Tagsets & Part-of-Speech Tagging

By:

B. SENTHIL KUMAR

Asst. Prof / CSE, SSNCE

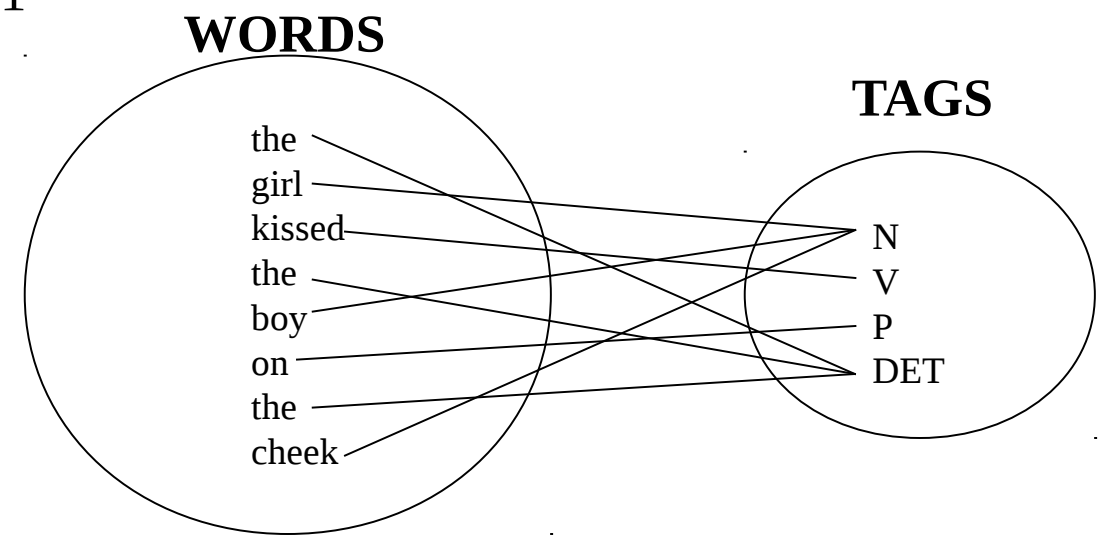
Overview

- Motivation
- Tagsets for English
- Part-of-Speech

POS Tagging – Definition

- The process of assigning a part-of-speech or other lexical class marker to each word in a corpus" (Jurafsky and Martin)

Example-1

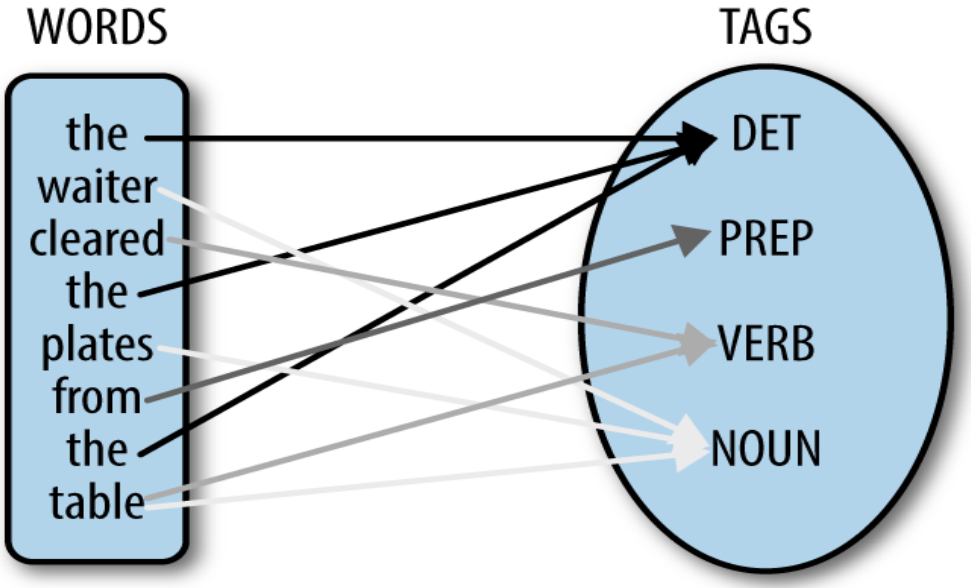


POS Tagging – Example

WORD	LEMMA	TAG
the	the	+DET
girl	girl	+NOUN
kissed	kiss	+VPAST
the	the	+DET
boy	boy	+NOUN
on	on	+PREP
the	the	+DET
cheek	cheek	+NOUN

POS Tagging – Example

Example-2



Motivation

- Speech synthesis — pronunciation
- Speech recognition — class-based N-grams
- Information retrieval — stemming, selection high-content words
- Word-sense disambiguation
- Corpus analysis of language & lexicography

Tagsets for English

- There are a small number of popular tagsets for English, many of which evolved from the 87-tag tagset used for the Brown corpus.
  - Three commonly used
    - **The small 45-tag Penn Treebank tagset**
    - The medium-sized 61 tag C5 tagset used by the Lancaster UCREL project's CLAWS tagger to tag the British National Corpus, and
    - The larger 146-tag C7 tagset

Tagsets for English

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*			

Penn Treebank POS tags

Tagsets for English

Tag the following sentence using Penn Treebank tagset:

Book that flight .

