

Parts of Speech and Sequence Tagging

CS-585

Natural Language Processing

Sonjia Waxmonsky

Sequence labeling

- We can treat a text as a bunch of lexical units (a bag of words)
- ...Or we can also treat it as a *sequence*
 - Typically for shorter texts - e.g., one sentence at a time
- Similarly, we can make predictions at the text level, or we can **make predictions associated with each word in the text**, incorporating information from the surrounding sequence
- These are ***sequence labeling*** tasks
 - Very common in NLP, but also in bioinformatics

Sequence labeling

- Sequence labeling involves associating a predicted label or tag with each word

Part-of-Speech (POS) tags:

AT0 AJ0 AJ0 NN1 VVD PRP AT0 AJ0 NN1

The quick brown fox jumped over the lazy dog

Language labels:

ES ES ES ES ES EN EN EN

bueno porque le gusta decir hi good morning

Sequences and Text Spans

- Sequence labeling methods can also be used to label *spans* of text, by specifying word-level tags that encode whether the word is inside or outside of the span

Named Entity recognition:

Place

Thank you for visiting the [Chicago Public Library] .

O O O O O I I I

- Problem – this is ambiguous

Place

Place

Thank you for visiting the [Chicago] [Public Library] .

IOB Representations

Tag type	Use
I	The word is <i>inside</i> the span (and more specific tags do not apply)
O	The word is <i>outside</i> the span
B	The word is at the <i>beginning</i> of the span
E/L	The word is at the <i>end</i> of the span (the <i>last</i> word)
S/U	The word is a <i>singleton / unit</i> span

IOB

○ ○ ○ ○ ○ B-Place I-Place I-Place ○
Thank you for visiting the Chicago Public Library .

I-Place ○ ○ ○ ○ B-Place I-Place ○
Albany is the capital of New York .

PART OF SPEECH TAGGING

Parts of Speech

- Generally speaking, the “grammatical type” of word:
 - Verb, Noun, Adjective, Adverb, Article, ...
- We also include inflections:
 - Verbs: Tense, number, ...
 - Nouns: Number, proper/common, ...
 - Adjectives: comparative, superlative, ...
 - ...

POS tagging key for NLP

- Ambiguity: book *can be a noun or a verb*
- Useful input for parsing
 - “*book a ticket*” vs. “*book of the century*”
- Disambiguate meaning even without parsing
 - *Is a text about literature or about travel?*

Tagsets

- *Tagsets*: canonical definitions of parts of speech
 - May be more or less granular
 - Typically language-specific (although there are universals – nouns, verbs, ...)

Tagset	Language	Number of Tags
CLAWS8	English	170
Penn Treebank	English	36
BNC/C5	English	57
Universal POS tags	“universal”	17
Stuttgart Tübingen Tagset	German	54
Chinese PTB	Chinese	34
...		

BNC Parts of Speech

Nouns:

NN0 Common noun, neutral for number (e.g. *aircraft*)

NN1 Singular common noun (e.g. *pencil, goose, time*)

NN2 Plural common noun (e.g. *pencils, geese, times*)

NP0 Proper noun (e.g. *London, Michael, Mars, IBM*)

Pronouns:

PN1 Indefinite pronoun (e.g. *none, everything, one*)

PNP Personal pronoun (e.g. *I, you, them, ours*)

PNQ *Wh*-pronoun (e.g. *who, whoever, whom*)

PNX Reflexive pronoun (e.g. *myself, itself, ourselves*)

Verbs:

VVB finite base form of lexical verbs (e.g. *forget, send, live, return*)

VVD past tense form of lexical verbs (e.g. *forgot, sent, lived*)

VVG *-ing* form of lexical verbs (e.g. *forgetting, sending, living*)

VVI infinitive form of lexical verbs (e.g. *forget, send, live, return*)

VVN past participle form of lexical verbs (e.g. *forgotten, sent, lived*)

VVZ *-s* form of lexical verbs (e.g. *forgets, sends, lives, returns*)

