Part-of-speech Tagging

Natural Language Processing

Master in Business Analytics and Big Data

acastellanos@faculty.ie.edu

Parts of Speech

Open class (lexical) words

- Nouns
  - Proper: *IBM*, *Italy*
  - Common: *cat / cats*, *snow*
- Verbs
  - Main: *see*, *registered*
  - Modals: *can*, *had*
- Adjectives *old older oldest*
- Adverbs *slowly*
- Numbers *122,312*, *one*
- *… more*

Closed class (functional)

- Determiners *the some*
- Conjunctions *and or*
- Pronouns *he its*
- Prepositions *to with*
- Particles *off up*
- Interjections *Ow Eh*
- *… more*

Parts of Speech

Brown/Penn Treebank tags

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*			

POS Tagging

- Words often have **more than one POS**: *back*
  - The *back* door = JJ
  - On my *back* = NN
  - Win the voters *back* = RB
  - Promised to *back* the bill = VB

- The POS tagging problem is **to determine the POS tag for a particular instance of a word**.
  - **Disambiguation Problem**
  - **You need the whole sentence**

POS Tagging

Plays	**well**	**with**	**others**
NNS/VBZ	UH/JJ/NN/RB	IN	NNS
Plays/VBZ	well/RB	with/IN	others/NNS

- **Uses:**
  - **Text-to-speech** (how do we pronounce "lead"?)
  - Can **write regexps** like (Det) Adj* N+ over the output for phrases to detect multiword expressions.
  - As input to or to speed up a **full syntax parser**
  - If you know the tag, you can **back off to it in other tasks**
    - **NER → focus on NN**
    - **Sentiment Analysis → Focus on ADJ**

POS Tagging

- **Probabilistic Modeling**

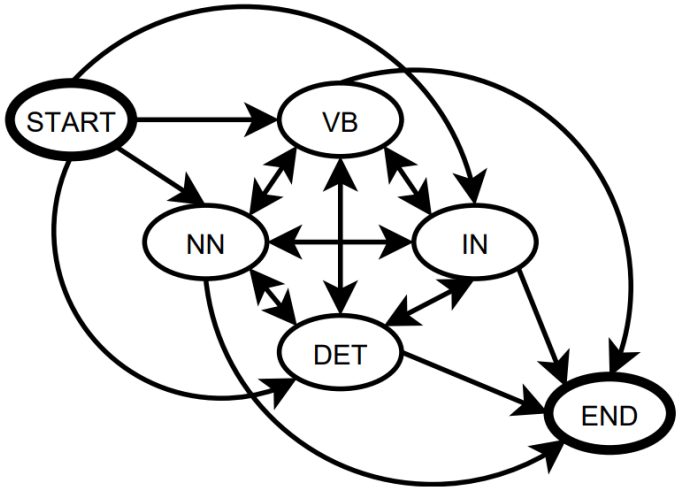
$$P(y \mid x; \beta) = \frac{\exp\left(x^\top \beta_y\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(x^\top \beta_{y'}\right)}$$

$$P(\mathrm{VBZ} \mid flies) = \frac{\exp\left(x^\top \beta_{\mathrm{VBZ}}\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(x^\top \beta_{y'}\right)}$$

POS Tagging

- **Probabilistic Modeling**

$$\text{argmax}_T \; P(T|S) = \prod_i P(w_i|t_i) \; P(t_i|t_{i-1})$$



POS Tagging Performance

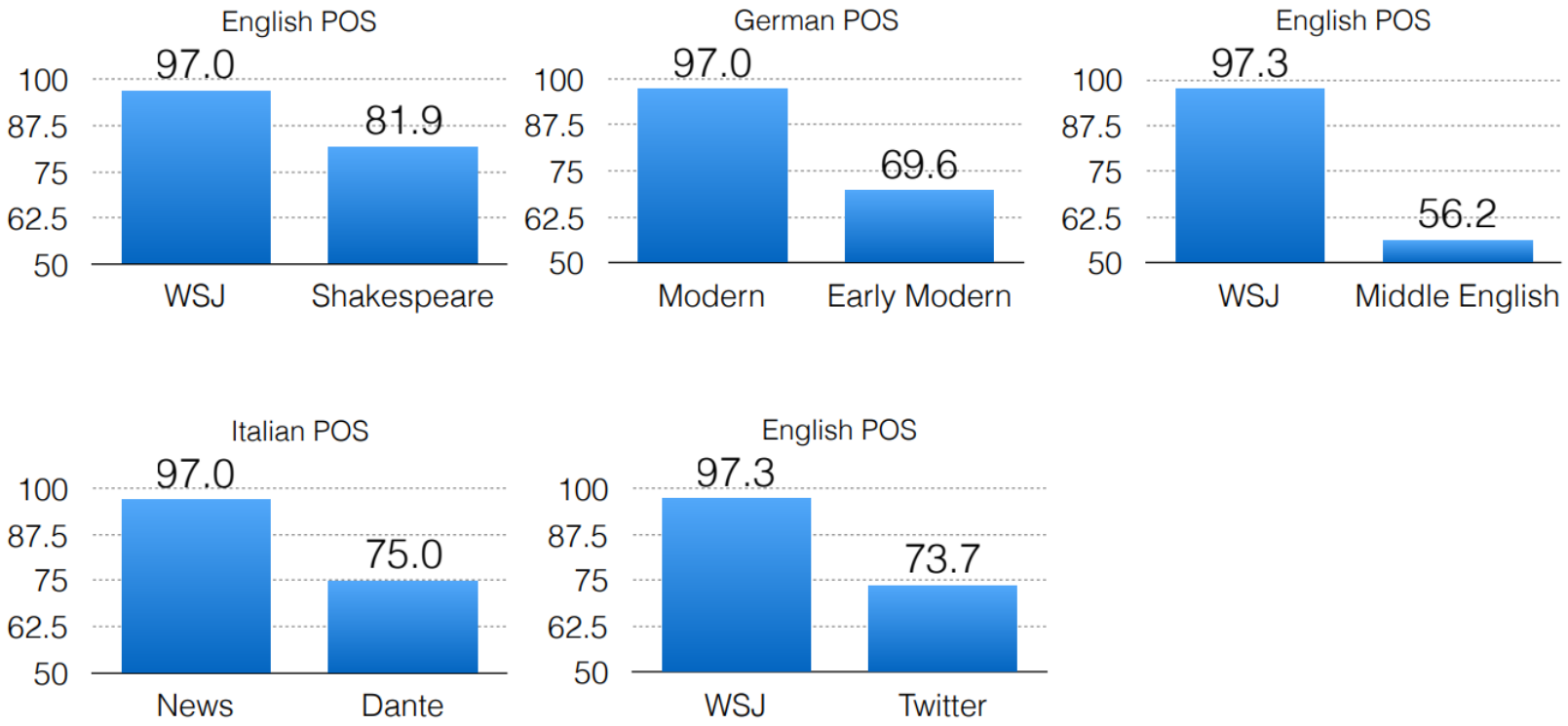
- How many tags are correct? (Tag accuracy)
  - **SOFTA**: about **94-97%** currently
  - Partly **easy** because
    - Many words are unambiguous
    - You get points for them (*the*, *a*, etc.) and for punctuation marks!

