LING439/539 - Statistical NLP
Chapter 10. Part-of-speech tagging

September 13-15 2016

The ultimate goal of research on natural language processing is to *parse* and *understand* language.

$\rightarrow$ Still far from achieving the goal

Much research in NLP has focused on **intermediate** tasks...

Part-of-speech tagging

POS tagsets:

- Brown POS tagset
- Penn POS tagset
- Universal POS tagset
- ...

Universal POS tags (2012)

A set of 12 universal part-of-speech tags:

VERB	- verbs (all tenses and modes)
NOUN	- nouns (common and proper)
PRON	- pronouns
ADJ	- adjectives
ADV	- adverbs
ADP	- adpositions (prepositions and postpositions)
CONJ	- conjunctions
DET	- determiners
NUM	- cardinal numbers
PRT	- particles or other function words
X	- other: foreign words, typos, abbreviations
.	- punctuation

Slav Petrov, Dipanjan Das and Ryan McDonald (2012). A Universal Part-of-Speech Tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, `https://github.com/slavpetrov/universal-pos-tags`

Universal POS tags (2016) from Universal Dependencies

A set of 17 universal part-of-speech tags:

VERB	- verbs (all tenses and modes)
AUX	- auxiliary verb
NOUN	- nouns (common)
PROPN	- proper noun
PRON	- pronouns
ADJ	- adjectives
ADV	- adverbs
ADP	- adpositions (prepositions and postpositions)
CONJ	- conjunctions
DET	- determiners
NUM	- cardinal numbers
PRT → **PART**	- particles or other function words
INTJ	- interjection
SCONJ	- subordinating conjunction
SYM	- symbol
X	- other: foreign words, typos, abbreviations
. → **PUNCT**	- punctuation

See `http://universaldependencies.org/u/pos`

POS tagging approaches

- rule-based tagging
- transformation-based tagging (Brill tagger)
- any sequence labeling algorithms...
  - HMM
  - ME
  - CRFs

POS tagging resources

- Scottish Gaelic http://datashare.is.ed.ac.uk/handle/10283/2011
- Tamil http://au-kbc.org/nlp/corpusrelease.html (requires license agreement by email)
- Afrikaans http://rma.nwu.ac.za/index.php/resource-catalogue/afribooms.html
- Turkish http://ii.metu.edu.tr/corpus (requires sending a digital copy of a signed license agreement)
- Persian http://stp.lingfil.uu.se/~mojgan/UPDT.html
- Norwegian http://www.nb.no/sprakbanken/show?serial=sbr-10
- BrazPortugese Newswire http://www.nltk.org/nltk_data
- Dutch Alpino https://www.let.rug.nl/vannoord/trees
- Spanish https://www.iula.upf.edu/recurs01_tbk_uk.htm
- Italian-TurinTree/Parallel http://www.di.unito.it/~tutreeb/treebanks.html
- Polish National Corpus http://nkjp.pl/index.php?page=14&lang=1
- Icelandic-Historical Corpus http://linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_(IcePaHC)
- Icelandic http://www.malfong.is/index.php?lang=en&pg=mim
- Slovene-English Parallel Corpus http://nl.ijs.si/elan/
- Finnish Treebank http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/
- German Tiger http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html
- German Hamburg Treebank https://corpora.uni-hamburg.de/drupal/en/islandora/object/treebank:hdt
- Russian Open Corpus http://opencorpora.org/?page=downloads
- Italian-Pisa http://www.corpusitaliano.it/en/contents/description.html
- English https://corpling.uis.georgetown.edu/gum/
- Coptic https://github.com/CopticScriptorium/corpora
- French https://deep-sequoia.inria.fr/corpus/
- French https://perso.limsi.fr/pap/free_multitag.tgz
- Danish https://code.google.com/p/copenhagen-dependency-treebank/
- Croatian http://nlp.ffzg.hr/resources/corpora/setimes-hr/
- Swedish Talbanken http://stp.lingfil.uu.se/%7Emojgan/UPDT.html
- English Ted Talk Treebank http://ahclab.naist.jp/resource/tedtreebank

