Applied Natural Language Processing

Info 256

Lecture 20: Multiword expressions (Nov 1, 2023)

David Bamman, UC Berkeley

Words

- One morning I shot an elephant in my pajamas
- I didn’t shoot an elephant
- Imma let you finish but Beyonce had one of the best videos of all time
- I do uh main- mainly business data processing
- 一天早上我穿着睡衣射了一只大象

- The White House pledged to cut down the red tape for access to public documents.

Multiword expressions

- The White House pledged to cut down the red tape for access to public documents.

Multiword expressions

The
White_House
pledged
to
cut_down
the
red_tape
for
access
to
public
documents

Multiword expressions

type	examples
MW compounds	red tape, motion picture, daddy longlegs, hot air balloon, trash talk
verb-particle	pick up, dry out, take over, cut short, hold hostage, take seriously
verb-noun	pay attention (to), go bananas, lose it, break a leg, make the most of
support verbs	make decisions, take breaks, take pictures, have fun, perform surgery
coordination	cut and dried/dry, more or less, up and leave
connective	as well as, let alone, in spite of, on the face of it/on its face
fixed phrase	easy as pie, scared to death, go to hell, bring home the bacon
proverbs	Beggars can't be choosers. The early bird gets the worm.

Schneider 2014

Multiword expressions

- Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity

Baldwin and Kim (2010); Sag (2002)

Predictability

- The meaning and behavior of multiword expressions is typically not predictable from the individual words that comprise it.

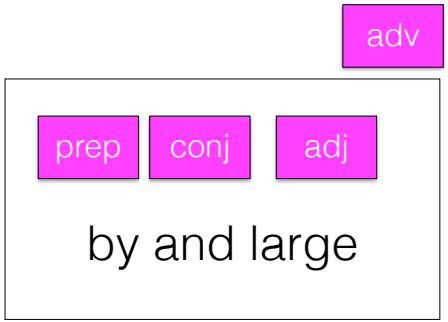
  - “dog”
  - “top”
  - “days”

  - “dog days”
  - “top dog”

Baldwin and Kim (2010)

Syntactic idiomaticity

- The syntax of the MWE isn't predictable from its components



Baldwin and Kim 2014

Semantic idiomaticity

- The meaning of a MWE is not predictable from its components

kick →

- S: (v) **kick** (drive or propel with the foot)
- S: (v) **kick** (thrash about or strike out with the feet)
- S: (v) **kick** (strike with the foot) *"The boy kicked the dog"; "Kick the door down"*
- S: (v) **kick** (kick a leg up)
- S: (v) kick back, recoil, **kick** (spring back, as from a forceful thrust) *"The gun kicked back into my shoulder"*
- S: (v) **kick**, give up (stop consuming) *"kick a habit"; "give up alcohol"*
- S: (v) **kick** (make a goal) *"He kicked the extra point after touchdown"*
- S: (v) complain, **kick**, plain, sound off, quetch, kvetch (express complaints, discontent, displeasure, or unhappiness) *"My mother complains all day"; "She has a lot to kick about"*

bucket →

- S: (n) **bucket**, pail (a roughly cylindrical vessel that is open at the top)
- S: (n) **bucket**, bucketful (the quantity contained in a bucket)

kick the bucket →

- S: (v) die, decease, perish, go, exit, pass away, expire, pass, **kick the bucket**, cash in one's chips, buy the farm, conk, give-up the ghost, drop dead, pop off, choke, croak, snuff it (pass from physical life and lose all bodily attributes and functions necessary to sustain life) *"She died from cancer"; "The children perished in the fire"; "The patient went peacefully"; "The old guy kicked the bucket at the age of 102"*

Baldwin and Kim 2014

Pragmatic idiomaticity

- An MWE is associated “with a fixed set of situations or particular context”

good morning!	Fixed greeting used at same time of day
all aboard!	used in specific situation of boarding a train/ship
shock and awe	fixed phrased associated with specific moment in Iraq War

Baldwin and Kim 2014

Lexical idiomaticity

- At least one component of the MWE doesn't appear in the vocabulary on its own.

ad hoc	"created or done for a particular purpose as necessary"

- Neither "ad" nor "hoc" are English words on their own.

