NATURAL LANGUAGE PROCESSING

LESSON 3: POS, SYNTAX, N-GRAMS

OUTLINE

- Part of Speech
  - Types of words and categories
- Syntax
  - Natural language rules, and context-free grammar
- N-Grams
  - Probabilistic data for NLP studies

PART OF SPEECH (POS) TAGGING

Generally it is divided into 3 groups:

- Noun, pronoun, adjective
- Verb, adverb, auxiliary verb (yardımcı fiil)
- Preposition (edat), conjuction (bağlaç), and exclamation (ünlem).

WHY WE NEED POS TAGGING?

For accurate stemming: if we know the class of word we can analyze the suffixes in right scope

- Bu ev mavi renge <u>boyanmış.</u> -> Verb boya(mak) – n (reflexive verb suffix)+(~past perfect continues)
- Ev için kullanılan senin boyanmış. -> Noun boya + n (2nd person possesive suffix) + (~past perfect continues)

WHY WE NEED POS TAGGING?

- For quickly finding names or other phrases
- For finding instances or frequencies of particular constructions in large corpora.

POS TAGGING CATEGORIES

It has two broad supercategories:

1.Closed class : fixed word types. Rarely new words gained.

- Most of the members are **Function Words**.
  - Ders kayıtı yaparken herhangi bir dersin saati bu dersin saati ile çakışmaması gerekir.

2.Open class: these types can gain new words from derived words or borrowed words from other languages.

Four major types: Noun, Adjective, Verb, Adverb

POS TAGGING CATEGORIES

Nouns:

- Proper Nouns: unique names for people, places, etc. Usually capitalized.
- Common Nouns: general names for everything.

Verbs:

- Verbs: Refer to actions, processes, occurrences. Koşmak, konuşmak, doğmak
- Auxiliary Verbs: help names to act as a verb group: etmek, -(e)bilmek

POS TAGGING CATEGORIES

Adjectives: describes the properties or qualities of nouns

- Chinese does not have adjectives
- Turkish has plenty of adjectives with plenty of subclasses

Adverbs:

- Most undetermined class: mostly modifies verbs, adverbs, entire verb phrases.
  - Bu önemli belgeyi, kurum kapanmadan, çok hızlı koşarak yetiştirmelisiniz.

POS TAGGING CATEGORIES

- Preposition, conjuction and exclamation: members of closed class

  Natural languages may have different sets of these closed class part of speech members.

  English has more:

  prepositions: on, under, over,near, by, at, from, to, with

  determiners : a, an, the

  particles: up, down, on, off, in, out, at, by

  numerals: one,two,three, first, second, third

POS TAGGING CATEGORIES

- Pronouns: They are member of closed class but they act as a kind of shorthand for referring to some noun phrases or entity or event.

- Ali okula geldi mi? Bugün onu göremedim.

PENN TREEBANK TAGS

- PENN Treebank
  - 45 Part-of-Speech tags
  - Operated between 1989-1996
  - approximately 7 million words of part-of-speech tagged text.

```
( (S
  (NP Martin Marietta Corp.)
  was
  (VP given
      (NP a
          $ 29.9
          million Air Force contract
          (PP for
              (NP low-altitude navigation
                  and
                  targeting equipment)))))
  .)
```

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	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	*#*
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	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*			

Figure 8.6 Penn Treebank Part-of-Speech Tags (Including Punctuation)

SYNTAX

- Language is not a bag of words
- Syntax -> «Setting out together or arrangement»
- Gramatical rules apply to categories and groups of words, not individual words.
- A sentence includes a subject and a predicate. The subject is a noun phrases and the predicate is a verb phrase.
- When people learn a new word, they learn its syntactic usage.
  - See in a sentence, you will easily find out unknown word's category.
    - «Students will be really zealous for this class.»

SYNTAX

- Constituents -> words or word phrases that has solid meaning
- Each word is a constituent
- Constituents are non-crossing, any two constituents has an itersection than one of them containing the other.

