Artificial Intelligence
EDA132

Lecture 13.2: Language Models, Part-of-Speech Tagging, and Named Entity Recognition

Pierre Nugues

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

March 1st, 2017

Corpora

A corpus is a collection of texts (written or spoken) or speech
Corpora are balanced from different sources: news, novels, etc.

	English	French	German
Most frequent words in a collection of contemporary running texts	*the*	*de*	*der*
	of	*le* (article)	*die*
	to	*la* (article)	*und*
	in	*et*	*in*
	and	*les*	*des*
Most frequent words in Genesis	*and*	*et*	*und*
	the	*de*	*die*
	of	*la*	*der*
	his	*à*	*da*
	he	*il*	*er*

Characteristics of Current Corpora

Big: The Bank of English (Collins and U Birmingham) has more than 500 million words

Available in many languages

Easy to collect: The web is the largest corpus ever built and within the reach of a mouse click

Parallel: same text in two languages: English/French (Canadian Hansards), European parliament (23 languages)

Annotated with part-of-speech or manually parsed (treebanks):

- Characteristics/`N` of/`PREP` Current/`ADJ` Corpora/`N`
- (`NP` (`NP` Characteristics) (`PP` of (`NP` Current Corpora)))

Lexicography

Writing dictionaries

Dictionaries for language learners should be build on real usage

- *They're just trying to score* ***brownie points*** *with politicians*
- *The boss is pleased – that's another* ***brownie point***

Bank of English: *brownie point* (6 occs) *brownie points* (76 occs)

Extensive use of corpora to:

- Find **concordances** and cite real examples
- Extract **collocations** and describe frequent pairs of words

Concordances

A word and its context:

Language	Concordances
English	`s beginning of miracles did Je`
	`n they saw the miracles which`
	`n can do these miracles that t`
	`ain the second miracle that Je`
	`e they saw his miracles which`
French	`le premier des miracles que fi`
	`i dirent: Quel miracle nous mo`
	`om, voyant les miracles qu'il`
	`peut faire ces miracles que tu`
	`s ne voyez des miracles et des`

Collocations

Word preferences: Words that occur together

	English	**French**	**German**
You say	*Strong tea*	*Thé fort*	*Schmales Gesicht*
	Powerful computer	*Ordinateur puissant*	*Enge Kleidung*
You don't say	*Strong computer*	*Thé puissant*	*Schmale Kleidung*
	Powerful tea	*Ordinateur fort*	*Enges Gesicht*

Word Preferences

Strong w			Powerful w		
strong w	*powerful w*	*w*	*strong w*	*powerful w*	*w*
161	0	showing	1	32	than
175	2	support	1	32	figure
106	0	defense	3	31	minority
...					

Corpora as Knowledge Sources

Short term:

- Describe usage more accurately
- Assess tools: part-of-speech taggers, parsers.
- Learn statistical/machine learning models for speech recognition, taggers, parsers

Longer term:

- Semantic processing and knowledge extraction
- Texts are the main repository of human knowledge

Counting Words and Word Sequences

Words have specific contexts of use.

Pairs of words like *strong* and *tea* or *powerful* and *computer* are not random associations.

Psychological linguistics tells us that it is difficult to make a difference between *writer* and *rider* without context

A listener will discard the improbable *rider of books* and prefer *writer of books*

A language model is the statistical estimate of a word sequence.

Originally developed for speech recognition

The language model component enables to predict the next word given a sequence of previous words: *the writer of books, novels, poetry*, etc. and not *the writer of hooks, nobles, poultry*, . . .

N-Grams

The types are the distinct words of a text while the tokens are all the words or symbols.

The phrases from *Nineteen Eighty-Four*

> *War is peace*
> *Freedom is slavery*
> *Ignorance is strength*

have 9 tokens and 7 types.