Multi Universial Dependencies http://universaldependencies.org

Evaluation and tagging accuracy

- Accuracy numbers currently reported for POS tagging are most often between 95% and 97%.
- How to calculate accuracy?
  - POS tagging accuracy $=$
    $$\frac{\text{The number of correct POS labels}}{\text{The number of all POS labels}} \times 100$$
- Evaluation data (or gold data) should be provided to calculate the POS tagging accuracy.

Example of POS tagging accuracy

Persian POS tagging results using cross-lingual projection

Baseline system

- assign randomly
- the most frequent one
- ...

English Data

- Universal Dependencies English Web Treebank v1.3 – 2016-05-15 `https://github.com/UniversalDependencies/UD_English`
- A Gold Standard Universal Dependencies Corpus for English, built over the source material of the English Web Treebank LDC2012T13 (`https://catalog.ldc.upenn.edu/LDC2012T13`).

Natalia Silveira and Timothy Dozat and Marie-Catherine de Marneffe and Samuel Bowman and Miriam Connor and John Bauer and Christopher D. Manning. 2014. **A Gold Standard Dependency Corpus for English**. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Most frequent (Universal) POS tags in training data:

35313	NOUN
27509	VERB
23679	PUNCT
18047	PRON
17639	ADP
16285	DET
12946	PROPN
12476	ADJ
10549	ADV
7893	AUX

Over 204,586 tokens

Rule-based part-of-speech tagging

1. assign each word a list of potential POS labels using the **dictionary**
2. winnow down the list to a single POS label for each word using **large lists of hand-written disambiguation rules**

Adverbial-that rule

Given input: "that"
if

(+1 A/ADV/QUANT); /* *if next word is adj, adverb, or quantifier* */
(+2 SENT-LIM); /* *and following which is a sentence boundary.* */
(NOT-1 SVOC/A); /* *and the previous word is not a verb like* */
/* *'consider' which allows adjs as object complements* */

then eliminate non-ADV tags
else eliminate ADV tag

Transformation-based tagging, Brill tagger

Transformation-based learning of tags

- a specification of which **error correcting** transformation are admissible
- the learning algorithm

Transformation

A rewrite rule, $t_1 \rightarrow t_2$ means replace tag t_1 by tag t_2.

schema	t_{t-3}	t_{t-2}	t_{t-1}	t_t	t_{t+1}	t_{t+2}	t_{t+3}
1			O	*			
2				*	O		
3		O	O	*			
4				*	O	O	
5	O	O	O	*			
6				*	O	O	O
7			O	*	O		
8			O	*		O	
9		O		*	O		

source tag	target tag	triggering environment
NN	VB	previous tag is TO `NN VB PREVTAG TO` to/TO race/NN → to/TO race/VB
VBR	VB	one of the previous three tags is MD `-`
JJR	RBR	next tag is JJ `JJR RBR NEXTTAG JJ`
VBP	VB	one of the previous two words is *n't* `VBP VB PREV1OR2WD n't`

Examples of some transformations learned in transformation-based tagging (`CONTEXTUALRULEFILE`)

CONTEXTUALRULEFILE

14 CURWD
5 LBIGRAM
1 NEXT1OR2OR3TAG
5 NEXT1OR2TAG
1 NEXT2TAG
7 NEXTBIGRAM
34 NEXTTAG
8 NEXTWD
6 PREV1OR2OR3TAG
6 PREV1OR2TAG
8 PREV1OR2WD
1 PREV2TAG
10 PREVBIGRAM
56 PREVTAG
19 PREVWD
11 RBIGRAM
45 SURROUNDTAG
2 WDAND2AFT
3 WDAND2TAGAFT
1 WDAND2TAGBFR
28 WDNEXTTAG
13 WDPREVTAG

284 rules learned from WSJ

Applying transformation

"A $\rightarrow$ B if the preceding tag is A"

AAAA $\rightarrow$ A???

