# POS, NER AND NLP

Gabriella Pasi
gabriella.pasi@unimib.it

# Text Representation

- We have presented in the previous lessons the preliminary phases for text processing and representation
- We have also seen that in several text mining tasks, terms and term weighting constitute the basis for text representation
- We have also seen that positional information of terms in a text can be obtained by extracting n-grams
- Now we will see how text representation can be enriched with simple NLP techniques.

# NLP: Why?

- Texts are objects with inherent complex structure. A simple BoW model is not good enough for text understanding.
- ***Natural Language Processing*** provides models that go deeper to uncover the meaning.
  - Part-of-speech tagging
  - NER
  - Syntactic analysis
  - Semantic analysis
  - Discourse structure

# Text Representation: Phrases

- Phrases are
  - More informative than single words
    - e.g., documents containing “black sea” vs. two words “black” and “sea”
  - Less ambiguous
    - e.g., “big apple” vs. “apple”

# Text Representation: Phrases

- Text processing issue: how are phrases recognized?

- Three possible approaches:
  - Use word *n-grams (we have seen this)*
  - Identify the syntactic role of words within phrases by using a *part-of-speech* (POS) tagger (*we will see it "at high level" today*)
  - Store word positions of indexes in texts, and use *proximity operators* in queries (*we will see this when we will introduce search engines*)

# POS and NER

We will present POS and Named Entity Recognition as *sequence labeling problems* or *tagging problems:* given a sequence of words in input the aim is to define a *model* that produces in output a sequence of labels (tags).

Either this model can be a rule-based model or it can be a supervised learning problem.

In particular, an important class of models for supervised learning problems is represented by *generative models*.

# Supervised learning

In supervised learning problems we assume the availability of training examples $(x^{(1)}, y^{(1)}) \dots (x^{(m)}, y^{(m)})$, where each example is a pair consisting of an input $x^{(i)}$ paired with a label $y^{(i)}$.

The task is to learn a function $f : X \rightarrow Y$ that maps any input $x$ to a label $f(x)$.

In tagging problems each $x^{(i)}$ is a sequence of words

$x_1^{(i)}, x_2^{(i)}, x_3^{(i)} \dots\dots x_{ni}^{(i)}$ and each $y^{(i)}$ is a sequence of tags $y_1^{(i)}, y_2^{(i)}, y_3^{(i)} \dots\dots y_{ni}^{(i)}$

(ni refers to the length of the i'th training example)

# Conditional and generative models

One way to define the function f(x) is through a *conditional model*. In this approach the model defines the conditional probability $p(y|x)$ for any $(x, y)$ pair.

An alternative approach (often used in NLP) is to define a *generative model*. Rather than directly estimating the conditional distribution $p(y|x)$, generative models estimate the *joint* probability $p(x, y)$ over $(x, y)$ pairs. The parameters of the model $p(x, y)$ are again estimated from the training examples $(x^{(i)}, y^{(i)})$ for $i = 1 \ldots n$.

Models that decompose a joint probability into into terms $p(y)$ and $p(x|y)$ are often called *noisy-channel* models.

# Part of Speech Tagging (POS)

# Part-of-Speech tagging (POS tagging)

- Once the preliminary text processing phases have been undertaken, POS tagging aims at marking up a word in a text (corpus) by a tag corresponding to a particular *part of speech* (POS tags can be of varying granularity)

A + dog + is + chasing + a + boy + on + the + playground

| A | + | dog | + | is | + | chasing | + | a | + | boy | + | on | + | the | + | playground |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Det | | Noun | | Aux | | Verb | | Det | | Noun | | Prep | | Det | | Noun |

# Word Classes

- Words that somehow ‘behave’ alike:
  - Appear in similar contexts
  - Perform similar functions in sentences
  - Undergo similar transformations
- ~ 9 traditional word classes of parts of speech for IndoEuropean languages
  - Noun, verb, pronoun, adjective, preposition, adverb, article, conjunction, interjections
  - Called: parts-of-speech, lexical categories, word classes, morphological classes, lexical tags, POS

# Some Examples of POS tags

| | | |
|---|---|---|
| • N | noun | chair, bandwidth, pacing |
| • V | verb | study, debate, read |
| • ADJ | adjective | purple, tall, ridiculous |
| • ADV | adverb | unfortunately, slowly |
| • P | preposition | of, by, to |
| • PRO | pronoun | I, me, mine |
| • DET | determiner | the, a, that, those |
| • CONG | conjunction | and, or |

