MORPHOLOGY

Luísa Coheur

TÉCNICO LISBOA

inesc id lisboa

Overview

- Learning objectives
- Topics
  - Linguistic Knowledge
  - Morphology
    - Word Classes
    - Part-of-speech tagging
    - Morphology Main Concepts
- Key takeaways
- Suggested readings

LEARNING OBJECTIVES

LEARNING OBJECTIVES

- Grasp fundamental concepts
- Learn several ways to perform Part-of-Speech (PoS) tagging

TOPICS

Overview

- Learning objectives
- Topics
  - Linguistic Knowledge
  - Morphology
    - Word Classes
    - Part-of-speech tagging
    - Morphology Main Concepts
- Key takeaways
- Suggested readings

LINGUISTIC KNOWLEDGE

- Phonetic knowledge: relates words to sounds
  - Example: meme. How to pronounce it? ;-)
  - (in Portuguese: Eu almoço o almoço. How to pronounce "almoço")
- Morphological knowledge: related to the study of the constituents of words
  - Example: if "almoço" is tagged as a verb, you will already know how to pronounce it...

This class is going to be dedicated to Morphology

LINGUISTIC KNOWLEDGE

- Syntactic: determines how words can be combined to form a sentence
- Semantic Knowledge: used to assign a meaning to each word and to a sentence (literal meaning)
  - Before: Logic
  - Now: vectors (and neural embeddings)
  - Problem: how to represent NL? How to map NL in this representation?

Next week we will have a class dedicated to Syntax and another one to Semantics

LINGUISTIC KNOWLEDGE

- Pragmatic Knowledge: takes into account the context in the interpretation of a sentence (non-literal meaning).
  - It is so dark! (maybe I want you to turn on the light)

No class dedicated to Pragmatic Knowledge

LINGUISTIC KNOWLEDGE

- Discourse knowledge: used to determine the influence of the preceding sentences on the interpretation of the current sentence (e.g., pronouns and temporal information)
  - John loves his sister. She is so nice.

No Discourse Knowledge

WORLD KNOWLEDGE (common sense)

- John was shot in the eye, and his brain came out of his ear
- => he died

No World Knowledge (but check Open Mind Common Sense… or ChatGPT)

Overview

- Learning objectives
- Topics
  - Linguistic Knowledge
  - Morphology
    - Word Classes
    - Part-of-speech tagging
    - Morphology Main Concepts
- Key takeaways
- Suggested readings

WORD CLASSES (or PART-OF-SPEECH)

- A Word Class is a category into which words are grouped considering their function within a sentence

WORD CLASSES

- Nouns: name people, places, things, ideas, or concepts
- Pronouns: take the place of nouns
- Verbs: express actions, states, or occurrences
  - Action verb: She writes a letter.
  - State verb: He knows the answer. She is tired.
  - Occurrence verb: The sun rose early today.
- Adjectives: describe or modify nouns or pronouns
- Adverbs: modify verbs, adjectives, or other adverbs

WORD CLASSES (cont.)

- Prepositions: typically indicating location, direction, time, or manner
  - Examples:
    - Location: The book is on the table.
    - Direction: He walked to the park.
    - Time: We will meet at 5 PM.
    - Manner: She spoke with confidence.
- Conjunctions: join words, phrases, clauses, or sentences
- Interjections: express emotions

Remember?

ACTIVE LEARNING MOMENT

EXERCISE

- Consider the sentence (by ChatGPT):

  "Wow", Sarah carefully hands her friend the red book from the shelf, and smiles.

  - Find:
    - Adjective(s):
    - Adverb(s):
    - Conjunction(s):
    - Common noun(s):
    - Determiner(s):
    - Interjection(s):
    - Preposition(s):
    - Proper noun(s):
    - Pronoun(s):
    - Verb(s):

EXERCISE

- Consider the sentence (by ChatGPT):

"Wow", Sarah carefully hands her friend the red book from the shelf, and smiles.

