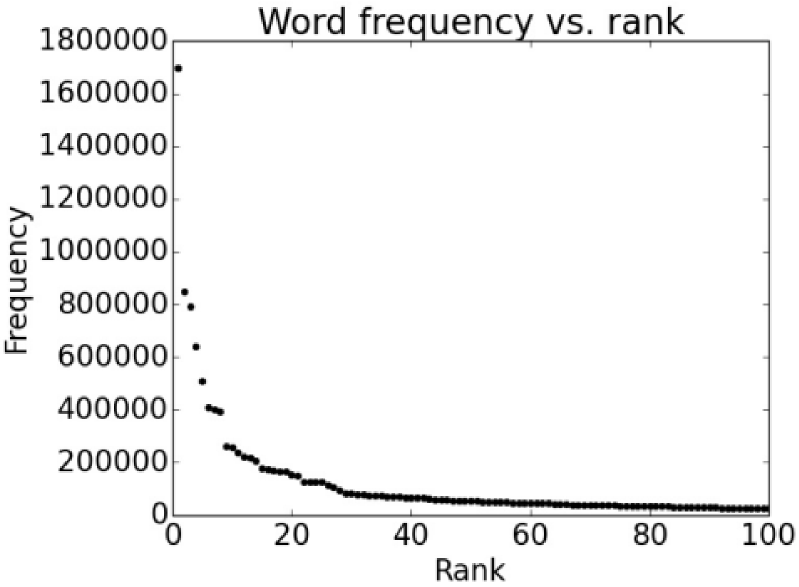


Figure 1: Words like “the” or “and” can have more than 100k in a single dataset, whilst words like “mathematician” or “cornflakes” appear once, or not at all.

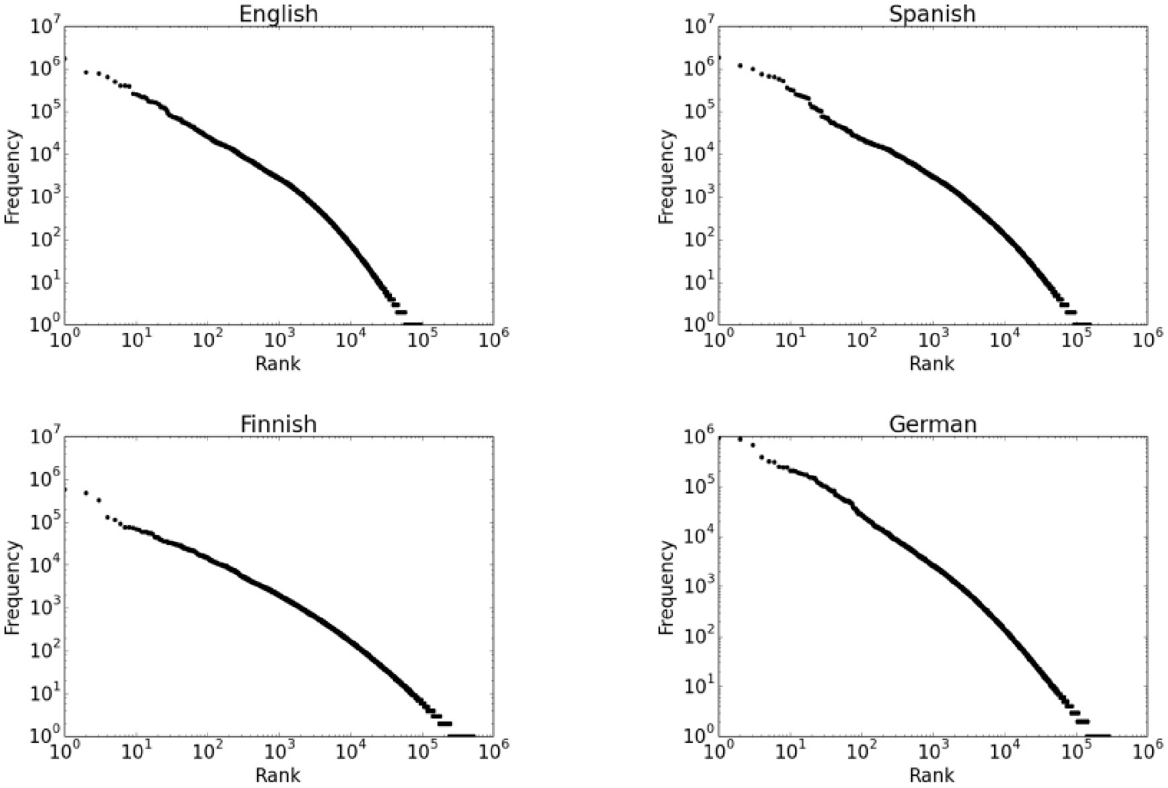


Figure 2: Zipf’s Law across many languages. Note the log scales.