VBB present tense of BE, except for *is*

...and so on: VBD VBG VBI VBN VBZ

VDB finite base form of DO: *do*

...and so on: VDD VDG VDI VDN VDZ

VHB finite base form of HAVE: *have, 've*

...and so on: VHD VHG VHI VHN VHZ

VM0 Modal auxiliary verb (e.g. *will, would, can, could, 'll, 'd*)

Articles

AT0 Article (e.g. *the, a, an, no*)

DPS Possessive determiner (e.g. *your, their, his*)

DT0 General determiner (*this, that*)

DTQ *Wh*-determiner (e.g. *which, what, whose, whichever*)

EX0 Existential *there*, i.e. occurring in “***there*** is...” or “***there*** are...”

Adjectives

AJ0 Adjective (general or positive) (e.g. *good, old, beautiful*)

AJC Comparative adjective (e.g. *better, older*)

AJS Superlative adjective (e.g. *best, oldest*)

Adverbs

AV0 General adverb (e.g. *often, well, longer (adv.), furthest*.)

AVP Adverb particle (e.g. *up, off, out*)

AVQ *Wh*-adverb (e.g. *when, where, how, why, wherever*)

Miscellaneous:

CJC Coordinating conjunction (e.g. *and, or, but*)

CJS Subordinating conjunction (e.g. *although, when*)

CJT The subordinating conjunction *that*

CRD Cardinal number (e.g. *one, 3, fifty-five, 3609*)

ORD Ordinal numeral (e.g. *first, sixth, 77th, last*)

ITJ Interjection or other isolate (e.g. *oh, yes, mhm, wow*)

POS The possessive or genitive marker 's or '

TOO Infinitive marker *to*

PUL Punctuation: left bracket - i.e. (or [

PUN Punctuation: general separating mark - i.e. . , ! , : ; - or ?

PUQ Punctuation: quotation mark - i.e. ' or ”

PUR Punctuation: right bracket - i.e.) or]

XX0 The negative particle *not* or *n't*

ZZ0 Alphabetical symbols (e.g. *A, a, B, b, c, d*)

Task: Part-Of-Speech Tagging

- **Goal:** Assign the correct part-of-speech to each word (and punctuation) in a text.
- Example:

Two	old	men	bet	on	the	game	.
CRD	AJ0	NN2	VVD	PP0	AT0	NN1	PUN

- Learn a *local* model of POS dependencies, usually from pretagged data
- **No** parsing!

Modeling considerations for POS tagging

- POS tagging is very **similar** to word-sense disambiguation, in that it involves making predictions at the word level, rather than the text level
 - One difference is that we need to assign a tag to *every* word
 - Another is that the predictive information tends to be much more “local”
- Not a bag-of-words problem!
 - Contextual features rather than document-level lexical features

Modeling considerations for POS tagging

- POS tagging (and sequence labeling in general) involves using
 - Bottom-up evidence related to the likelihood of tags
 - Top-down information about what tag sequences are valid/likely

PNP I ADV0 VVD AT0 NN0
only saw the one three
CRD CRD NN0

"three" is usually a cardinal number

Pronouns rarely occur after articles

Nouns occur more often than pronouns or numbers after "the"

Nouns occur frequently after cardinal numbers

Global constraints for sequence labeling

○ ○ ○ ○ ○ B-Place I-Place I-Place ○
Thank you for visiting the Chicago Public Library .

○ ○ ○ ○ ○ I-Place I-Place B-Place ○
Thank you for visiting the Chicago Public Library .

○ ○ ○ ○ ○ B-Place I-Company I-Place ○
Thank you for visiting the Chicago Public Library .

Evaluation

- Part-of-speech tagging is “easy”
 - Just guessing the most frequent tag for each word (and “noun” for unknown words) gets over 90% accuracy*
 - As always, it’s important to evaluate accuracy measures relative to some baseline
- POS tagging accuracy can suffer when models are applied to a different domain or genre than that used for training. E.g., performance for a model trained on Wall Street Journal text :

Domain	WSJ	news-groups	reviews	blogs	answers	emails
Accuracy	96.8%	89.1%	91.4%	94.1%	88.9%	88.7%

*Jurafsky & Martin. *Speech and Language Processing*.