  - Find:
    - Adjective(s): "red"
    - Adverb(s): "carefully"
    - Conjunction(s): "and"
    - Common noun(s): "friend", "book" and "shelf"
    - Determiner(s): "the"
    - Interjection(s): "Wow"
    - Preposition(s): "from"
    - Proper noun(s): "Sarah"
    - Pronoun(s): "her"
    - Verb(s): "hands" and "smiles"

Overview

- Learning objectives
- Topics
  - Linguistic Knowledge
  - Morphology
    - Word Classes
    - Part-of-speech tagging
    - Morphology Main Concepts
- Key takeaways
- Suggested readings

PART-OF-SPEECH TAGGING

- Part-of-Speech (PoS) tagging is the task of assigning a word class to each word in a text

- By the way: not to confuse with Morphological Analysis – we will talk about this next

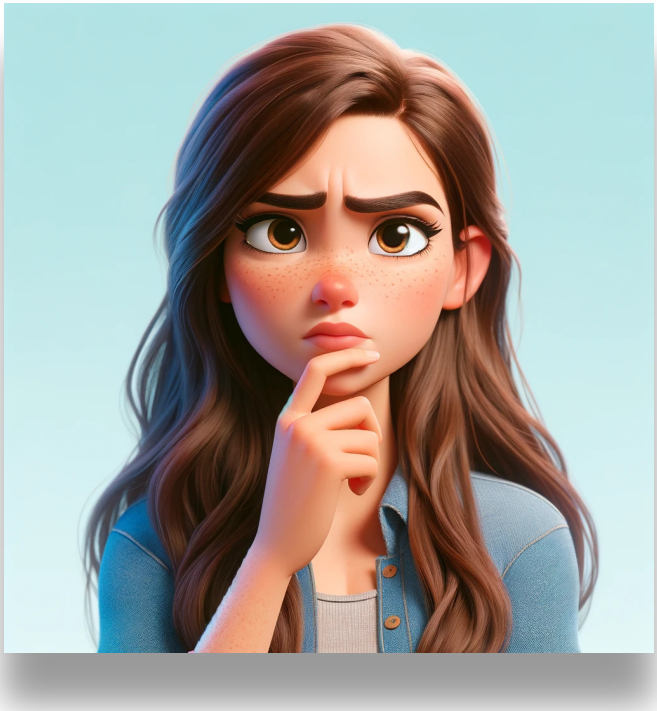
PART-OF-SPEECH TAGGING

- Challenges:
  - There is not a single tag set for part-of-speech tags
  - Words are ambiguous (Example: book)
- Good News:
  - Just a small percentage of words are ambiguous!
- Bad news:
  - Ambiguous words are the most frequent words!
- Example from the Brown Corpus:
  - 11,5% of the words are ambiguous (form)
  - 40% of the words in the corpus are ambiguous.

SOME APPROACHES TO PART-OF-SPEECH TAGGING

- Rule-based
- Stochastic
- Deep Learning

Where have I seen this before?



RULE-BASED PART-OF-SPEECH TAGGING

- The algorithms operate in two steps:
  - STEP 1: with the help of a dictionary, a list of potential tags is assigned to each word
  - STEP 2: the tag to be assigned is chosen based on sets of rules usually designed by humans (otherwise automatically extracted from data annotated by humans)

RULE-BASED PART-OF-SPEECH TAGGING

- Example: He had a book
  - STEP 1:
    - He he/pronoun
    - Had have/verb
    - A a/article
    - Book book/noun book/verb
  - STEP 2
    - "Rule XPTO: If the previous tag is an article, then eliminate the verb tag."
    - Thus: Book book/noun

SOME APPROACHES TO PART-OF-SPEECH TAGGING

- ~~Rule-based~~
- Stochastic
- Deep Learning

STOCHASTIC PART-OF-SPEECH TAGGING

- Goal: choose the best sequence of tags

$$T = t_1 t_2 \ldots t_n$$

for a given sentence (sequence of words)

$$W = w_1 w_2 \ldots w_n$$

  - That is: calculate the most likely (highest probability) tag sequence for a sequence of words

STOCHASTIC PART-OF-SPEECH TAGGING

- We will estimate P(T | W) with an HMM tagger
- HMM taggers make two simplifying assumptions:

$$P(T|W) \approx \prod_{i=1}^{n} P(t_i|t_{i-1}) \times P(w_i|t_i)$$

Transition: bigram assumption

Emission: the probability of a word depends only on its own tag

STOCHASTIC PART-OF-SPEECH TAGGING

What is an HMM tagger?

STOCHASTIC PART-OF-SPEECH TAGGING

- A Hidden Markov Model (HMM) is a statistical model that describes a system with unobservable (hidden) states (in our case, the tags) through observable sequences (in our case the words), with the transitions among states being characterized by certain probabilities.

