EDAN20

Language Technology

http://cs.lth.se/edan20/

Chapter 7: Part-of-Speech Tagging Using Rules

Pierre Nugues

Lund University
Pierre.Nugues@cs.lth.se
http://cs.lth.se/pierre_nugues/

September 10, 2015



The Parts of Speech

The parts of speech (POS) are classes that correspond to the lexical – or word – categories

Plato made a distinction between the verb and the noun.

After him, the word categories further evolved and grew in number until Dionysus Thrax formulated and fixed them.

Aelius Donatus popularized the list of the eight parts of speech: noun, pronoun, verb, participle, conjunction, adverb, preposition, and interjection. Grammarians have adopted these POS for most European languages although they are somewhat arbitrary



Part-of-speech Annotation

Sentence:

That round table might collapse

Annotation:

Words	Parts of speech	POS tags
that	Determiner	DT
round	Adjective	JJ
table	Noun	NN
might	Modal verb	MD
collapse	Verb	VB

The automatic annotation uses predefined POS tagsets such as the Penn Treebank tagset for English

Word Ambiguity

	English	French	German
Part of speech	can modal	<i>le</i> article	<i>der</i> article
	can noun	<i>le</i> pronoun	der pronoun
Semantic	<i>great</i> big	grand big	groß
	great notable	grand notable	groß



POS Tagging

Words	Possible tags	Example of use
that	Subordinating conjunction	That he can swim is good
	Determiner	That white table
	Adverb	It is not that easy
	Pronoun	That is the table
	Relative pronoun	The table that collapsed
round	Verb	Round up the usual suspects
	Preposition	Turn round the corner
	Noun	A big round
	Adjective	A round box
	Adverb	He went round
table	Noun	That white table
	Verb	I table that
might	Noun	The might of the wind
	Modal verb	She might come
collapse	Noun	The collapse of the emp
	Verb	The empire can collapse

Part-of-Speech Ambiguity in Swedish

The word som in the Norstedts svenska ordbok, 1999, has three entries:

- Om jag vore lika vacker som du, skulle jag vara lycklig. (konjunktion)
- 2 Bilen som jag köpte i fjol. (pronomen)
- 3 Som jag har saknat dig. (adverb)

The part-of-speech difference can be significant:

- Swedish. Compare the pronunciation of *vaken*, adjective, as in *Han är* aldrig vaken innan klockan sju and vaken, noun, as in *Vi* fiskade i vaken i sjön
- English. Compare *object* in *I object to violence*, verb, or *I could see an object*, noun.

Simple Grammatical Constraints are not Satisfying

Although, it makes no sense,

I see a bird

can be tagged as:

I/noun see/noun a/noun bird/noun

Because sequences of four nouns are possible in English as in:

city school committee meeting.

The disambiguation methods are based on

- Handcrafted rules
- Automatically learned rules
- Statistical methods

Currently disambiguation accuracy is greater than 95% for many



POS Annotation with Rules

The phrase *The can rusted* has two readings Let's suppose that *can*/modal is more frequent than *can*/noun in our corpus

First step: Assign the most likely POS

The/art can/modal rusted/verb

Second step: Apply rules

Change the tag from modal to noun if one of the two

previous words is an article

The/art can/noun rusted/verb

This is the idea of Brill's tagger.



EDAN20 Language Technology http://cs.lth.se/edan20/

Rule Templates

Rules	Explanation
alter(A, B, prevtag(C))	Change A to B if preceding tag is C
<pre>alter(A, B, nexttag(C))</pre>	Change A to B if the following tag is C
<pre>alter(A, B, prev2tag(C))</pre>	Change A to B if tag two before is C
<pre>alter(A, B, next2tag(C))</pre>	Change A to B if tag two after is C
<pre>alter(A, B, prev1or2tag(C))</pre>	Change A to B if one of the two preceding tags is C
<pre>alter(A, B, next1or2tag(C))</pre>	Change A to B if one of the two following tags is C
<pre>alter(A, B, surroundingtag(C, D))</pre>	Change A to B if surrounding tags are C and D
<pre>alter(A, B, nextbigram(C, D))</pre>	Change A to B if next bigram tag is C D
<pre>alter(A, B, prevbigram(C, D))</pre>	Change A to B if previous bigram tag is C D

Learning Rules Automatically

Compare the hand-annotation of the reference corpus with the automatic one

```
Automatic tagging Hand annotation: gold standard The/art can/modal rusted/verb The/art can/noun rusted/verb
```

For each error instantiate the templates Rules correcting the error

```
alter(modal, noun, prevtag(art)).
alter(modal, noun, prev1or2tag(art)).
alter(modal, noun, nexttag(verb))
alter(modal, noun, surroundingtag(art, verb))
```

Rules introduce good and bad transformations Select the rule that has the greatest error reduction and apply it

Part-of-Speech Ambiguity in Swedish

The Swedish word *den* can be a determiner or a pronoun. It corresponds to two entries in the *Nordstedts svenska ordbok* (1999, page 187):

- den artikel . . . som här antas vara känd . . . : den nya bilen
- den pron. personen eller företeelsen som är omtalad i sammanhanget
 ...: Var har du köpt kameran? Jag har fått den i present.