"A $\rightarrow$ B if the preceding tag is A"

AAAA $\rightarrow$ ABAB (immediate effect, influence each other)

AAAA $\rightarrow$ ABBB (delayed effect used in Brill tagger)

Lexical information in Brill tagger

- `LEXICON`
- `LEXICALRULEFILE`

`LEXICON`:

- overthrown VBN
- grand JJ
- unfortunate JJ NN
- Veiling VBG

LEXICALRULEFILE

- `0 haspref 1 CD x`: if a word has prefix "0" (of length 1 character), tag it as a "CD"
- `VBN un fhaspref 2 JJ x`: if a word has prefix "un" (of length 2 characters), and it is currently tagged as "VBN", then change the tag to "JJ".
- `- char JJ x`: If the character "-" appears anywhere in the word, tag it as "JJ".
- `ly hassuf 2 RB x`: If a word has suffix "ly", tag it as "RB".
- `ly addsuf 2 JJ x`: If adding the letters "ly" to the end of a word results in a word (the new word appears in LEXICON or the extended wordlist), tag it as "JJ"

Brill tagger

Brill tagger is available at
`http://www.tech.plym.ac.uk/soc/staff/guidbugm/software/RULE_BASED_TAGGER_V.1.14.tar.Z`

Eric Brill. 1992. **A simple rule-based part of speech tagger**. In *Proceedings of the third conference on Applied natural language processing (ANLC '92)*. Stroudsburg, PA, USA, 152-155.

HMM POS tagging

The best sequence of tags

- We want to choose the tag sequence that is most probable give the observation sequence of n word w_1^n $(w_1, w_2, ..., w_n)$.
- In other words, we want out of all sequence of n tags t_1^n the single tag sequence such that $P(t_1^n|w_1^n)$ is highest.

$$\hat{t}_1^n = \arg\max_{t_1^n} P(t_1^n|w_1^n) \tag{1}$$

where we want the particular tag sequence t_1^n that maximize the $\hat{t}_1^n$.

The function $\arg\max_x f(x)$ means "the x such that $f(x)$ is maximized".

Bayes' rule

Bayes' rule:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \tag{2}$$

We don't know how to directly compute $P(t_1^n|w_1^n)$. Therefore,

$$\begin{aligned}
\hat{t}_1^n &= \arg\max_{t_1^n} P(t_1^n|w_1^n) \\
&= \arg\max_{t_1^n} \frac{P(w_1^n|t_1^n)P(t_1^n)}{P(w_1^n)} \\
&= \arg\max_{t_1^n} P(w_1^n|t_1^n)P(t_1^n)
\end{aligned}$$

$P(w_1^n)$ doesn't change for each tag sequence: we are always asking about the most likely tag sequence fro the same observation w_1^n, which mush have the same probability $P(w_1^n)$.

Two assumptions

1.

The probability of a word appearing depends only on its own POS tag: that is, it is **independent** of other words and other tags around it:

$$P(w_1^n|t_1^n) \approx \prod_{i=1}^{n} P(w_i|t_i) \tag{3}$$

2.
The probability of a tag appearing is **dependent** only on the previous tag (bigram assumption), rather than the entire tag sequence.

$$P(t_1^n) \approx \prod_{i=1}^{n} P(t_i|t_{i-1}) \tag{4}$$

HMM bigram tagger

$$
\begin{aligned}
\hat{t}_1^n &= \arg\max_{t_1^n} P(t_1^n|w_1^n) \\
&= \arg\max_{t_1^n} \frac{P(w_1^n|t_1^n)P(t_1^n)}{P(w_1^n)} \\
&= \arg\max_{t_1^n} P(w_1^n|t_1^n)P(t_1^n) \\
&\approx \arg\max_{t_1^n} \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})
\end{aligned}
$$