Lecture 3: Part-of-Speech Tagging and Name Entity Recognition

Tianpei Xie

Jun. 27th., 2022

Contents

	Tag	Description	Example
Open Class	**ADJ**	Adjective: noun modifiers describing properties	*red, young, awesome*
	ADV	Adverb: verb modifiers of time, place, manner	*very, slowly, home, yesterday*
	NOUN	words for persons, places, things, etc.	*algorithm, cat, mango, beauty*
	VERB	words for actions and processes	*draw, provide, go*
	PROPN	Proper noun: name of a person, organization, place, etc..	*Regina, IBM, Colorado*
	INTJ	Interjection: exclamation, greeting, yes/no response, etc.	*oh, um, yes, hello*
Closed Class Words	**ADP**	Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation	*in, on, by, under*
	AUX	Auxiliary: helping verb marking tense, aspect, mood, etc.,	*can, may, should, are*
	CCONJ	Coordinating Conjunction: joins two phrases/clauses	*and, or, but*
	DET	Determiner: marks noun phrase properties	*a, an, the, this*
	NUM	Numeral	*one, two, first, second*
	PART	Particle: a preposition-like form used together with a verb	*up, down, on, off, in, out, at, by*
	PRON	Pronoun: a shorthand for referring to an entity or event	*she, who, I, others*
	SCONJ	Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement	*that, which*
Other	**PUNCT**	Punctuation	; , ()
	SYM	Symbols like $ or emoji	$, %
	X	Other	asdf, qwfg

Figure 8.1 The 17 parts of speech in the Universal Dependencies tagset (Nivre et al., 2016a). Features can be added to make finer-grained distinctions (with properties like number, case, definiteness, and so on).

Figure 1: The common part-of-speech in English.

1 Concepts

This chapter mainly covers the **part-of-speech**, i.e. **POS**. We also covers the term **named entity** for, roughly speaking, anything that can be referred to with a proper name: a person, a location, an organization, although as well see the term is commonly extended to include things that arent entities per se. In this chapter well introduce the task of **part-of-speech tagging**, taking a sequence of words and assigning each word a part of speech like NOUN or VERB, and the task of **named entity recognition (NER)**, assigning words or phrases tags like PERSON, LOCATION, or ORGANIZATION.

Such tasks in which we assign, to each word x_i in an input word sequence, a label y_i, so that the output sequence Y has the same length as the input sequence X are called **sequence labeling tasks**.

1.1 Word Classes in English

The common part-of-speech tag can be seen in Figure 1. This includes **closed class** words and **open class** words: Closed classes are those with relatively fixed membership, such as prepositionsnew prepositions are rarely coined. By contrast, nouns and verbs are open classes, since new names and actions are invented all the time.

The open class words include:

- **noun (NOUN)**: words for person, things, places, etc. can be used for concrete terms, or abstract terms. Common nouns can be divided into **count nouns** or **mass nouns**. Count nouns can occur in the ***singular*** and ***plural*** (*goat/goats*, *relationship/relationships*) and can be *counted* (one goat, two goats). Mass nouns are used when something is conceptualized as a *homogeneous group*.
- **verb (VERB)**: words for action and processes. English verbs have ***inflections*** (***non-third-person-singular*** (eat), ***third-person-singular*** (eats), ***progressive*** (eating), ***past participle*** (eaten)).
- **adjective (ADJ)**: noun modifiers, describing noun
- **adverb (ADV)**: verb modifiers, of manner, place, time etc. **Directional adverbs** or **locative adverbs** (*home, here, downhill*) specify the direction or location of some action; **degree adverbs** (*extremely, very, somewhat*) specify the extent of some action, process, or property; **manner adverbs** (*slowly, slinkily, delicately*) describe the manner of some action or process; and **temporal adverbs** describe the time that some action or event took place (yesterday, Monday).
- **proper noun (PROPN)**: name of place, person etc.
- **interjection (INTJ)**: exclamation, greeting, yes/no response, etc.

