Sequence Labeling for Part of Speech and Named Entities

Part of Speech Tagging
Parts of Speech

From the earliest linguistic traditions (Yaska and Panini 5th C. BCE, Aristotle 4th C. BCE), the idea that words can be classified into grammatical categories

- part of speech, word classes, POS, POS tags

8 parts of speech attributed to Dionysius Thrax of Alexandria (c. 1st C. BCE):

- noun, verb, pronoun, preposition, adverb, conjunction, participle, article

- These categories are relevant for NLP today.
Two classes of words: Open vs. Closed

**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*
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

Closed class ("function")

Determiners: the, some

Conjunctions: and, or

Pronouns: they, its

Auxiliary
- can, had

Prepositions
- to, with

Particles
- off, up

... more
Part-of-Speech Tagging

Assigning a part-of-speech to each word in a text. Words often have more than one POS.

book:
- VERB: *(Book that flight)*
- NOUN: *(Hand me that book)*.
Part-of-Speech Tagging

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

\[
y_1 \\
\text{NOUN} \\
\uparrow \\
\text{Janet} \\
x_1
\]
\[
y_2 \\
\text{AUX} \\
\uparrow \\
\text{will} \\
x_2
\]
\[
y_3 \\
\text{VERB} \\
\uparrow \\
\text{back} \\
x_3
\]
\[
y_4 \\
\text{DET} \\
\uparrow \\
\text{the} \\
x_4
\]
\[
y_5 \\
\text{NOUN} \\
\uparrow \\
\text{bill} \\
x_5
\]

Part of Speech Tagger
8.1 (Mostly) English Word Classes

Until now we have been using part-of-speech terms like *noun* and *verb* rather freely. In this section we give more complete definitions. While word classes do have semantic tendencies—*adjectives*, for example, often describe *properties* and *nouns*—parts of speech are defined instead based on their grammatical relationship with neighboring words or the morphological properties about their affixes.

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

There/PRO were/VERB 70/NUM children/NOUN there/ADV ./PUNC

Preliminary/ADJ findings/NOUN were/AUX reported/VERB in/ADP today/NOUN ’s/PART New/PROP new England/PROP Journal/PROP of/ADP Medicine/PROP
Why Part of Speech Tagging?

- Can be useful for other NLP tasks
  - Parsing: POS tagging can improve syntactic parsing
  - MT: reordering of adjectives and nouns (say from Spanish to English)
  - Sentiment or affective tasks: may want to distinguish adjectives or other POS
  - Text-to-speech (how do we pronounce “lead” or "object"?)
- Or linguistic or language-analytic computational tasks
  - Need to control for POS when studying linguistic change like creation of new words, or meaning shift
  - Or control for POS in measuring meaning similarity or difference
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 performance in English

How many tags are correct? (Tag accuracy)

- About 97%
  - Hasn't changed in the last 10+ years
  - HMMs, CRFs, BERT perform similarly.
  - Human accuracy about the same

But baseline is 92%!

- Baseline is performance of stupidest possible method
  - "Most frequent class baseline" is an important baseline for many tasks
    - Tag every word with its most frequent tag
    - (and tag unknown words as nouns)

- Partly easy because
  - Many words are unambiguous
Sources of information for POS tagging

Janet will back the bill

AUX/NOUN/VERB? NOUN/VERB?

Prior probabilities of word/tag
- "will" is usually an AUX

Identity of neighboring words
- "the" means the next word is probably not a verb

Morphology and wordshape:
- Prefixes: unable: un- → ADJ
- Suffixes: importantly: -ly → ADJ
- Capitalization: Janet: CAP → PROPN
Standard algorithms for POS tagging

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

All required a hand-labeled training set, all about equal performance (97% on English)

All make use of information sources we discussed
- Via human created features: HMMs and CRFs
- Via representation learning: Neural LMs
Part of Speech Tagging
Sequence Labeling for Part of Speech and Named Entities

Named Entity Recognition (NER)
Named Entities

- **Named entity**, in its core usage, means anything that can be referred to with a proper name. Most common 4 tags:
  - **PER** (Person): “Marie Curie”
  - **LOC** (Location): “New York City”
  - **ORG** (Organization): “Stanford University”
  - **GPE** (Geo-Political Entity): "Boulder, Colorado"
- Often multi-word phrases
- But the term is also extended to things that aren't entities:
  - dates, times, prices
Named Entity tagging

The task of named entity recognition (NER):
- find spans of text that constitute proper names
- tag the type of the entity.
The most-frequent-tag baseline has an accuracy of about 92%. The baseline thus differs from the state-of-the-art and human ceiling (97%) by only 5%.

8.3 Named Entities and Named Entity Tagging

Part of speech tagging can tell us that words like Janet, Stanford University, and Colorado are all proper nouns; being a proper noun is a grammatical property of these words. But viewed from a semantic perspective, these proper nouns refer to different kinds of entities: Janet is a person, Stanford University is an organization, and Colorado is a location.