Unigrams are single words

Bigrams are sequences of two words

Trigrams are sequences of three words

Trigrams

Word	Rank	More likely alternatives
We	9	*The This One Two A Three Please In*
need	7	*are will the would also do*
to	1	
resolve	85	*have know do...*
all	9	*the this these problems...*
of	2	*the*
the	1	
important	657	*document question first...*
issues	14	*thing point to...*
within	74	*to of and in that...*
the	1	
next	2	*company*
two	5	*page exhibit meeting day*
days	5	*weeks years pages months*

Probabilistic Models of a Word Sequence

$$
\begin{aligned}
P(S) &= P(w_1, ..., w_n), \\
&= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)...P(w_n|w_1, ..., w_{n-1}), \\
&= \prod_{i=1}^{n} P(w_i|w_1, ..., w_{i-1}).
\end{aligned}
$$

The probability $P(\mathit{It\ was\ a\ bright\ cold\ day\ in\ April})$ from *Nineteen Eighty-Four* corresponds to

It to begin the sentence, then *was* knowing that we have *It* before, then *a* knowing that we have *It was* before, and so on until the end of the sentence.

$$
\begin{aligned}
P(S) = & P(It) \times P(was|It) \times P(a|It, was) \times P(bright|It, was, a) \times ... \\
& \times P(April|It, was, a, bright, ..., in).
\end{aligned}
$$

Approximations

Bigrams:

$$P(w_i|w_1, w_2, ..., w_{i-1}) \approx P(w_i|w_{i-1}),$$

Trigrams:

$$P(w_i|w_1, w_2, ..., w_{i-1}) \approx P(w_i|w_{i-2}, w_{i-1}).$$

Using a trigram language model, $P(S)$ is approximated as:

$$\begin{aligned} P(S) &\approx P(It) \times P(was|It) \times P(a|It, was) \times P(bright|was, a) \times ... \\ &\quad \times P(April|day, in). \end{aligned}$$

Maximum Likelihood Estimate

Bigrams:

$$P_{\mathrm{MLE}}(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i)}{\sum\limits_{w} C(w_{i-1}, w)} = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}.$$

Trigrams:

$$P_{\mathrm{MLE}}(w_i|w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})}.$$

Part-of-Speech Tagging

The annotation of the words with their part of speech: article, noun, verb, adjective, etc.

Given the sentence:

> *That round table might collapse.*

The correct part-of-speech tagging is:

> *That/**determiner** round/**adjective** table/**noun** might/**modal verb** collapse/**verb**.*

Part-of-speech tagging (POS tagging) is a compulsory step to most NLP applications.

Not as Simple as it Seems

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 empire*
	Verb	*The empire can collapse*

Part-of-Speech Ambiguity in Swedish

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

1. *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.

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	*wh*-pronoun
11.	MD	Modal	35.	WP$	Possessive *wh*-pronoun
12.	NN	Noun, singular or mass	36.	WRB	*wh*-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.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	‘	Left open single quote
22.	RBS	Adverb, superlative	46.	“	Left open double quote
23.	RP	Particle	47.	’	Right close single quote
24.	SYM	Symbol	48.	”	Right close double quote

Training Set

Part-of-speech taggers use a training set where every word is hand-annotated (Penn Treebank and CoNLL 2008).

Index	Word	Hand annotation	Index	Word	Hand annotation
1	Battle	JJ	19	of	IN
2	-	HYPH	20	their	PRP$
3	tested	JJ	21	countrymen	NNS
4	Japanese	JJ	22	to	TO
5	industrial	JJ	23	visit	VB
6	managers	NNS	24	Mexico	NNP
7	here	RB	25	,	,
8	always	RB	26	a	DT
9	buck	VBP	27	boatload	NN
10	up	RP	28	of	IN
11	nervous	JJ	29	samurai	FW
12	newcomers	NNS	30	warriors	NNS
13	with	IN	31	blown	VBN
14	the	DT	32	ashore	RB
15	tale	NN	33	375	CD
16	of	IN	34	years	NNS
17	the	DT	35	ago	RB
18	first	JJ	36	.	.

Part-of-Speech Tagging with Linear Classifiers

Linear classifiers are efficient devices to carry out part-of-speech tagging:

1. The lexical values are the input data to the tagger.
2. The parts of speech are assigned from left to right by the tagger.

Given the feature vector: $(w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, t_{i-2}, t_{i-1})$, the classifier will predict the part-of-speech tag t_i at index i.