2.4 Model Robustness

- **Why is robustness a challenge in NLP?**
    - NLP systems are typically trained and evaluated in "clean" settings, over data without significant noise.
    - a general-purpose, real world model needs to be able to deal with this
    - also needs to account with the fact that not everyone agrees that the same phrase is grammatically correct
    - for example, consider a POS-tagger:
        - trained on clean, proof-read WSJ financial articles ("The chairman, Mrs. Jane Doe, thanked stakeholders for their commitment")
        - applied to tweets, prone to spelling mistakes and abbreviations ("ikr smh fam, i h8 yo ass")

2.5 Context Dependence

- **Why is context dependence a challenge in NLP?**
    - context oftentimes indicates the correct interpretation of a sentence
    - having "world knowledge" is important

2.6 Unknown Representation

- **Why are unknown representations a challenge in NLP?**
    - to have a "world knowledge" we need to be able to **represent** this knowledge
    - need to model meaning, context, general notions, etc ...

2.7 Language Diversity

- **Why is language diversity a challenge in NLP?**
    - we focus on English, but the structure of language, and the meaning of words vary across languages
    - for example:
        - word order used (i.e subject-verb-object order)
        - cases (accusative, dative, reflexive, etc ...)
        - languages can be more flexible, in terms of what a well-fromed phrase is (i.e in russian, the same phrase can be said by changing the position of certain words)

3 NLP Models

- **What is the main challenge that an NLP model needs to handle?**
    - challenges in NLP tend to be related to **uncertainty**
        - *ambiguity*: interpretation uncertainty
        - *variability*: semantic construction uncertainty
        - *robustness*: input uncertainty
        - *sparsity/context dependence*: plainly, uncertainty

  - to deal with **uncertainty**, ML harnesses the power of **probability**

- **Can we deal with uncertainty non-probabilistically?**
  - **FSMs** for morphology
  - **CKY** for syntax parsing
  - however, since non-probabilistic, it provides **all** interpretations

- **What probabilistic methods can be used to manage uncertainty?**
  - models, such as **Hidden Markov Models** (POS tagging) and **Probabilistic Context Free Grammars** (syntax)
  - algorithms, such as **Viterbi** or **Probabilistic CKY**
  - these methods return the **most probable interpretation**, according to the model

- **Why use statistical NLP?**
  1. *Robustness* (not hindered by tons of rules)
  2. *Scalable*
  3. *Data-Driven* (i.e relevant statistics/probabilities)

- **What is a big issue with statistical NLP?**
  - need a lot of data to produce a representative sample
  - it is computationally expensive to select the most likely output

- **What does a probabilistic model look like?**
  - define **input-output** pairs, based on a **formal grammar** (i.e set of rules)
    - for example, for POS-tagging, we can use $(the, \underline{list})$ as input, and *noun* as output (we predict "list" given a context word "the")
  - determine the **probabilities** of output, given an input
  - use **corpora** to obtain these **statistics**
  - select an efficient **algorithm**
  - measure success of model using **evaluation methods**

4 Text Corpora

- **What is a corpus?**
  - a body of utterances, as words or sentences, assumed to be representative of and used for lexical, grammatical, or other linguistic analysis

- **How are corpora used in NLP?**
  - provide a **realistic sample** of language
  - need to be **naturally occurring**
  - include **metadata** (i.e author, publishing date, topic, etc ...)
  - provide an **empirically grounded** approach to learn

- **What are linguistic annotations?**

    - annotations of corpora made by humans
    - include info on categories/syntax/meaning
    - can sometimes be derived directly from data (i.e reviews can be annotated with the rating given)

- **What is a key issue with linguistic annotations?**
    - inconsistency in human annotators
    - can use a “big rule book” to mitigate

4.1 Motivation: Sentiment Analysis

- **What is sentiment analysis?**
    - the task of identifying the overall sentiment of a piece of text
    - sentiment can be given as binary (positive or negative) or as a rating system (i.e 3/5 stars)

- **What do we have to consider when constructing a sentiment analyser?**
    1. Type of input? (sentence, full review or text + metadata)
    2. Type of output? (binary or rating)
    3. How to decide on output?
    4. How to evaluate the decisions?