# Closed and open class words

## Closed class words

- Relatively fixed membership
- Usually **function** words: short, frequent words with grammatical function
  - determiners: ***a, an, the***
  - pronouns: ***she, he, I***
  - prepositions: ***on, under, over, near, by, …***

## Open class words

- Usually **content** words: Nouns, Verbs, Adjectives, Adverbs
  - Plus interjections: **oh, ouch, uh-huh, yes, hello**
- New nouns and verbs like *iPhone* or *to fax*

# Closed and open class words

**Open class** ("content") words

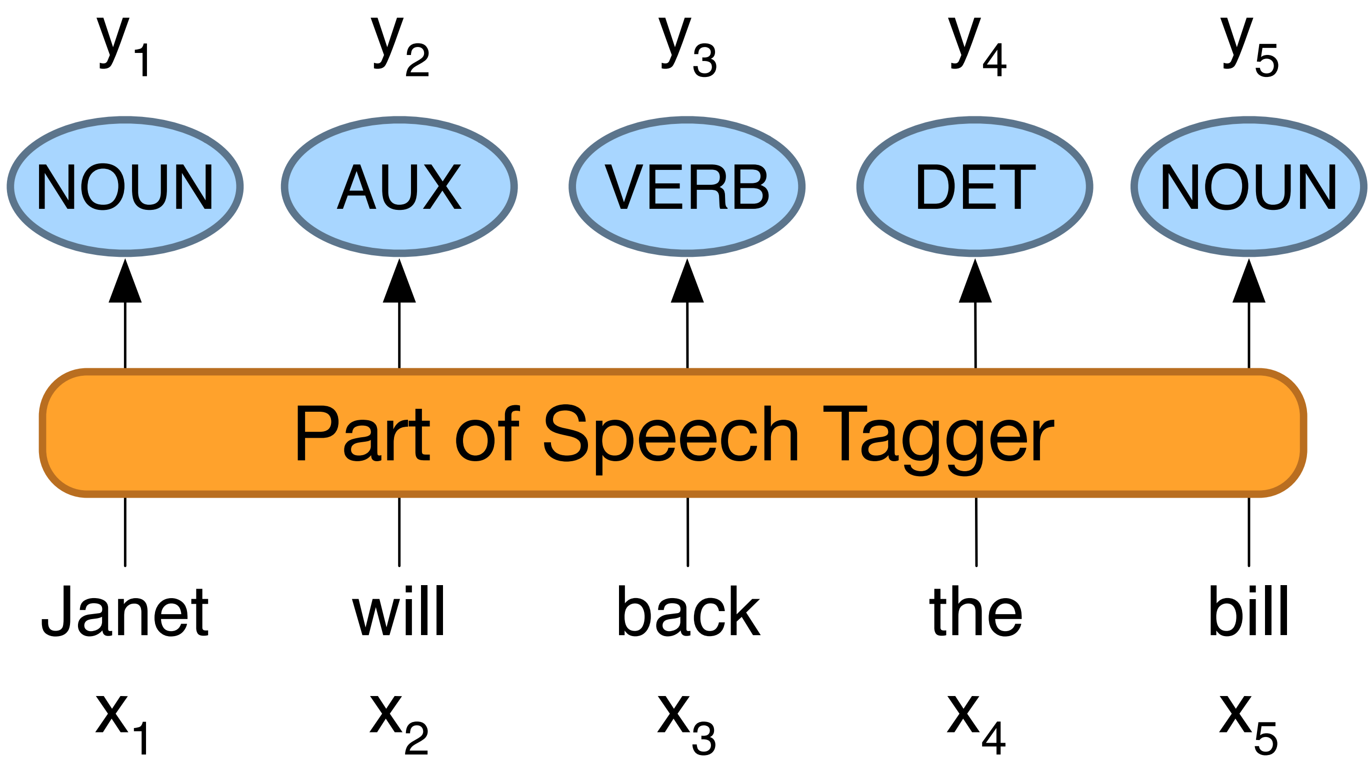
- Nouns
  - Proper: *Janet*, *Italy*
  - Common: *cat, cats*, *mango*
- Verbs
  - Main: *eat*, *went*
- Adjectives: *old green tasty*
- Adverbs: *slowly yesterday*
- Numbers: *122,312*, *one*
- Interjections: *Ow hello*
- *… more*

**Closed class** ("function")

- Determiners: *the some*
- Conjunctions: *and or*
- Pronouns: *they its*
- Auxiliary (verbs): *can*, *had*
- Prepositions: *to with*
- Particles: *off up*
- *… more*