**Schnabel & Schutze (2013). *FLORS: Fast and Simple Domain Adaptation for Part-of-Speech Tagging*.

Evaluation

- *Accuracy* is a high-level measure of model performance for classification tasks, aligned with the loss function used for optimization (generally, cross-entropy)
 - *But it doesn't tell us how the model is doing on specific categories*
 - *And as we have seen, it can be misleading in cases with very unbalanced class distributions or simple baselines – e.g., a mediocre model gives us 90%, a great model give us 92%*
- Applies for a range of tasks (sequence labeling, text categorization, etc.)

Evaluation: confusion matrix

A tool to provide more insight into model performance is the confusion matrix: a table comparing the model's predictions to the true labels class by class

Sample confusion matrix for **sentiment**:

		Predicted class		
		Positive	Neutral	Negative
True class	Positive	100	50	5
	Neutral	40	150	30
	Negative	10	60	90

Total correct = $100+150+90=340$

Total items = $100+50+5+40+150+30+10+60+90=535$

Accuracy = $340/535=63.6\%$

Evaluation: precision

Precision: The proportion of predictions for a given class that are correct

$$\text{Precision} = \text{True Positives} \div (\text{True Positives} + \text{False Positives})$$

For Neutral sentiment:

		Predicted class		
		Positive	Neutral	Negative
True class	Positive	100	50	5
	Neutral	40	150	30
	Negative	10	60	90

$$\text{Precision} = 150 \div (150 + 50 + 60) = 57.7\%$$

Evaluation: recall

Recall: The proportion of true labels for a given class that are correctly predicted

$$\text{Recall} = \text{True Positives} \div (\text{True Positives} + \text{False Negatives})$$

For Neutral sentiment:

		Predicted class		
		Positive	Neutral	Negative
True class	Positive	100	50	5
	Neutral	40	150	30
	Negative	10	60	90

$$\text{Recall} = 150 \div (150 + 40 + 30) = 68.2\%$$

Evaluation: F measure

F-measure is a balanced weighting of precision and recall, combining the two using the harmonic mean:

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R}$$

Technically, we refer to it as F_1 , because there are a family of F measures with different weighting of precision vs. recall:

$$F_\beta = (1 + \beta^2) \frac{PR}{\beta^2 P + R}$$

$\beta > 1$ gives more importance to recall; $0 < \beta < 1$ gives more importance to precision

So for $P=57.7\%$, $R=68.2\%$: $F_1 = \frac{2 \times .577 \times .682}{.577 + .682} = .625$

Confusion matrix for POS tagging

	NN0	NN1	NN2	NP0	PNI	PNP	PNQ	PNX	...
NN0									
NN1									
NN2									
NP0									
PNI									
PNP									
PNQ									
PNX									
VVB									
VVD									
...									

Inference and search for sequence modeling

- Note that our search space has blown up, relative to text categorization tasks
 - For text categorization tasks, we might have to choose one of 2 classes for a text, or one of 200
 - For POS tagging, we might have to select a tag from a list of ~ 80 , for each word in a sentence $\rightarrow 80^N$ possible tag assignments for a sentence of length N !
 - Generally for sequence labeling tasks, the search space is exponential in the length of the sequence
- Of course, most of these tag combinations are very unlikely.

Inference and search for sequence modeling

- We can make the search tractable in a few ways
 - Greedy search: commit to tag assignments one by one, and use them as context for the remaining assignments
 - Beam search: consider only a limited number of hypotheses for partial tag assignments, and discard the rest
 - Dynamic programming: store intermediate results in a data structure to reduce backtracking and transform the exponential search into a quadratic one

Brill tagger

- We'll get to probabilistic approaches to POS tagging next lecture (HMMs, Viterbi)
- Alternate approach: **Brill tagger** (1993)
 - Example of ***transformation-based learning***
 - Iteratively apply rules to change incorrect labels

Transformation-based learning

- Idea: to get a good model
 1. Start with a mediocre model
 2. Identify the most common errors it makes
 3. Fix them
 4. Go to 2