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 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 aren't entities per se, including dates, times, and other kinds of temporal expressions, and even numerical expressions like prices. Here's an example of the output of an NER tagger:

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].

<table>
<thead>
<tr>
<th>Type Tag</th>
<th>Sample Categories</th>
<th>Example sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>PER</td>
<td>Turing is a giant of computer science.</td>
</tr>
<tr>
<td>Organization</td>
<td>ORG</td>
<td>The IPCC warned about the cyclone.</td>
</tr>
<tr>
<td>Location</td>
<td>LOC</td>
<td>Mt. Sanitas is in Sunshine Canyon.</td>
</tr>
<tr>
<td>Geo-Political Entity</td>
<td>GPE</td>
<td>Palo Alto is raising the fees for parking.</td>
</tr>
</tbody>
</table>

Figure 8.5 A list of generic named entity types with the kinds of entities they refer to. Named entity tagging is a useful first step in lots of natural language understanding tasks. In sentiment analysis we might want to know a consumer's sentiment toward a particular entity. Entities are a useful first stage in question answering, or for linking text to information in structured knowledge sources like Wikipedia. And named entity tagging is also central to natural language understanding tasks of building semantic representations, like extracting events and the relationship between participants.
Why NER?

Sentiment analysis: consumer’s sentiment toward a particular company or person?

Question Answering: answer questions about an entity?

Information Extraction: Extracting facts about entities from text.
Why NER is hard

1) Segmentation
   • In POS tagging, no segmentation problem since each word gets one tag.
   • In NER we have to find and segment the entities!

2) Type ambiguity

[PER Washington] was born into slavery on the farm of James Burroughs.
[ORG Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [LOC Washington] for what may well be his last state visit.
In June, [GPE Washington] passed a primary seatbelt law.
BIO Tagging

How can we turn this structured problem into a sequence problem like POS tagging, with one label per word?

[PER Jane Villanueva] of [ORG United], a unit of [ORG United Airlines Holding], said the fare applies to the [LOC Chicago] route.
BIO Tagging


<table>
<thead>
<tr>
<th>Words</th>
<th>BIO Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane Villanueva</td>
<td>B-PER</td>
</tr>
<tr>
<td>of United Airlines Holding discussed the Chicago route</td>
<td>I-PER, O, O, B-LOC</td>
</tr>
</tbody>
</table>

Now we have one tag per token!!!
BIO Tagging

B: token that *begins* a span
I: tokens *inside* a span
O: tokens outside of any span

# of tags (where n is #entity types):
1 O tag,
n B tags,
n I tags
total of $2n+1$

<table>
<thead>
<tr>
<th>Words</th>
<th>BIO Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>B-PER</td>
</tr>
<tr>
<td>Villanueva</td>
<td>I-PER</td>
</tr>
<tr>
<td>of</td>
<td>O</td>
</tr>
<tr>
<td>United Airlines</td>
<td>B-ORG</td>
</tr>
<tr>
<td>Holding</td>
<td>I-ORG</td>
</tr>
<tr>
<td>discussed</td>
<td>I-ORG</td>
</tr>
<tr>
<td>the</td>
<td>O</td>
</tr>
<tr>
<td>Chicago route</td>
<td>O</td>
</tr>
<tr>
<td>.</td>
<td>B-LOC</td>
</tr>
<tr>
<td></td>
<td>O</td>
</tr>
</tbody>
</table>
BIO Tagging variants: IO and BIOES

[PER Jane Villanueva] of [ORG United], a unit of [ORG United Airlines Holding], said the fare applies to the [LOC Chicago] route.

<table>
<thead>
<tr>
<th>Words</th>
<th>IO Label</th>
<th>BIO Label</th>
<th>BIOES Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>I-PER</td>
<td>B-PER</td>
<td>B-PER</td>
</tr>
<tr>
<td>Villanueva</td>
<td>I-PER</td>
<td>I-PER</td>
<td>E-PER</td>
</tr>
<tr>
<td>of</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>United</td>
<td>I-ORG</td>
<td>B-ORG</td>
<td>B-ORG</td>
</tr>
<tr>
<td>Airlines</td>
<td>I-ORG</td>
<td>I-ORG</td>
<td>I-ORG</td>
</tr>
<tr>
<td>Holding</td>
<td>I-ORG</td>
<td>I-ORG</td>
<td>E-ORG</td>
</tr>
<tr>
<td>discussed</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>the</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>Chicago</td>
<td>I-LOC</td>
<td>B-LOC</td>
<td>S-LOC</td>
</tr>
<tr>
<td>route</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>.</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>
Standard algorithms for NER

Supervised Machine Learning given a human-labeled training set of text annotated with tags

• Hidden Markov Models
• Conditional Random Fields (CRF)/ Maximum Entropy Markov Models (MEMM)
• Neural sequence models (RNNs or Transformers)
• Large Language Models (like BERT), finetuned
Named Entity Recognition (NER)