STOCHASTIC PART-OF-SPEECH TAGGING

- To calculate transitions' probabilities, we use:

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}$$

Number of times t_i follows t_{i-1}

- To calculate the emissions' probabilities, we use:

$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$

Number of times t_i is the tag of word W_i

ACTIVE LEARNING MOMENT

EXERCISE

- In the Brown corpus:
  - The DT tag occurs 116,454 times and appears before an NN 56,509 times. Then:
    - P(NN | DT) =
  - The VBZ tag occurs 21,627 times, and VBZ is the tag for "is" 10,073 times. Then:
    - P(is | VBZ) =

EXERCISE

- In the Brown corpus:
  - The DT tag occurs 116,454 times and appears before an NN 56,509 times. Then:
    - P(NN | DT) = C(DT, NN)/C(DT) = 56,509/116,454 = 0.49
  - The VBZ tag occurs 21,627 times, and VBZ is the tag for "is" 10,073 times. Then:
    - P(is | VBZ) = C(VBZ, is)/C(VBZ) = 10,073/21,627 = 0.47

STOCHASTIC PART-OF-SPEECH TAGGING

- After the counts, HMMs usually take advantage of the Viterbi algorithm for decoding

VITERBI

STOCHASTIC PART-OF-SPEECH TAGGING

- Viterbi:
  - uses dynamic programming
  - seeks the best path for a given observation, using:
    - the probability of the previous path
    - the transition probability

VITERBI

$i \leftarrow 1$
while $i < N$ **do**
  $\mathrm{SS(i, 1)} = \mathrm{P(w_1 \mid L_i)}*\mathrm{P(L_i \mid} < s >)$
  $\mathrm{BP(i, 1)} = 0$
  $i++$
end while
$t \leftarrow 2$
while $t < n$ **do**
  $i \leftarrow 1$
  while $i < N$ **do**
    $\mathrm{SS(i, t)} = max_{j=1,\ldots,N}\ \mathrm{SS(j, t\text{-}1)}*\mathrm{P(L_i \mid L_j)}*\mathrm{P(w_t \mid L_i)}$
    $\mathrm{BP(i, t)} =$ j that resulted in the maximum score
    $i++$
  end while
  $t++$
end while
$\mathrm{C(n)} =$ i that maximizes $\mathrm{SS(i, n)}$
$i \leftarrow n - 1$
while $i > 1$ **do**
  $i--$
  $\mathrm{C(n)} = \mathrm{BP(C(i+1), i+1)}$
end while

- N = number of tags
- n = number of words in the sequence

- Data structures:
  - SS (sequence score) – records the score of the best sequence found up to a given position with category L.
  - BP (Back Pointer) – records the previous state to a given state
  - C – records the best sequence of tags.

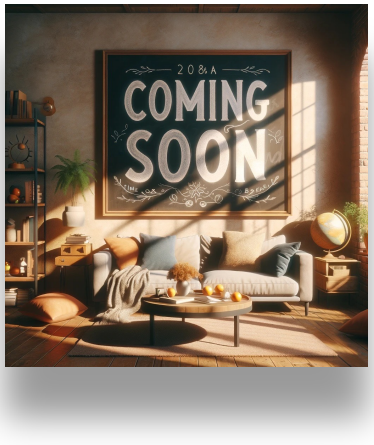
A quick look at Viterbi

Note: Fictitious values

SS	John	likes	Mary
Noun	0.6	0.015	0.6
Verb	0.3	0.045	0.0735

BP			
Noun	0	1 or 2	2
Verb	0	1	1

- Soon, in a lab near you



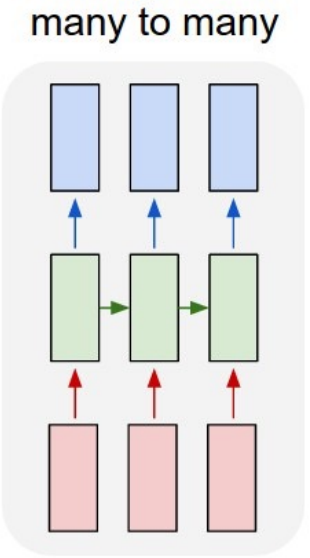
SOME APPROACHES TO PART-OF-SPEECH TAGGING

- ~~Rule-based~~
- ~~Stochastic~~
- Deep Learning

DEEP LEARNING PART-OF-SPEECH TAGGING

- PoS is a sequence labelling task
- RNNs, LSTMs and other architectures can be used to train a PoS model:
  - The model is trained on a labelled dataset
    - Each word is tagged with its correct PoS
  - The model learns to predict the tag of each word based on the context and in the word itself

Remember?