SYNTAX

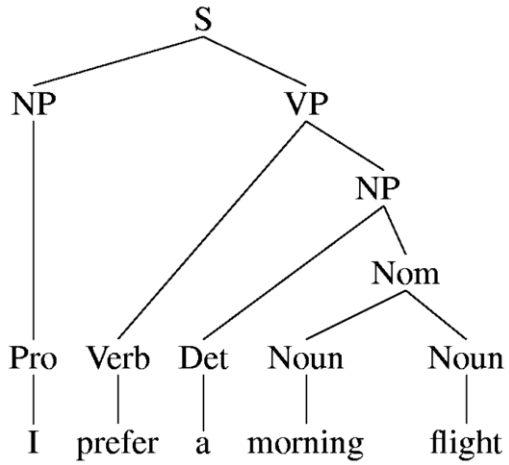
- Constituents Tests
- Pronoun test: change constituent candidate with a pronoun if it sounds right then it is correct
  - A small dog is barking in the park -> It is barking in the park. ✓
- Question test:
  - I have seen blue elephants.
  - What I have seen? -> blue elephants

CONTEXT FREE GRAMMARS FOR NLP

- English has two main group: Noun Phrase (NP) and Verb Phrase (VP)
  - S -> NP VP
  - NP -> DT CN
  - NP -> PN
  - VP -> V | NP
  - DT -> a | the
  - CN -> child | cat | dog
  - PN -> Samantha | Jorge | Min
  - V -> took | saw | liked | scared |chased

NP -> PN
NP CP -> DT CN
VP -> V | NP

CONTEXT FREE GRAMMARS FOR NLP



CONTEXT FREE GRAMMARS FOR NLP

- Wherever N is allowed in a sentence,
  - DT N -> the cat
  - JJ N -> white cat
  - DT JJ N -> the white cat

are also allowed

- We can use the notation for alternatives
  - NP -> N | DT N | JJ N | DT JJ N
- Optional categories can be also marked using parantheses:
  - NP -> (DT) (JJ) N

CONTEXT FREE GRAMMARS FOR NLP

- Verb Phrases
  - Samanta <u>ran</u>. -> VP ->V
  - Samanta <u>ran to the park</u>. -> VP -> V P NP
  - Samanta <u>ran away</u>. -> VP -> V P
  - Samanta bought a cookie. -> VP -> V NP
  - Samanta bought a cookie for John -> V NP P NP
  - Overall structure: VP -> V (NP) (P) (NP)

CONTEXT FREE GRAMMARS FOR NLP

- Adding Prepositional Phrases (on, under, over,near, by, at, from, to, with) end of Noun Phrases
  - S -> NP VP
  - NP -> (DT) (JJ) N (PP)
  - VP -> V (NP) (PP)
  - PP -> P (NP)
- Whenever a preposition is allowed, it can be followed by a noun phrase.
- NP can contain any number of PPs but only up to two NPs.
- Humans prefer PPs less than 4 in a NP.

CONTEXT FREE GRAMMARS FOR NLP

- The boy saw the woman with the telescope.
  - Did the boy saw
    - The woman that has a telescope with his bare eyes?
    - The woman with his telescope?
- Çocuk kadını teleskopla gördü.
- Acaba çocuk
  - Kadını teleskopla birlikte mi gördü? -> Çocuk, teleskoplu kadını gördü.
  - Kadını teleskoptan mı gördü? -> Çocuk teleskopla kadını gördü.

CONTEXT FREE GRAMMARS FOR NLP

- Word order and flexible inflectional suffix may resolve out most of the ambiguous sentences in English.
- Turkish actually uses Subject – Object – Verb order but in theory it has a free word order for sentence. This makes the language use constituent orders to emphasize the info in the sentence.
- Any phrase or word before predicate has the emphasized information for that sentence.

