NATURAL LANGUAGE PROCESSING

LESSON 5: PART OF SPEECH (POS) TAGGING

OUTLINE

- **Part of Speech (POS) Tagging**
  - What is POS Tagging?
  - Types of words and categories
- **POS Tagging Methods**
  - Rule-Based
  - Stochastic
  - Neural
- **Some POS Taggers**

WHAT IS PART OF SPEECH TAGGING?

Part-of-speech tagging, also called grammatical tagging, is the process of marking a word in a text as corresponding to a specific part of speech based on both its definition and context.

A simplified form of this is usually taught by identifying words as noun, pronoun, adjective, verb, adverb, auxiliary verb, preposition, conjuction and exclamation.

WHY WE NEED POS TAGGING?

POS tags make it possible for automatic text processing tools to take into account which part of speech each word is.

The ambiguous words can have two or more different parts of speech. POS tags are used to distinguish between the occurrences of the word when used as a noun or verb.

Besides let's think about the syntax we model English with Context Free Grammar. Using such a system, we can define sentences by sorting POS tags, regardless of what the word is.

WHY WE NEED POS TAGGING?

If we know the part of the speech tag of a word, we can analyze the suffixes in the right scope.

- Bu ev, mavi renge **boyanmış**.
  boyanmış -> boya (verb) + n (reflexive verb suffix) + mış (past)
- Ev için kullanılan, senin **boyanmış**.
  boyanmış -> boya (noun) + n (2nd person suffix) + mış (past)

POS TAGGING CATEGORIES

Part of speech tags can be categorized as open or closed classes.

Open classes can grow with new words derived from a known word or borrowed from other languages. The major open classes are nouns, verbs, adjectives and adverbs.

Closed classes have a fixed numbers of members which are usually function words. For example, determiners, pronouns, prepositions.

POS TAGGING CATEGORIES

Nouns

- A word used to identify any of a class of people, places, things or to name a particular one of these. There are two types of nouns: Proper Nouns and common nouns.

Verbs

- Refer to actions, processes, occurrences. Auxiliary Verbs help names to act as a verb group.

Adjectives

- Describe the properties or qualities of nouns. Chinese does not have adjectives. Turkish has plenty of adjectives with plenty of subclasses.

POS TAGGING CATEGORIES

Adverbs

- Most undetermined class: mostly modifies verbs, adverbs, entire verb phrases.

Preposition, conjunction and exclamation

Its classes depend the natural language. 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, entities or events.

Ali okula geldi mi? Bugün onu göremedim.

TREEBANKS AND POS TAGS

PENN Treebank

- 45 POS tags
- 1989-1996
- ~7 million words

☺

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*			

TREEBANKS AND POS TAGS

- Brown Corpus & Lancaster-Oslo-Bergen (LOB) Corpus
  - has 85 tags
- British National Corpus
  - has 61 tags
- PENN Treebank
  - has 45 tags

PART OF SPEECH TAGGING METHODS

Although the method to be chosen changes depending on whether the text is tagged or not, the part of speech tagging methods are usually told under three titles as to computation approach:

- Rule-Based
- Stochastic
- Neural

PART OF SPEECH TAGGING METHODS

Rule-Based Part-of-Speech Tagging

- The Rule-Based Part-of-Speech Tagging methods start from 1960s with two stage architecture.
- In the first stage, a dictionary is used to assign each word a list of potential parts-of-speech.
- In the second stage, a large list of hand-written disambiguation rules is used to winnow down ambiguous words to a single POS-Tag.

PART OF SPEECH TAGGING METHODS

But there are some problems about the rule based models.

- They may not be practical for active natural languages.
- Rules will never cover all situations, because natural languages has complex evolution history:

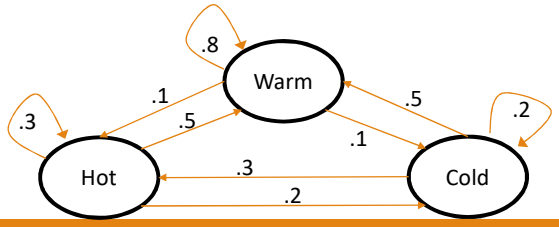
  For example, Turkish includes three main era: Old Turkish, Ottoman Turkish and new Turkish.