The closed class words include:

- **preposition/postposition (ADP)**: mark a noun's spacial, temporal, or other relation
- **pronoun (PRON)**: a shorthand for referring to an entity or event such as *I, me, he, she, you*. **Wh-pronouns** (*what, who, whom, whoever*) are used in certain question forms, or act as **complementizers** (*Frida, who married Diego...*).
- **auxiliary (AUX)**: such as *can, make, need, may, should* etc. Auxiliary verbs mark *semantic* features of a main verb such as its tense, whether it is completed (aspect), whether it is negated (polarity), and whether an action is necessary, possible, suggested, or desired (mood). English auxiliaries include the **copula verb** *be*, the two verbs *do* and *have*, forms, as well as **modal verbs** used to mark the mood associated with the event depicted by the main verb: *can* indicates ability or possibility, *may* permission or possibility, *must* necessity.
- **determiner (DET)**: marks noun phrase properties such as *this, that, a, an, the*.
- **coordinating conjunction (CCONJ)**: such as *and, but, however,*
- **subordinating conjunction (SCONJ)**: such as *which, where, that etc.* Subordinating conjunctions like *that which* link a verb to its argument in this way are also called **complementizers**
- **numeral (NUM)**: numbers, *one, two, three* etc.
- **particle (PART)**: a proposition-like form used together with verb, *down, up, off, on, at, by, out, in*. A particle resembles a preposition or an adverb and is used in combination with a verb. Particles often have **extended meanings** that arent quite the same as the prepositions they resemble, as in the particle *over* in *she turned the paper over*. A verb and a particle acting as *a single unit* is called a **phrasal verb**. The meaning of phrasal verbs is often **non-compositional**not predictable from the individual meanings of the verb and the

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coord. conj.	*and, but, or*	NNP	proper noun, sing.	*IBM*	TO	"to"	*to*
CD	cardinal number	*one, two*	NNPS	proper noun, plu.	*Carolinas*	UH	interjection	*ah, oops*
DT	determiner	*a, the*	NNS	noun, plural	*llamas*	VB	verb base	*eat*
EX	existential 'there'	*there*	PDT	predeterminer	*all, both*	VBD	verb past tense	*ate*
FW	foreign word	*mea culpa*	POS	possessive ending	*'s*	VBG	verb gerund	*eating*
IN	preposition/ subordin-conj	*of, in, by*	PRP	personal pronoun	*I, you, he*	VBN	verb past participle	*eaten*
JJ	adjective	*yellow*	PRP$	possess. pronoun	*your, one's*	VBP	verb non-3sg-pr	*eat*
JJR	comparative adj	*bigger*	RB	adverb	*quickly*	VBZ	verb 3sg pres	*eats*
JJS	superlative adj	*wildest*	RBR	comparative adv	*faster*	WDT	wh-determ.	*which, that*
LS	list item marker	*1, 2, One*	RBS	superlatv. adv	*fastest*	WP	wh-pronoun	*what, who*
MD	modal	*can, should*	RP	particle	*up, off*	WP$	wh-possess.	*whose*
NN	sing or mass noun	*llama*	SYM	symbol	+,%, &	WRB	wh-adverb	*how, where*

Figure 8.2 Penn Treebank part-of-speech tags.

Figure 2: The part-of-speech tag in Penn Treebank.

particle.

Closed class words are generally **function words** like *of, it, and, or you*, which tend to be very short, occur frequently, and often have *structuring* uses in grammar.

Figure 2 shows the symbol of part-of-speech tag from Penn Treebank [Jurafsky and Martin, 2014].

2 Part-of-speech tagging

Part-of-speech tagging is the process of assigning a part-of-speech to each word in a text. The input is a sequence $x_1, x_2, \ldots x_n$ of (tokenized) words and a tagset, and the output is a sequence $y_1, y_2, \ldots, y_n$ of tags, each output y_i corresponding exactly to one input x_i, as shown in the intuition in Figure 3.

Tagging is a **disambiguation** task; words are ambiguous have more than one possible part-of-speech and the goal is to find the correct tag for the situation. The goal of POS-tagging is to resolve these ambiguities, choosing the proper tag for the context (i.e. **ambiguity resolution**). Nonetheless, many words are easy to disambiguate, because their different tags arent equally likely. For example, "a" can be a determiner or the letter a, but the determiner sense is much more likely.

