Unit 3
Part of Speech Tagging
PoS, PoS tagging, PoSTagset

Natural Language Processing (NLP)
MDS 555

Objective

- PoS
- PoS tagging
- PoSTagset
- Hidden Markup Model
- Rule Based PoSTagging
- Stochastic PoS Tagging
- Transformation based Tagging

What is PoS?

- A category to which a word is assigned in accordance with its syntactic functions
- The role a word plays in a sentence denotes what part of speech it belongs to
- In English the main parts of speech are
  - noun, pronoun, adjective, determiner, verb, adverb, preposition, conjunction, and interjection

Noun

- Common and Proper
  - She bought a pair of shoes. (thing)
  - I have a pet. (animal)
  - Is this your book? (object)
  - Many people have a fear of darkness. (ideas/abstract nouns)
  - He is my brother. (person)
  - This is my school. (place)
- Singular and Plural Nouns

Pronouns

- words that are used to substitute a noun in a sentence
  - **I** reached home at six in the evening. (1st person singular pronoun)
  - Did **someone** see a red bag on the counter? (Indefinite pronoun)
  - Is this the boy **who** won the first prize? (Relative pronoun)
  - That is **my** mom. (Possessive pronoun)
  - I hurt **myself** yesterday when we were playing cricket. (Reflexive pronoun)

Verb

- denote an action that is being performed by the noun or the subject in a sentence
  - She **plays** cricket every day.
  - Darshana and Arul are **going** to the movies.
  - My friends **visited** me last week.
  - Did you **have** your breakfast?
  - My name **is** Meenakshi Kishore

Adverb

- used to provide more information about verbs, adjectives and other adverbs used in a sentence
  - Did you come **here** to buy an umbrella? (Adverb of place)
  - I did not go to school **yesterday** as I was sick. (Adverb of time)
  - Savio reads the newspaper **everyday**. (Adverb of frequency)
  - Can you please come **quickly**? (Adverb of manner)
  - Tony was so sleepy that he could **hardly** keep his eyes open during the meeting. (Adverb of degree)

Adjectives

- used to describe or provide more information about the noun or the subject in a sentence
- include good, ugly, quick, beautiful, late
  - The place we visited yesterday was **serene**.
  - Did you see how **big** that dog was?
  - The weather is **pleasant** today.
  - The **red** dress you wore on your birthday was lovely.
  - My brother had only **one** bread for breakfast.

Prepositions

- used to link one part of the sentence to another
- Prepositions show the position of the object or subject in a sentence
  - in, out, besides, in front of, below, opposite

Conjunctions

- used to connect two different parts of a sentence, phrases and clauses
  - and, or, for, yet, although, because, not only
    - Meera **and** Jasmine had come to my birthday party.
    - Jane did not go to work **as** she was sick.
    - **Unless** you work hard, you cannot score good marks.
    - I have not finished my project, **yet** I went out with my friends.

Interjunctions

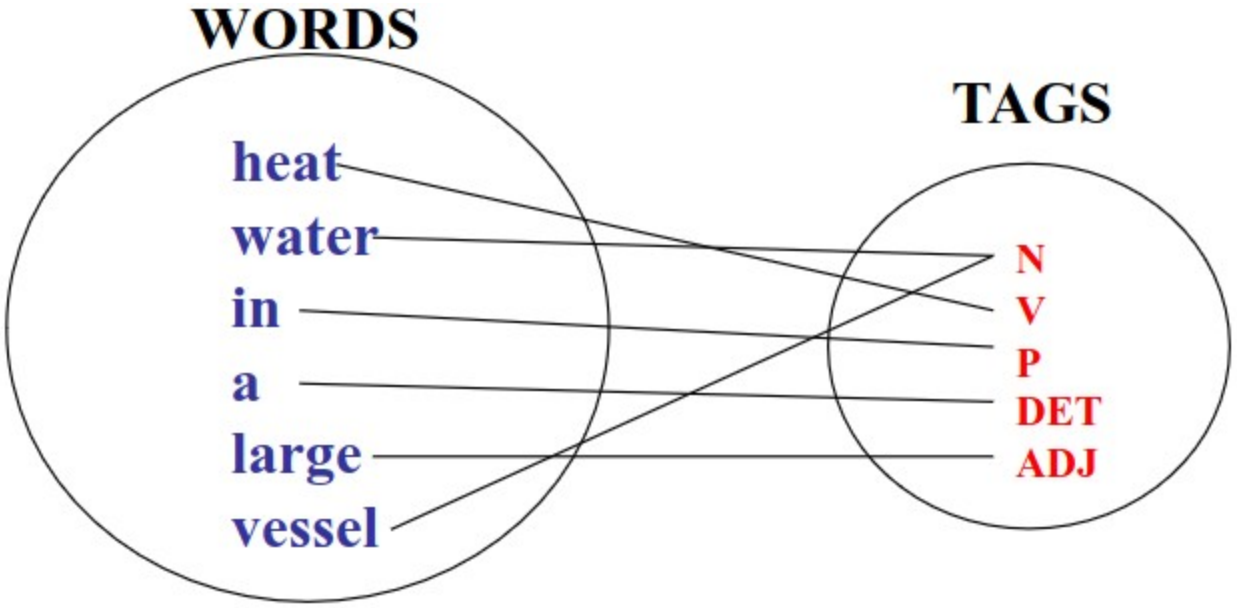
- used to convey strong emotions or feelings.
- oh, wow, alas, yippee, etc.
- It is always followed by an exclamation mark.
  - wow! What a wonderful work of art.
  - Alas! That is really sad.
  - Yippee! We won the match.

