

EDAN95

Applied Machine Learning

<http://cs.lth.se/edan95/>

Lecture 8: Word and Segment Categorization

Pierre Nugues

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

November 26, 2018

Word Categorization: The Parts of Speech

Sentence:

That round table might collapse

Annotation:

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

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

Ambiguity

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

Segment Recognition

Group detection – chunking –:

Brackets: [_{NG} The government _{NG}] has [_{NG} other agencies and instruments _{NG}] for pursuing [_{NG} these other objectives _{NG}] .

Tags: *The/I government/I has/O other/I agencies/I and/I instruments/I for/O pursuing/O these/I other/I objectives/I ./O*

Brackets: Even [_{NG} Mao Tse-tung _{NG}] [_{NG} 's China _{NG}] began in [_{NG} 1949 _{NG}] with [_{NG} a partnership _{NG}] between [_{NG} the communists _{NG}] and [_{NG} a number _{NG}] of [_{NG} smaller, non-communists parties _{NG}] .

Tags: *Even/O Mao/I Tse-tung/I 's/B China/I began/O in/O 1949/I with/O a/I partnership/I between/O the/I communists/I and/O a/I number/I of/O smaller/I ,/I non-communists/I parties/I ./O*

Segment Categorization

Tages extendible to any type of chunks: nominal, verbal, etc.

For the IOB scheme, this means tags such as I.Type, O.Type, and B.Type, Types being NG, VG, PG, etc.

In CoNLL 2000, ten types of chunks

Word	POS	Group	Word	POS	Group
<i>He</i>	PRP	B-NP	<i>to</i>	TO	B-PP
<i>reckons</i>	VBZ	B-VP	<i>only</i>	RB	B-NP
<i>the</i>	DT	B-NP	<i>£</i>	#	I-NP
<i>current</i>	JJ	I-NP	<i>1.8</i>	CD	I-NP
<i>account</i>	NN	I-NP	<i>billion</i>	CD	I-NP
<i>deficit</i>	NN	I-NP	<i>in</i>	IN	B-PP
<i>will</i>	MD	B-VP	<i>September</i>	NNP	B-NP
<i>narrow</i>	VB	I-VP	<i>.</i>	.	O

Noun groups (NP) are in red and verb groups (VP) are in blue.

















IOB Annotation for Named Entities

CoNLL 2002		CoNLL 2003			
Words	Named entities	Words	POS	Groups	Named entities
Wolff	B-PER	U.N.	NNP	I-NP	I-ORG
,	O	official	NN	I-NP	O
currently	O	Ekeus	NNP	I-NP	I-PER
a	O	heads	VBZ	I-VP	O
journalist	O	for	IN	I-PP	O
in	O	Baghdad	NNP	I-NP	I-LOC
Argentina	B-LOC	.	.	O	O
,	O				
played	O				
with	O				
Del	B-PER				
Bosque	I-PER				
in	O				
the	O				
final	O				
years	O				
of	O				
the	O				
seventies	O				
in	O				
Real	B-ORG				
Madrid	I-ORG				
.	O				

Evaluation

Accuracy, precision, and recall.

For noun groups with the predicted output:

Word	POS	Group	Word	POS	Group
 <i>He</i>	PRP	B-NP	 <i>to</i>	TO	B-PP
 <i>reckons</i>	VBZ	B-VP	 <i>only</i>	RB	B-NP
 <i>the</i>	DT	B-NP	 <i>£</i>	#	I-NP
 <i>current</i>	JJ	B-NP	 <i>1.8</i>	CD	B-NP
 <i>account</i>	NN	I-NP	 <i>billion</i>	CD	I-NP
 <i>deficit</i>	NN	I-NP	 <i>in</i>	IN	B-PP
 <i>will</i>	MD	B-VP	 <i>September</i>	NNP	B-NP
 <i>narrow</i>	VB	I-VP	 <i>.</i>	.	O

Accuracy = $\frac{14}{16}$, recall = $\frac{2}{4} = 0.5$, precision = $\frac{2}{6} = 0.33$

harmonic mean = $2 \times \frac{0.33 \times 0.5}{0.33 + 0.5} = 0.4$

Complete Code Example

Jupyter Notebook: Chollet 6.1 <https://github.com/fchollet/deep-learning-with-python-notebooks/6.1-using-word-embeddings.ipynb>