Frequency information:

```
egrep -i "den dt" talbanken.txt | wc -l
820
egrep -i "den pn" talbanken.txt | wc -l
256
```

EDAN20 Language Technology http://cs.lth.se/edan20/



Ambiguity Resolution in Swedish: The Baseline

Let us suppose that *den* is the only word to tag in the corpus and that it has two possible parts of speech: dt and pn.

Using the most frequent part of speech produces the annotations:

If the POS tagger is restricted to den, out of 820 + 256 = 1076 POS assignments,

$$\frac{820}{1076} = 76\%$$

are correct.



Ambiguity Resolution in Swedish: The Rule Templates

Let us use two rules templates alter(A, B, prev(C)) and alter(A, B, next(C)) and instantiate them with the error on Jag har fått den i present.

It yields:

- Change dt to pn if previous POS tag is vb: alter(dt, pn, prev(vb))
- Change dt to pn if next POS tag is pp: alter(dt, pn, next(pp))

Both rules produce a correct annotation on the training example.

Ambiguity Resolution in Swedish: Selecting the Rules

Let us apply the two rules to all the occurrences of *den* in the corpus and ignore all the other words:

- The first rule corrects 15 wrong annotations of *den* and introduces 59 mistakes: 15-59=-44
- The second rule corrects 20 wrong annotations and introduces 5 mistakes: 20-5=+5

The training step of Brill's tagger selects the most efficient rule, here alter(dt, pn, next(pp)).

Of course, this step is applied to all the ambiguous words and not only *den*. We iterate the procedure until the error rate is below a certain threshold.



Brill's Learning Algorithm

St.	Operation	Input	Output
1.	Annotate each word of the corpus with its most likely part of speech	Corpus	AnnotatedCorpus(1)
2.	Compare pairwise the part of speech of each word of the <i>AnnotationReference</i> and <i>AnnotatedCorpus(i)</i>	AnnotationReference AnnotatedCorpus(i	
3.	For each error, instantiate the rule templates to correct the error	List of errors	List of tentative rules
4.	For each instantiated rule, compute on <i>AnnotatedCorpus(i)</i> the number of good transformations minus the number of bad transformations the rule yields) Scored tentative rules

Brill's Learning Algorithm

St.	Operation	Input	Output
5.	Select the rule that has the greatest error reduction and append it to the ordered list of transformations	Tentative rules	Rule(i)
6.	Apply Rule(i) to Annotated-Corpus(i)	AnnotatedCorpus(i) Rule(i)	AnnotatedCorpus(i+1)
7.	If number of errors is under predefined threshold, end the algorithm else go to step 2.	_	List of rules



First Brill's Rules

	Chang	e	
#	From	To	Condition
1	NN	VB	Previous tag is TO
2	VBP	VB	One of the previous three tags is MD
3	NN	VB	One of the previous two tags is MD
4	VB	NN	One of the previous two tags is DT
5	VBD	VBN	One of the previous three tags is VBZ

In the table, rules consider parts of speech only. This is the normal case and they are called unlexicalized.

Rules can also consider word values and they are called lexicalized.

Chang	e		* *	
From	To	Condition	J	A 10.10
IN	RB	The word two positions to the right is	as 🖤	See Line
			A STATE OF THE PARTY OF THE PAR	S

Standard POS Tagsets: The Penn Treebank

1.	CC	Coordinating conjunction	25.	TO	to
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Verb, base form
4.	EX	Existential there	28.	VBD	Verb, past tense
5.	FW	Foreign word	29.	VBG	Verb, gerund/present participle
6.	IN	Preposition/sub. conjunction	30.	VBN	Verb, past participle
7.	JJ	Adjective	31.	VBP	Verb, non-third pers. sing. pres.
8	JJR	Adjective, comparative	32.	VBZ	Verb, third-pers. sing. present
9.	JJS	Adjective, superlative	33.	WDT	wh-determiner
10.	LS	List item marker	34.	WP	<i>wh</i> -pronoun
11.	MD	Modal	35.	WP\$	Possessive wh-pronoun
12.	NN	Noun, singular or mass	36.	WRB	<i>wh</i> -adverb
13.	NNS	Noun, plural	37.	#	Pound sign
14.	NNP	Proper noun, singular	38.	\$	Dollar sign
15.	NNPS	Proper noun, plural	39.		Sentence final punctuation
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semicolon
18.	PRP	Personal pronoun	42.	(Left bracket character
19.	PRP\$	Possessive pronoun	43.)	Right bracket character
20.	RB	Adverb	44.	II	Straight double quotes
21.	RBR	Adverb, comparative	45.	4	Left open single 🏈
22.	RBS	Adverb, superlative	46.	**	Left open double quality
23.	RP	Particle	47.	1	Right close single quote
24.	SYM	Symbol	48.	" ∢	□ Right close double quee