***Remember form ML2: Accuracy is sometimes misleading**

POS Tagging Performance

Peen Treebank (45-tag corpus)		
Unambiguous (1 tag)	38,857	(81%)
Ambiguous (2-7 tags)	8,844	(19%)
Details: 2 tags	6,731	
3 tags	1,621	
4 tags	357	
5 tags	90	
6 tags	32	
7 tags	6	*(well, set, round, open, fit, down)*
8 tags	4	*('s, half, back, a)*
9 tags	3	*(that, more, in)*

POS Tagging Performance



http://people.ischool.berkeley.edu/~dbamman/nlp18.html

POS Tagging Performance

Lexicon gap	4.5%	A 60% slash/NN the common stock dividend
Unknown word	4.5%	blaming the disaster on substandard/JJ construction
Could plausibly get right	16.0%	market players overnight/RB in Tokyo began bidding up oil prices
Difficult linguistics	19.5%	They set/VBP up absurd situations, detached from r eality
Underspecified/unclear	12.0%	a $ 10 million fourth quarter charge against/IN disco ntinued/JJ operations
Inconsistent/no standard	28.0%	Orson Welles 's Mercury Theater in the '30s/NNS
Gold standard wrong	15.5%	Our market got hit/VB a lot harder on Monday than the listed market

Source: https://nlp.stanford.edu/~manning/papers/CICLing2011-manning-tagging.pdf

POS Tagging Performance

- How many tags are correct? (Tag accuracy)
  - **SOFTA**: about **94-97%** currently
  - Partly **easy** because
    - Many words are unambiguous
    - You get points for them (*the*, *a*, etc.) and for punctuation marks!

      ***Remember form ML2: Accuracy is sometimes misleading**
  - **Baseline**
    - Tag every word with its most frequent tag
    - Tag unknown words as nouns

90%

Improving Performance

- Main sources of information for POS tagging:
  - Knowledge of **neighboring words**
    - Bill saw that man yesterday

      NNP NN DT NN NN

      VB VB(D) IN VB NN
  - Knowledge of **word probabilities**
    - *man* is rarely used as a verb….

- The latter proves the most useful, but the former also helps

Model	Features	Token	Unknown	Sentence
Baseline	56,805	**93.69%**	82.61%	26.74%
3Words	239,767	**96.57%**	86.78%	48.27%

Improving Performance

- **Create new word-based features**
  - Word — the: the → DT
  - Lowercased word — Importantly: importantly → RB
  - Prefixes — unfathomable: un- → JJ
  - Suffixes — Importantly: -ly → RB
  - Capitalization — Meridian: CAP → NNP
  - Word shapes — 35-year: d-x → JJ
- Then build a model to predict tag
  - $P(T|w)$: 93.7% overall / 82.6% unknown

NLTK POS Tagging

```
import nltk
text = nltk.word_tokenize("Bill    saw     that  man yesterday")
nltk.pos_tag(text)
```

```
[('Bill', 'NNP'),
 ('saw', 'VBD'),
 ('that', 'IN'),
 ('man', 'NN'),
 ('yesterday', 'NN')]
```

Parsey McParseface

- Based on Google's SyntaxNet
- The World's Most Accurate Parser (according to Google)
  - 94% Accuracy on **Penn Treebank**
  - **Human performance** expected to be around 96%
- https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html
- https://algorithmia.com/algorithms/deeplearning/Parsey

Additional Resources

- **Implement your own parser**
  - https://nelsonmanohar.wordpress.com/2015/07/08/a-part-of-speech-2nd-order-classification-taggger/
- **Stanford POS Tagger**
  - https://nlp.stanford.edu/software/tagger.shtml
- **State of the art review**
  - https://arxiv.org/ftp/arxiv/papers/1708/1708.00241.pdf
- **Twitter Part-of-Speech Tagging**
  - http://www.derczynski.com/sheffield/papers/twitter_pos.pdf
- **NLTK Tagged Corpora for Training**
  - http://www.nltk.org/howto/corpus.html#tagged-corpora