Book/VB that/DT flight/NN .

Tagsets for English

Tag the following sentence using Penn Treebank tagset:

The grand jury commented on a number of other topics.

Tagsets for English

Tag the following sentence using Penn Treebank tagset:

The grand jury commented on a number of other topics.

The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.

Tagsets for English

Tag the following sentence using Penn Treebank tagset:

There are 70 children there

There/EX are/VBP 70/CD children/NNS there/RB

Tagsets for English

- Some tagging distinctions are quite hard.
- Prepositions (IN), particles (RP), and adverbs (RB) can have a large overlap. Word like *around* can be all three:

1. Mr./NNP John/NNP never/RB got/VBD around/RP to/TO joining/VBG
2. All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB around/IN the/DT corner/NN
3. Apples/NNP costs/VBZ around/RB 250/CD

Tagsets for English

- Particles often can either precede or follow a noun phrase object
- Prepositions can not follow their noun phrase

She told off/RP her friends

She told her friends off/RP

She stepped off/IN the train

*She stepped the train off/IN

- Words that can be adjective, proper nouns are tagged as common nouns when acting as modifiers:

Chinese/NN cooking/NN

Pacific/NN waters/NNS

Tagsets for English

- Distinguishing past participles (VBN) from adjectives (JJ)

They were married/VBN by the Justice of the Peace yesterday.

At the time, he was already married/JJ

- Certain syntactic distinctions were not marked in the Penn Treebank tagset
  - *prepositions* and *subordinating conjunctions* were combined into the single tag *IN*, since the tree-structure of the sentence disambiguated them

Tagsets for English

- Brown and C5 tagsets distinguish prepositions (IN) from subordinating conjunctions (CS)

> after/CS spending/VBG a/AT few/AP days/NNS at/IN the/AT hotel/NN
> after/IN a/AT holiday/NN trip/NN to/IN Canada/NP

- Also contains two tags for word to – infinitive use as TO, prepositional use as IN

> to/TO give/VB priority/NN to/IN teacher/NN pay/NN raises/NNS

- Which tagset to use depends on how much information the application needs

Part-of-Speech Tagging

- **POS tagging (tagging)**
  - The process of assigning a part-of-speech or other lexical marker to each word in a corpus.
  - Also applied to punctuation marks
  - Thus, tagging for NL is the same process as **tokenization** for computer language, although tags for NL are much more ambiguous.
  - Taggers play an increasingly important role in speech recognition, NL parsing and IR

Part-of-Speech Tagging

- Input to tag algorithm: a string of words (ex. *Book that flight*, Penn Treebank tagset)
- Output: a single best tag for each word

VB Book that flight.	NN Hand me that book
DT Does that flight serve dinner ?	CS I thought that your flight was earlier

- Automatically assigning a tag to a word is not trivial
  - For example, *book* is **ambiguous**: it can be a verb or a noun
  - Similarly, *that* can be a determiner, or a complementizer
- The problem of POS-tagging is to resolve these ambiguities, choosing the proper tag for the context.

Part-of-Speech Tagging

- Many of the 40% ambiguous tokens are easy to disambiguate, because
  - The various tags associated with a word are not equally likely.
  - For example, *a* can be a determiner, or the letter *a* (part of acronym or initial), but the determiner sense of *a* is much more likely

Part-of-Speech Tagging

Types:		WSJ	Brown
Unambiguous	(1 tag)	44,432 (**86%**)	45,799 (**85%**)
Ambiguous	(2+ tags)	7,025 (**14%**)	8,050 (**15%**)
Tokens:			
Unambiguous	(1 tag)	577,421 (**45%**)	384,349 (**33%**)
Ambiguous	(2+ tags)	711,780 (**55%**)	786,646 (**67%**)

Figure 10.2 The amount of tag ambiguity for word types in the Brown and WSJ corpora, from the Treebank-3 (45-tag) tagging. These statistics include punctuation as words, and assume words are kept in their original case.

Part-of-Speech Tagging

- Many tagging algorithms fall into two classes:
  - Rule-based taggers
    - Involve a large database of hand-written disambiguation rule specifying, for example, that *an ambiguous word is a noun rather than a verb if it follows a determiner*. [EngCG tagger]
  - Stochastic taggers
    - Resolve tagging ambiguities by using a training corpus to *count the probability of a given word having a given tag in a given context*.
  - The Brill tagger, called the transformation-based tagger, shares features of both tagging architecture.

References

- *Speech and Language Processing,* Jurafsky and H.Martin

Thank you!