PoS Categories: Open vs. Closed Classes

- Closed
  - limited number of words, do not grow usually
    - e.g., Auxiliary, Article, Determiner, Conjuction, Pronoun, Preposition, Particle, Interjection
- Open
  - unlimited number of words
    - e.g., Noun, Verb, Adverb, Adjective

PoS Tagging

- The process of assigning a part-of-speech to each word in a sentence



Word	Tag
heat	verb (noun)
water	noun (verb)
in	prep (noun, adv)
a	det (noun)
large	adj (noun)
vessel	noun

PoS Tagsets

- There are many parts of speech tagsets
- Tag types
  - Coarse-grained
    - Noun, verb, adjective, …
  - Fine-grained
    - noun-proper-singular, noun-proper-plural, noun-common-mass, ..
    - verb-past, verb-present-3rd, verb-base, …
    - adjective-simple, adjective-comparative, ...

PoS Tagsets

- Brown tagset (87 tags)
  - Brown corpus
  - https://en.wikipedia.org/wiki/Brown_Corpus
- C5 tagset (61 tags)
- C7 tagset (146 tags!)
- Penn TreeBank (45 tags) – most used
  - A large annotated corpus of English tagset
  - https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
- UPenn TreeBank II - 36 tags

PoS Tag: Challenge

- words often have more than one POS
  - On my back[NN] (noun)
  - The back[JJ] door (adjective)
  - Win the voters back[RB] (adverb)
  - Promised to back[VB] the bill (verb)

Challenge: Ambiguity in POS tags

- 45-tags Brown corpus (word types)
  - Unambiguous (1 tag): 38,857
  - Ambiguous: 8,844
    - 2 tags: 6,731
    - 3 tags: 1,621
    - 4 tags: 357
    - 5 tags: 90
    - 6 tags: 32
    - 7 tags: 6 (well, set, round, open, fit, down)
    - 8 tags: 4 ('s, half, back, a)
    - 9 tags: 3 (that, more, in)

Challenge: OOV in POS

- Words that are not present in the training data of a POS tagger can be difficult to tag accurately, especially if they are rare or specific to a particular domain.

Challenge: Complex grammatical structures

- Languages with complex grammatical structures, such as languages with many inflections or free word order, can be more challenging to tag accurately

Challenge: Lack of annotated training data

- Some languages or domains may have limited annotated training data, making it difficult to train a high-performing POS tagger

Challenge: Inconsistencies in annotated data

- Annotated data can sometimes contain errors or inconsistencies, which can negatively impact the performance of a POS tagger

Baseline Method for PoS Tag

- Very basic PoS Tagging method
  - Tagging unambiguous words with the correct label
  - Tagging ambiguous words with their most frequent label
  - Tagging unknown words as a noun
- This method performs around 90% precision

Use of PoS Tagging

- Information extraction:
  - POS tagging can be used to identify specific types of information in a text, such as names, locations, and organizations.
  - This is useful for tasks such as extracting data from news articles or building knowledge bases for artificial intelligence systems.

Use of PoS Tagging

- Named entity recognition
  - POS tagging can be used to identify and classify named entities in a text, such as people, places, and organizations.
  - This is useful for tasks such as building customer profiles or identifying key figures in a news story.

Use of PoS Tagging

- Text classification
  - POS tagging can be used to help classify texts into different categories, such as spam emails or sentiment analysis.
  - By analyzing the POS tags of the words in a text, algorithms can better understand the content and tone of the text

Use of PoS Tagging

- Machine translation
  - POS tagging can be used to help translate texts from one language to another by identifying the grammatical structure and relationships between words in the source language and mapping them to the target language.

Use of PoS Tagging

- Natural language generation
  - POS tagging can be used to generate natural-sounding text by selecting appropriate words and constructing grammatically correct sentences.
  - This is useful for tasks such as chatbots and virtual assistants.

Methods of PoS Tagging

- Rule-Based POS tagging
  - e.g., ENGTWOL [Voutilainen, 1995]
  - large collection (> 1000) of constraints on what
  - sequences of tags are allowable

Methods of PoS Tagging

- Transformation-based tagging
  - e.g.,Brill's tagger [Brill, 1995]

Methods of PoS Tagging

- Stochastic (Probabilistic) tagging
  - e.g., TNT [Brants, 2000]

Thank you