An Example of Tagged Text from the Penn Treebank

Battle-tested/JJ Japanese/JJ industrial/JJ managers/NNS here/RB always/RB buck/VBP up/RP nervous/JJ newcomers/NNS with/IN the/DT tale/ NN of/IN the/DT first/JJ of/IN their/PP\$ countrymen/NNS to/TO visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN samurai/FW warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./. "/" From/IN the/DT beginning/NN ,/, it/PRP took/VBD a/DT man/NN with/IN extraordinary/JJ qualities/NNS to/TO succeed/VB in/IN Mexico/NNP "/" says/VBZ Kimihide/NNP Takimura/NNP ,/, president/NN of/IN the/DT Mitsui/NNP group/NN 's/POS Kensetsu/NNP Engineering/NNP Inc./NNP unit/NN ./.



EDAN20 Language Technology http://cs.lth.se/edan20/

Measuring Quality: The Confusion Matrix

From Franz (1996, p. 124)

↓Correct	Tagg	er o								
	DT	IN	JJ	NN	RB	RP	VB	VBD	VBG	VBN
DT	99.4	0.3	_	_	0.3	_	_	_	_	_
IN	0.4	97.5	_	_	1.5	0.5	_	_	_	_
JJ	_	0.1	93.9	1.8	0.9	_	0.1	0.1	0.4	1.5
NN	_	_	2.2	95.5	_	_	0.2	_	0.4	_
RB	0.2	2.4	2.2	0.6	93.2	1.2	_	_	_	_
RP	_	24.7	_	1.1	12.6	61.5	_	_	_	_
VB	_	_	0.3	1.4	_	_	96.0	_	_	0.2
VBD	_	_	0.3	_	_	_	_	94.6	_	4.8
VBG	_	_	2.5	4.4	_	_	_	_	93.0	_
VBN	_	_	4.6	_	_	_	_	4.3	56年1	90.6

Recognizing Parts of Speech

Parts of speech denomination is comparable in Western European languages and roughly corresponds
They follow Donatus' teaching
(http://htl2.linguist.jussieu.fr:8080/CGL/text.jsp?id=T28)

Two common mistakes in the labs:

If you are not sure, look up in a dictionary

- Confusion between noun and the Swedish word namn.
 - A common noun, or more simply a noun, corresponds to *substantiv*
 - Proper noun, or name, (or proper name) corresponds to namn or egennamn.
- Possessive pronouns like my, your, his, her, ... are not real pronouns.
 They should be called possessive adjectives or determiners.

Multext and Google's Universal POS tagset

Part of speech	Multext	Universal POS
Noun	N	Noun
Verb	V	Verb
Adjective	Α	Adj
Pronoun	Р	Pron
Determiner	D	Det
Adverb	R	Adv
Adposition (Preposition)	S	Adp
Conjunction	C	Conj
Numeral	M	Num
Interjection	1	-
Residual	X	X
Particle	-	Prt
Ponctuation mark	-	

Attributes for Nouns (Multext)

Position	Attribute	Value	Code
1	Туре	Common	С
		Proper	р
		Masculine	m
2	Gender	Feminine	f
		Neuter	n
3	Number	Singular	S
		Plural	р
		Nominative	n
4	Case	Genitive	g
		Dative	d
		Accusative	а



Annotation for Swedish: Tokens

<token id="10">så</token>
<token id="11">att</token>

Bilen framför justitieministern svängde fram och tillbaka över vägen så att hon blev rädd.

'The car in front of the Justice Minister swung back and forth and she was frightened.'

Parts of Speech for Swedish

```
<taglemmas>
  <taglemma id="1" tag="nn.utr.sin.def.nom" lemma="bil"/>
  <taglemma id="2" tag="pp" lemma="framför"/>
  <taglemma id="3" tag="nn.utr.sin.def.nom" lemma="justitieminister"</pre>
  <taglemma id="4" tag="vb.prt.akt" lemma="svänga"/>
  <taglemma id="5" tag="ab" lemma="fram"/>
  <taglemma id="6" tag="kn" lemma="och"/>
  <taglemma id="7" tag="ab" lemma="tillbaka"/>
  <taglemma id="8" tag="pp" lemma="över"/>
  <taglemma id="9" tag="nn.utr.sin.def.nom" lemma="väg"/>
  <taglemma id="10" tag="ab" lemma="så"/>
  <taglemma id="11" tag="sn" lemma="att"/>
  <taglemma id="12" tag="pn.utr.sin.def.sub" lemma="hon"/>
  <taglemma id="13" tag="vb.prt.akt.kop" lemma="bli"/>
  <taglemma id="14" tag="jj.pos.utr.sin.ind.nom" lemma="p</pre>
  <taglemma id="15" tag="mad" lemma="."/>
</taglemmas>
```