WORD ORDER

Word order	English equivalent	Proportion of languages		Example languages
SOV	"She him loves."	45%		Proto-Indo-European, Sanskrit, Hindi, Ancient Greek, Latin, Japanese, Korean
SVO	"She loves him."	42%		English, French, Hausa, Indonesian, Malay, Mandarin, Russian
VSO	"Loves she him."	9%		Biblical Hebrew, Arabic, Irish, Filipino, Tuareg-Berber, Welsh
VOS	"Loves him she."	3%		Malagasy, Baure, Proto-Austronesian
OVS	"Him loves she."	1%		Apalaí, Hixkaryana
OSV	"Him she loves."	0%		Warao

Frequency distribution of word order in languages surveyed by Russell S. Tomlin in 1980s[10][11] (v · t · e)

WORD ORDER

- Adjective Ordering in English
  - Det
  - Number
  - Strength
  - Size
  - Age
  - Shape
  - Color
  - Origin
  - Material
  - Purpose
  - Noun
- Adjective Ordering in Turkish
  - Number
  - Personal Opinion
  - Size
  - Age
  - Shape
  - Color
  - Origin
  - Material
  - Purpose
  - Noun

There is no strict order. It is unsual to use 4 or more adjectives for a noun.

N-GRAMS

- Why we need them?
  - They are leaving in about fifteen *minuets* to go to her house.
  - He is trying to *fine* out.

  İf we know the probabilities of

  minute and minuet with other words, or find and fine

  We can assume **fifteen minute** or **find out** is more sensible in the context of the sentences.

N-GRAMS

- Probabilities are based on counting things and Natural Language Processing works with **words** as countable data.
- Bigram -> looks one word into the past
  -> first-order Markov Model
- Trigram -> looks two words into the past
  ->second-order Markov Model
- N-gram -> looks N-1 words into the past

N-GRAMS MODELS

- Can be trained by counting and normalizing
- For bigrams, a particular bigram can be calculated as

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)}$$

- Can be simplified as:

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

N-GRAMS MODELS

- Bigram counts for 7 of the words (out of 1.616 total word types) in Berkeley Restaurant Project Corpus of ~10.000 sentences.

	I	want	to	eat	Chinese	food	lunch
I	8	1087	0	13	0	0	0
want	3	0	786	0	6	8	6
to	3	0	10	860	3	0	12
eat	0	0	2	0	19	2	52
Chinese	2	0	0	0	0	120	1
food	19	0	17	0	0	0	0
lunch	4	0	0	0	0	1	0

N-GRAMS MODELS

- from Turkish Dictionary

2-Gram	**olma durumu**	*4359*
	-Hayırsız olma durumu	
	-Uçarı olma durumu	
	bir biçimde	*1196*
	-Çekimsere yakışır bir biçimde	
	-Tedbirsiz bir biçimde, tedbirsiz olarak	
	yaptığı iş	*907*
	-Telgrafçının yaptığı iş	
	-Kapıcının yaptığı iş	

N-GRAMS MODELS

- from Turkish Dictionary

3-Gram	**işine konu olmak**	*416*
	-Başlama işine konu olmak	
	-Aktarma işine konu olmak	
	Bu renkte olan	*214*
	-Bu renkte olan	

N-GRAMS MODELS

- from Turkish Dictionary

4-Gram	**ihtimali veya imkânı bulunmak**	*1788*
	-Yavaşlama ihtimali veya imkânı bulunmak	
	-Tutulma ihtimali veya imkânı bulunmak	
	iline bağlı ilçelerden biri	*1061*
	-Adana iline bağlı ilçelerden biri	
	-Ankara iline bağlı ilçelerden biri	
	yapan veya satan kimse	*202*
	-Tatlı yapan veya satan kimse	
	-Yoğurt yapan veya satan kimse	

JUST GUESS: QUESTION

How about last one from me?

- If we look appearance in English written books from 1800 to 2000 for 3 N-Grams: [‘Alber Einstein’, ‘Sherlock Holmes’, ‘Frankenstein’] what will be the graph of these three N-Grams?
- Hint:
  - Sherlock Holmes first appears in ***A Study in Scarlet,*** 1887
  - Frankenstein first published at 1818
  - Alber Einstein published his paper about general relativity at 1916 and win Nobel Prize of Physics at 1921
- Cheat: try https://books.google.com/ngrams

JUST GUESS: ANSWER