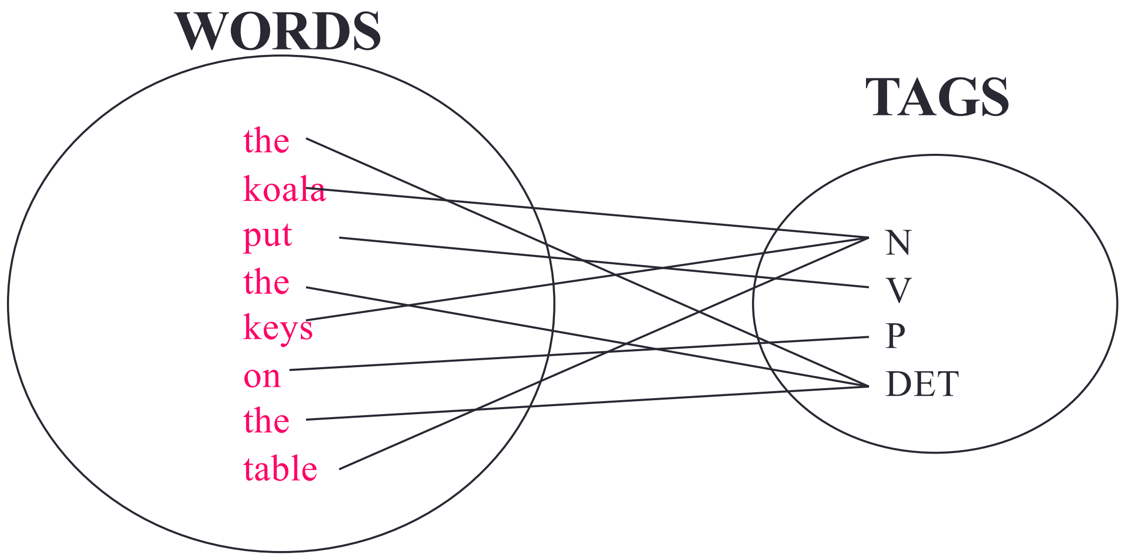
# POS Tagging

Map from sequence $x_1,\ldots,x_n$ of words to $y_1,\ldots,y_n$ of POS tags



# Defining POS Tagging

- The process of assigning a part-of-speech or lexical class marker (tag) to each word in a corpus:



# Applications for POS Tagging

- Parsing: e.g. *Time flies like an arrow*
  - Is *flies* an N or V?
- Word prediction in speech recognition
  - Possessive pronouns (*my, your, her*) are likely to be followed by nouns
  - Personal pronouns (*I, you, he*) are likely to be followed by verbs
- Machine Translation

# Pos Tagging Example

**Original text:**

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales.

**Brill tagger:**

Document/NN will/MD describe/VB marketing/NN strategies/NNS carried/VBD out/IN by/IN U.S./NNP companies/NNS for/IN their/PRP agricultural/JJ chemicals/NNS ,/, report/NN predictions/NNS for/IN market/NN share/NN of/IN such/JJ chemicals/NNS ,/, or/CC report/NN market/NN statistics/NNS for/IN agrochemicals/NNS ,/, pesticide/NN ,/, herbicide/NN ,/, fungicide/NN ,/, insecticide/NN ,/, fertilizer/NN ,/, predicted/VBN sales/NNS ,/, market/NN share/NN ,/, stimulate/VB demand/NN ,/, price/NN cut/NN ,/, volume/NN of/IN sales/NNS ./.

Eric Brill. 1992. A simple rule-based part of speech tagger. In Proceedings of the third conference on Applied natural language processing (ANLC '92). Association for Computational Linguistics, Stroudsburg, PA, USA, 152-155

# Choosing a POS Tagset

- To do POS tagging, first need to choose a set of tags
- Could pick very coarse (small) tagsets
  - N, V, Adj, Adv.

- More commonly used: Brown Corpus (general corpus, Francis & Kucera ‘82), 1 Million words, 87 tags – more informative but more difficult to tag.