  Notice, at the very least, the evaluation step requires **data**

- **Why is evaluation data-driven?**
    1. Ensures **controlled experimentation**
    2. Use as **benchmarks** (i.e compare current system to previous system)
    3. Ensures we detach from our **intuitions** (we can’t analyse every single aspect of a task)

- **How many datasets do we use?**
    - typically 3:
        * Training (80%)
        * Development (10%)
        * Testing (10%)
    - this ensures that we don’t simply perform **data analysis**
    - allows us to **estimate** model performance on future data

- **What are *gold labels*?**
    - labels used to determine how **effective** the sentiment analyser is
    - tell us the **sentiment** of a text (either from **metadata** or from **human annotators**)
    - if human labels, need to ensure consistency (i.e similar pieces of text get similar annotations)

- **Can we use word counting as a measure for sentiment?**
    - strategy: count +ve and -ve words in the text. Predict whichever is greater.
    - this has 3 issues:
        * text is not a **bag of words**: order matters. For example, **semantic modifiers**:

            - “not good” would be perceived as neutral, since it has 1 +ve and 1 -ve word
        - doesn’t account for the sense conveyed by words:
            - *sarcasm* (“totally great”)
            - *sense ambiguity* (“fantastically horrible”)
        - doesn’t account for:
            - *descriptions* (“X acts as a horrible person, which is great”)
            - *expectations* (“I was looking forward to this, it was meant to be great, but...”)

- **What can we use to measure model performance?**
    - *accuracy*: proportion of correct predictions, out of all predictions

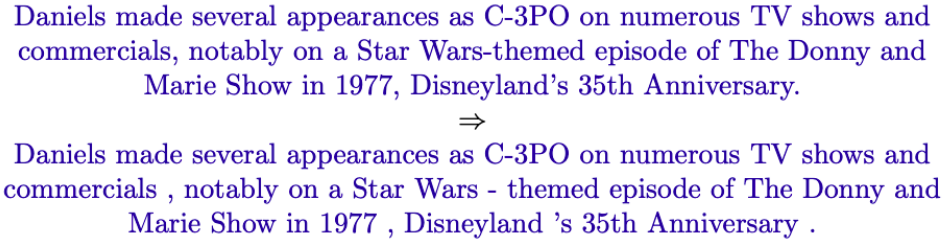
4.2 Preparing Corpora

To prepare corpora, many preprocessing steps can be taken, including:

- (word) tokenisation
- encoding conversion
- removal of markup (i.e italics)
- insertion of markup
- case conversion (i.e capitalising the first letter)
- sentence boundary detection (sentence tokenisation)

4.2.1 Tokenisation

- **What is tokenisation?**
    - the process of adding **logical boundaries** between separate word/punctuation tokens (occurrences) not already separated by spaces
    - this can sometimes be **automatised** using rules



Daniels made several appearances as C-3PO on numerous TV shows and commercials, notably on a Star Wars-themed episode of The Donny and Marie Show in 1977, Disneyland’s 35th Anniversary.

⇒

Daniels made several appearances as C-3PO on numerous TV shows and commercials , notably on a Star Wars - themed episode of The Donny and Marie Show in 1977 , Disneyland ’s 35th Anniversary .

- **What are some tokenisation conventions in English?**
    - these can vary
    - *clitics* (contracted forms, like “’s” or “n’t”) are separated
    - *hyphens* in compound words are separated

4.3 Domain Adaptation

- **What is domain in NLP?**
  - the **contextual factors** which affect how language is expressed
- **What are examples of domains?**
  - *mode of communication*:
    - speech (i.e videos, telephone, in-person)
    - text (i.e email, SMS, websites)
  - *topic*
    - small talk
    - politics
    - science
  - *genre*
    - tweet
    - political address
    - novel
  - *audience*
    - formality
    - audience
    - complexity
- **Why is it important that training and testing data is distributed similarly?**
  - most ML methods expect this
  - sampling from testing and training should give similar results
- **What differences can there be between test and training data ?**
  - vocabulary
  - label distribution
  - style (colloquial, serious)
  - sentence length
- **What is domain adaptation?**
  - the ability to apply an algorithm trained in one or more "source domains" to a different (but related) "target domain"
  - if the test data is **different**, **domain adaptation** tries to correct the assumption of similar test/training data distribution