BY THE WAY…

- You can also use ChatGPT (and friends)
  - Some people consider that these “old” NLP tasks are good for evaluating current LLMs

Generated by DALL-E

Overview

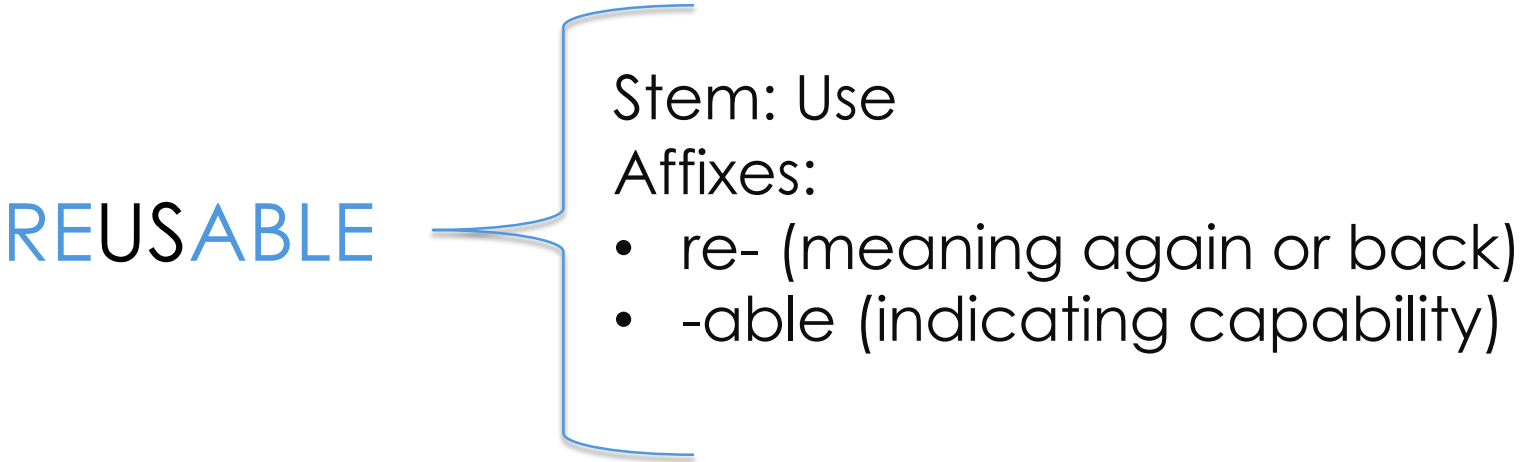
- Learning objectives
- Topics
  - Word Classes
  - Part-of-speech tagging
  - Morphology Main Concepts
    - Morphemes
    - Building Words
    - Morphological analysis
- Key takeaways
- Suggested readings

MAIN CONCEPTS

- Morphology is the linguistics field dedicated to the study of the internal structure of words (morph = shape, logos = word)
- Words are constituted by (meaningful) units called morphemes

MAIN CONCEPTS

- There are two types of morphemes:
  - Stems (or lexical morphemes): carry the primary of meaning of words
  - Affixes (or grammatical morphemes): change stems meaning and/or have grammatical functions
    - Example:



MAIN CONCEPTS

- Affixes' types:
  - Prefixes: beginning of the word
    - Examples:
      - Adding "un-" to "happy" creates "unhappy"
      - Adding "re-" to "write" creates "rewrite"
  - Suffixes: end of the word
    - Example:
      - Adding "-ness" to "happy" creates "happiness"
      - Adding "-ly" to "quick" creates "quickly"

MAIN CONCEPTS

- More affixes' types:
  - Infixes: inserted inside the stem
    - Example:
      - Inserting "-freaking-" into "unbelievable" as in "un-freaking-believable" (English slang example by ChatGPT)
      - Editor-in-chief + s -> Editors-in-chief

MAIN CONCEPTS

- More affixes' types:
  - Circumfixes: precede and follow the stem; inserted at the same time
    - Examples:
      - In Portuguese: entardecer, amanhecer, embelezar
      - German (example by ChatGPT):
        - Root: lieb ("love" or "dear")
        - Circumfix: ge- … -t
        - Word: geliebt ("loved")
  - Clitics: function like a word, but do not appear alone
    - Example:
      - Eu contei-os. (I counted them)

MAIN CONCEPTS

- In some languages words can contain an impressive number of morphemes
  - Example: Turkish, for instance, has many words with 9 or 10 morphemes

EXAMPLE

(from Prof. Nuno Mamede slides)