This idea suggests a useful **baseline**: given an ambiguous word, choose the tag which is most **frequent** in the training corpus. This is a key concept:

Most Frequent Class Baseline*: Always compare a classifier against a baseline at least as good as the most frequent class baseline (assigning each token to the class it occurred in most often in the training set).*

The most-frequent-tag baseline has an accuracy of about 92%. The standard technique behinds POS tagging is **sequence labeling**, i.e. for an input sequence of texts, output a sequence of labels,

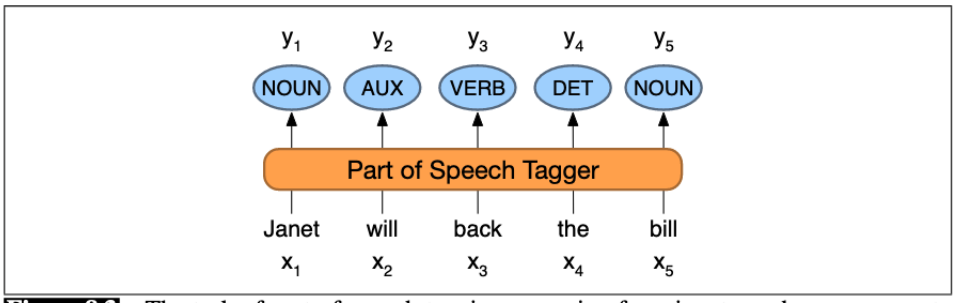

Figure 8.3 The task of part-of-speech tagging: mapping from input words $x_1, x_2, ..., x_n$ to output POS tags $y_1, y_2, ..., y_n$.

Figure 3: The part-of-speech tagging process.

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

Figure 4: An example of name entities in a paragraph.

each for a text.

3 Named Entities and Named Entity Tagging

A **named entity** is, roughly speaking, anything that can be referred to with a proper name: a *person*, a *location*, an *organization*. The task of **named entity recognition (NER)** is to find *spans* of text that constitute proper names and tag the type of the entity. **Four** entity tags are most **common**: **PER (person)**, **LOC (location)**, **ORG (organization)**, or **GPE (geo-political entity)**. However, the term named entity is commonly extended to include things that arent entities per se, including **dates**, **times**, and other kinds of **temporal expressions**, and even **numerical expressions** like prices. Figure 4 shows name entities in a paragraph.

Named Entity Recognition can be used as pre-processing steps for *sentiment analysis*, *question answering* as well as learning *semantic representations* such as event extraction and relationship inference.

NER is more challenging than POS tagging. For POS tagging, each word has only one part-of-speech tag but for NER, there is a *segmentation* problem. The task of NER is to find and label the *span* of text. The ambiguity of segmentation brings additional challenge: we need to decide whats an entity and what isnt, and where the boundaries are.

The standard approach that convert the span-recognition to sequence labeling is **BIO tagging** [Jurafsky and Martin, 2014]. This is a method that allows us to treat NER like a word-by-word sequence labeling task, via **tags** that *capture both the boundary and the named entity type*. Similarly,

Words	IO Label	BIO Label	BIOES Label
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O

Figure 8.7 NER as a sequence model, showing IO, BIO, and BIOES taggings.

Figure 5: BIO, IO and BIOES tagging for NER.

there are **IO tagging** and **BIOES tagging**.

In BIO tagging we label any token that begins a span of interest with the label **B**, tokens that occur inside a span are tagged with an **I**, and any tokens outside of any span of interest are labeled **O**. There is only *one* **O** tag to indicate not a named entity. On the other hand, we can have distinct **B** and **I** for each named entity type, i.e. **B-PER**, I-PER, B-ORG, I-ORG etc. For n distinct named entity types, there are $2n+1$ distinct tags. BIO tagging can represent exactly the same information as the bracketed notation, but has the advantage that we can represent the task in the same simple sequence modeling way as part-of-speech tagging. In BIOES tagging, we add **E** tag for end of a span, and a span tag **S** for a span consist of only one word.

A **sequence labeler** (HMM, CRF, RNN, Transformer, etc.) is trained to label each token in a text with tags that indicate the presence (or absence) of particular kinds of named entities.

References

Dan Jurafsky and James H Martin. Speech and language processing. vol. 3. *US: Prentice Hall*, 2014.