  - Ottoman Turkish era: Rich interaction with Arabic (Semitic) and Persian(Indo-European) languages.
  - New Turkish era: Interaction with French, English and German.

PART OF SPEECH TAGGING METHODS

Stochastic Part-of-Speech Tagging

The most common approach is Hidden Markov Model (HMM). HMMs are especially used successfully in converting speech to text in speech recognition. HMMs are based on Markov chains. A Markov chain is a model that describes a sequence of potential events in which the probability of an event is dependent only on the previous event.



PART OF SPEECH TAGGING METHODS

Trigram Hidden Markov Model

- Let the sentence be $x_1 \dots x_m$ and their POS tags are $y_1 \dots y_{m+1}$ respectively. The probability of matching the tags of the whole sentence is calculated as follows.

$$p(x_1 \dots x_m, y_1 \dots y_{m+1}) = \prod_{i=1}^{m+1} q(y_i | y_{i-2}, y_{i-1}) \prod_{i=1}^{m} e(x_i | y_i)$$

(assuming $y_0 = y_{-1} = *$)

PART OF SPEECH TAGGING METHODS

Let's give an example with a 3 words sentence.

Sentence is «**the dog laughs.**»

And the tag sequence of the sentence is «**D N V STOP**»

Here, the probability of the sentence is $p(x_1 \dots x_n, y_1 \dots y_{n+1}) = Q \times E$

Where

$$Q = q(D|*,*) \times q(N|*,D) \times q(V|D,N) \times q(STOP|N,V)$$
$$E = e(the|D) \times e(dog|N) \times e(laughs|V)$$

PART OF SPEECH TAGGING METHODS

- The value of Q is the prior probability of seeing the tag sequence «**D N V STOP**»

$$q(D|*,*) \times q(N|*,D) \times q(V|D,N) \times q(STOP|N,V)$$

- The value E can be interpreted as the conditional probability. Here, p(the dog laughs|D N V STOP)

$$e(the|D) \times e(dog|N) \times e(laughs|V)$$

PART OF SPEECH TAGGING METHODS

Neural POS Tagging

Recently, Neural POS taggers are being implemented as potential solutions to efficiently identify words in a given sentence across a paragraph.

All of these solutions aim to learn the word order of the tagged sentence and thus be able to predict it, as in the probabilistic labeling process. The success of neural systems in POS tagging, as in other areas, is based on an artificial neural network-based learning system.

PART OF SPEECH TAGGERS

Some NLTK POS Taggers

- FeaturesetTagger (Stochastic)
- NGramTagger (Stochastic)
- BrillTagger (Transitional-hybrid)
- CRFTagger (Stochastic)
- HiddenMarkovModelTagger (Stochastic)
- PerceptronTagger (Default - Neural)

PART OF SPEECH TAGGERS

- Default NLTK POSTagger for English (PerceptronTagger):

```
$ python
Python 3.6.3 (v3.6.3:2c5fed8, Oct  3 2017, 17:26:49) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> sentence = "This is a simple test sentence for part of speech tagging in English."
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['This', 'is', 'a', 'simple', 'test', 'sentence', 'for', 'part', 'of', 'speech', 'tagging', 'in',
 'English', '.']
>>> tagged = nltk.pos_tag(tokens)
>>> tagged
[('This', 'DT'), ('is', 'VBZ'), ('a', 'DT'), ('simple', 'JJ'), ('test', 'NN'), ('sentence', 'NN')
, ('for', 'IN'), ('part', 'NN'), ('of', 'IN'), ('speech', 'NN'), ('tagging', 'VBG'), ('in', 'IN')
, ('English', 'NNP'), ('.', '.')]
>>>
```

PART OF SPEECH TAGGING ISSUES

- Words can be ambiguous among multiple tags. In order to solve it, the researchers still try to improve methods. But the recent trends focus on context-dependent methods.

- Unknown words, especially new words that are not found in the dictionary data, create serious problems. To solve this, a periodically updated dictionary is needed for all living languages.