ID	FORM	PPOS	
	BOS	BOS	*Padding*
	BOS	BOS	
1	Battle	NN	
2	-	HYPH	
3	tested	NN	
...	...	...	
17	the	DT	
18	first	JJ	
19	of	IN	
20	their	PRP$	
21	countrymen	NNS	*Input features*
22	to	TO	
23	visit	VB	*Predicted tag*
24	Mexico		↓
25	,		
26	a		
27	boatload		
...	...	...	
34	years		
35	ago		
36	.		
	EOS		*Padding*
	EOS		

POS Annotation with Machine Learning

The feature vectors to predict the parts of speech

ID	Feature vectors							PPOS
	w_{i-2}	w_{i-1}	w_i	w_{i+1}	w_{i+2}	t_{i-2}	t_{i-1}	
1	BOS	BOS	Battle	-	tested	BOS	BOS	NN
2	BOS	Battle	-	tested	Japanese	BOS	NN	HYPH
3	Battle	-	tested	Japanese	industrial	NN	HYPH	JJ
...	...	...	...	...	...	...	...	...
19	the	first	of	their	countrymen	DT	JJ	IN
20	first	of	their	countrymen	to	JJ	IN	PRP$
21	of	their	countrymen	to	visit	IN	PRP$	NNS
22	their	countrymen	to	visit	Mexico	PRP$	NNS	TO
23	countrymen	to	visit	Mexico	,	NNS	TO	VB
24	to	visit	Mexico	,	a	TO	VB	NNP
25	visit	Mexico	,	a	boatload	VB	NNP	,
...	...	...	...	...	...	...	...	...
34	ashore	375	years	ago	.	RB	CD	NNS
35	375	years	ago	.	EOS	CD	NNS	RB
36	years	ago	.	EOS	EOS	NNS	RB	.

Tagging Techniques to Extract Groups

Group detection – chunking – can be reframed as a tagging operation.

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

To: *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

From: 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 $_{NG}$] .

To: *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

Other Chunking Schemes

Tjong and Venstra (1999) created 3 other schemes: IOB1, IOB2, IOE1, and IOB2:

IOB1 : Inside, Outside, Between

IOB2 : Begin, Inside, Outside

IOE1 : Inside, Outside, End (between two chunks)

IOE2 : Inside, Outside, End

Other Chunking Schemes

IOB1 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, non-communists/I parties/I

IOB2 Even/O Mao/B Tse-tung/I 's/B China/I began/O in/O 1949/B with/O a/B partnership/I between/O the/B communists/I and/O a/B number/I of/O smaller/B, non-communists/I parties/I

IOE1 Even/O Mao/I Tse-tung/E 's/I 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, non-communists/I parties/I

IOE2 Even/O Mao/I Tse-tung/E 's/I China/E began/O in/O 1949/E with/O a/I partnership/E between/O the/I communists/E and/O a/I number/E of/O smaller/I, non-communists/I parties/E

Multiple Categories of Chunks

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

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

IOB Annotation for Named Entities

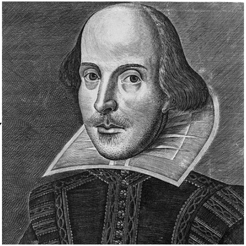
CoNLL 2002

Words	Named entities
Wolff	B-PER
,	O
currently	O
a	O
journalist	O
in	O
Argentina	B-LOC
,	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

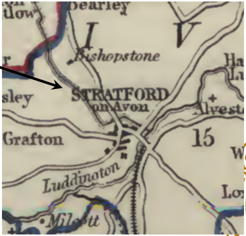
CoNLL 2003

Words	POS	Groups	Named entities
U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

Named Entities: Proper Nouns



William Shakespeare was born and brought up in Stratford-upon-Avon



Others Entities: Common Nouns

Meeting with our guest on the landing at lunchtime

Supervised Learning: A Summary

Needs a manually annotated corpus called the **gold standard**
The gold standard may contain errors (*errare humanum est*) that we ignore
A classifier is trained on a part of the corpus, the **training set**, and evaluated on another part, the **test set**, where automatic annotation is compared with the *gold standard*
N-fold cross validation is used avoid the influence of a particular division
Some algorithms may require additional optimization on a development set
Classifiers can use statistical or symbolic methods