- Most commonly used: Penn Treebank – 45 tags: hand-annotated corpus of *Wall Street Journal*

*https://www.sketchengine.eu/penn-treebank-tagset/*

# Penn Treebank Tagset

| Tag | Description | Example | Tag | Description | Example |
|---|---|---|---|---|---|
| CC | Coordin. Conjunction | *and, but, or* | SYM | Symbol | +,%, & |
| CD | Cardinal number | *one, two, three* | TO | “to” | *to* |
| DT | Determiner | *a, the* | UH | Interjection | *ah, oops* |
| EX | Existential ‘there’ | *there* | VB | Verb, base form | *eat* |
| FW | Foreign word | *mea culpa* | VBD | Verb, past tense | *ate* |
| IN | Preposition/sub-conj | *of, in, by* | VBG | Verb, gerund | *eating* |
| JJ | Adjective | *yellow* | VBN | Verb, past participle | *eaten* |
| JJR | Adj., comparative | *bigger* | VBP | Verb, non-3sg pres | *eat* |
| JJS | Adj., superlative | *wildest* | VBZ | Verb, 3sg pres | *eats* |
| LS | List item marker | *1, 2, One* | WDT | Wh-determiner | *which, that* |
| MD | Modal | *can, should* | WP | Wh-pronoun | *what, who* |
| NN | Noun, sing. or mass | *llama* | WP$ | Possessive wh- | *whose* |
| NNS | Noun, plural | *llamas* | WRB | Wh-adverb | *how, where* |
| NNP | Proper noun, singular | *IBM* | $ | Dollar sign | $ |
| NNPS | Proper noun, plural | *Carolinas* | # | Pound sign | # |
| PDT | Predeterminer | *all, both* | “ | Left quote | (‘ or “) |
| POS | Possessive ending | *’s* | ” | Right quote | (’ or ”) |
| PRP | Personal pronoun | *I, you, he* | ( | Left parenthesis | ( [, (, {, <) |
| PRP$ | Possessive pronoun | *your, one’s* | ) | Right parenthesis | ( ], ), }, >) |
| RB | Adverb | *quickly, never* | , | Comma | , |
| RBR | Adverb, comparative | *faster* | . | Sentence-final punc | (. ! ?) |
| RBS | Adverb, superlative | *fastest* | : | Mid-sentence punc | (: ; ... – -) |
| RP | Particle | *up, off* | | | |

# Tag Ambiguity

- Words often have more than one POS: e.g., *back*
  - The *back* door = JJ (adjective)
  - On my *back* = NN (singular noun)
  - Win the voters *back* = RB (adverb)
  - Promised to *back* the bill = VB (verb)

- The POS tagging problem is ***to determine the POS tag for a particular instance of a word***

# Tagging Whole Sentences with POS is Hard

- Ambiguous POS contexts
  - E.g., Time flies like an arrow.

- Possible POS assignments
  - Time/[V,N] flies/[V,N] like/[V,Prep] an/Det arrow/N
  - Time/N flies/V like/Prep an/Det arrow/N
  - Time/V flies/N like/Prep an/Det arrow/N
  - Time/N flies/N like/V an/Det arrow/N
  - …..

# How Do We Disambiguate POS?

- Many words have only one POS tag (e.g. is, Mary, very, smallest)
- Others have a single ***most likely*** tag (e.g. a, dog)
- Tags also tend to *co-occur* regularly with other tags (e.g. Det, N)
- In addition to conditional probabilities of words $P(w_n|w_{n-1})$, we can look at POS likelihoods ($P(t_n|t_{n-1})$) to disambiguate sentences and to assess sentence likelihoods

# Some Ways to do POS Tagging

- Rule-based tagging
  - E.g. EnCG ENGTWOL tagger

- Supervised Machine Learning algorithms
  - HMM (Hidden Markov Models)
  - Conditional Random Fields/Maximum Entropy Random Models
  - Neural sequence models (RNNs or Transformers)
  - Large Language Models (like BERT), finetuned

I will not detail these methods.

# Rule based tagging

- Start with a dictionary
- Assign all possible tags to words from the dictionary
- Write rules by hand to selectively remove tags
- Leaving the correct tag for each word.

# How difficult is POS tagging in English?

Roughly 15% of word types are ambiguous

- Hence 85% of word types are unambiguous
- *Janet* is always PROPN, *hesitantly* is always ADV

But those 15% tend to be very common.

So ~60% of word tokens are ambiguous

E.g., *back*

earnings growth took a back/ADJ seat
a small building in the back/NOUN
a clear majority of senators back/VERB the bill
enable the country to buy back/PART debt
I was twenty-one back/ADV then

# POS Tagging and sentences

- POS tagging too slow for large collections
- Simpler definition – phrase is any sequence of *n* words – *n-grams.*
- Recall:
  - *bigram*: 2 words sequence, *trigram*: 3 words sequence, *unigram*: single word
  - N-grams also used at character level for applications such as OCR
- N-grams typically formed from *overlapping* sequences of words
  - i.e. move n-word "window" one word at a time in document