Turkish has lots of afixes

Avrupa	Europe
Avrupalı	of Europe / European
Avrupalılaş	become European
Avrupalılaştır	Europeanise
Avrupalılaştırama	be unable to Europeanise
Avrupalılaştıramadık	we couldn't Europeanise
Avrupalılaştıramadık	one that is unable to be…
Avrupalılaştıramadıklar	unable to be Europeanised ones
Avrupalılaştıramadıklarımız	they, who we couldn't manage to…
Avrupalılaştıramadıklarımızdan	of them who we couldn't manage to Europeanise
Avrupalılaştıramadıklarımızdanmış	is reportedly of ours that we were unable to Europeanise
Avrupalılaştıramadıklarımızdanmışsınız	you are reportedly of ours that we were unable to Europeanise
Avrupalılaştıramadıklarımızdanmışsınızcasına	as if you were reportedly of ours that we were unable to Europeanise

EXAMPLE

(generated by ChatGPT – use edit distance to compare with Wikipedia example and check its veracity)

- Turkish Word:

Muvaffakiyetsizleştiricileştiriveremeyebileceklerimizdenmişsinizcesine

  - Meaning: As if you are among those whom we may not be able to easily make into a maker of unsuccessful ones
  - Breaking it down:
    - Stem: Muvaffak (successful)
    - Affixes: -iyet (noun-forming suffix related to doing the action), -siz (without), -leş (become), -tir (cause/make), -ici (agent or doer), -leş (become), -tir (cause/make), -iver (sudden action), -eme (cannot), -yebil (ability), -ecek (future tense), -ler (plural), -imiz (our), -den (from), -miş (past participle), -siniz (you are), -cesine (as if)

ACTIVE LEARNING MOMENT

EXERCISE

- Find (if possible) words in your own native language in which one of the morphemes is a:
  - prefix
  - suffix
  - infix (inside the stem)
  - circumfix (inserted at the same time before and after the stem)
  - clitic

Overview

- Learning objectives
- Topics
  - Word Classes
  - Part-of-speech tagging
  - Morphology Main Concepts
    - Morphemes
    - Building Words
    - Morphological analysis
- Key takeaways
- Suggested readings

MAIN CONCEPTS

- Building words from a word stem:
  - Inflection: Doesn't change the word class or the meaning of the word, considering the original stem
    - Examples:
      - eats from eat (both verbs) and gatas (female cats) from gato (cat) (both nouns).
  - Derivation: Results in a word from a different word class or with a "different meaning"
    - Examples:
      - do from undo (opposite meanings) and amigável (friendly) from amigo (friend) (adjective and noun).

MAIN CONCEPTS

- More ways of building words from a word stem:
  - Compounding: Combination of multiple word stems
    - Examples:
      - doghouse (dog + house) and chapéu-de-chuva (chapéu + de + chuva) (umbrella – something like "hat for rain")
  - Cliticization: Words with clitics
    - Example: apagou-o (erased it or turned it off).
  - ...

ACTIVE LEARNING MOMENT

EXERCISE

- Find (if possible) words in your own native language that are formed based on:
  - Inflection (no changes in the word class/meaning, considering the stem)
  - Derivation (changes)
  - Compounding (multiple word stems)
  - Cliticization (words with clitics)

Overview

- Learning objectives
- Topics
  - Word Classes
  - Part-of-speech tagging
  - Morphology Main Concepts
    - Morphemes
    - Building Words
    - Morphological analysis
- Key takeaways
- Suggested readings

MORPHOLOGICAL ANALYSIS

- Analyzes the structure of a word to understand its components (stem, prefixes, etc.), and how these contribute to the word's meaning and grammatical function.
- Subtasks:
  - Lemmatization: finding the dictionary form of a word
    - Example:
      - has → have
      - running → run
  - Stemming: reducing a word to its stem/root form (may not be a valid word on its own)
    - Example:
      - running" → runn
  - …

MORPHOLOGICAL ANALYSIS

- We will not study any form of Morphological Analysis

KEY TAKEAWAYS

KEY TAKEAWAYS

- Define Part-of-Speech
- Explain how the following can be performed:
  - Rule-based part-of-speech tagging
  - HMM part-of-speech tagging
- Define morpheme and stem
- Identify different types of affixes in a word
- Explain the difference between inflectional and derivational morphology
- Understand the difference between stemming and lemmatization

SUGGESTED READINGS

READINGS

- Sebenta:
  - Morphology? What is it? Is there a cure?
- Jurafsky:
  - Chapter 8 (Sequence Labelling for …):
    - 8.1-8.4