Baldwin and Kim 2014

Statistical idiomaticity

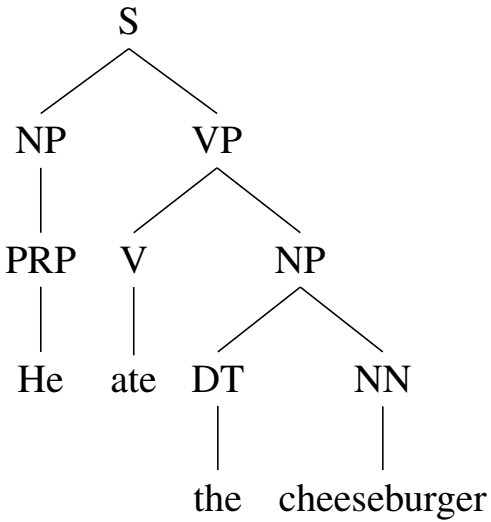
- The words in a MWE occur unusually frequently together compared to their individual frequency.

	flawless	immaculate	impeccable	spotless
condition	+	−	+	+
credentials	−	−	+	−
hair	−	+	?	−
house	?	+	?	+
logic	+	−	+	−
timing	?	+	+	−

Note: "+" = strong lexical affinity, "?" = neutral lexical affinity, "−" = negative lexical affinity.

Baldwin and Kim 2014; Cruse 1986

Compositionality

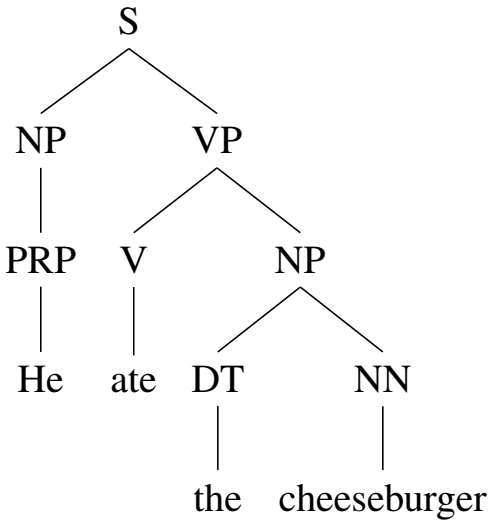


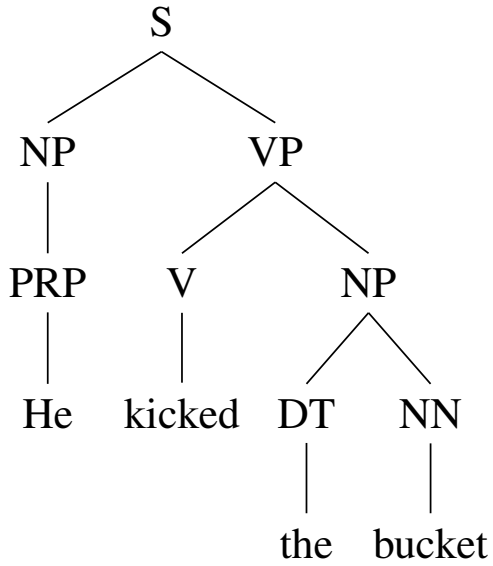
We can build up the meaning of a sequence by the combination of its parts

He
the cheeseburger
ate
ate the cheeseburger
He ate the cheeseburger

Compositionality

idioms resist compositionality





MWE dictionaries

- WordNet contains multiword entries

Random sample of WordNet MWEs:

arctic willow
blade apple
cardiac valve
de bakey
glycerol tristearate
line of descent
madagascar cat
vaginal discharge
western red cedar
works program

Wiktionary
The free dictionary

Main Page
Community portal
Preferences
Requested entries
Recent changes
Random entry
Help
Glossary
Donations
Contact us

Tools

What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Cite this page

Visibility

Show translations
Show quotations

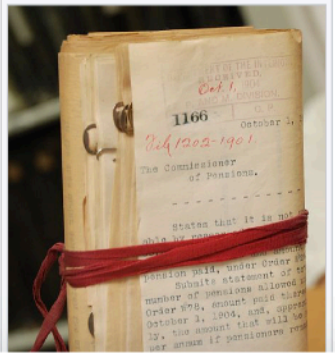
In other languages

red tape

English [edit]

Etymology [edit]

- Thought to allude to the former practice of binding government documents in red-coloured tape



Noun [edit]

red tape (*uncountable*)

1. The binding tape once used for holding important documents together. [quotations ▼]
2. (*metonymically*, *idiomatic*) Time-consuming regulations or bureaucratic procedures.

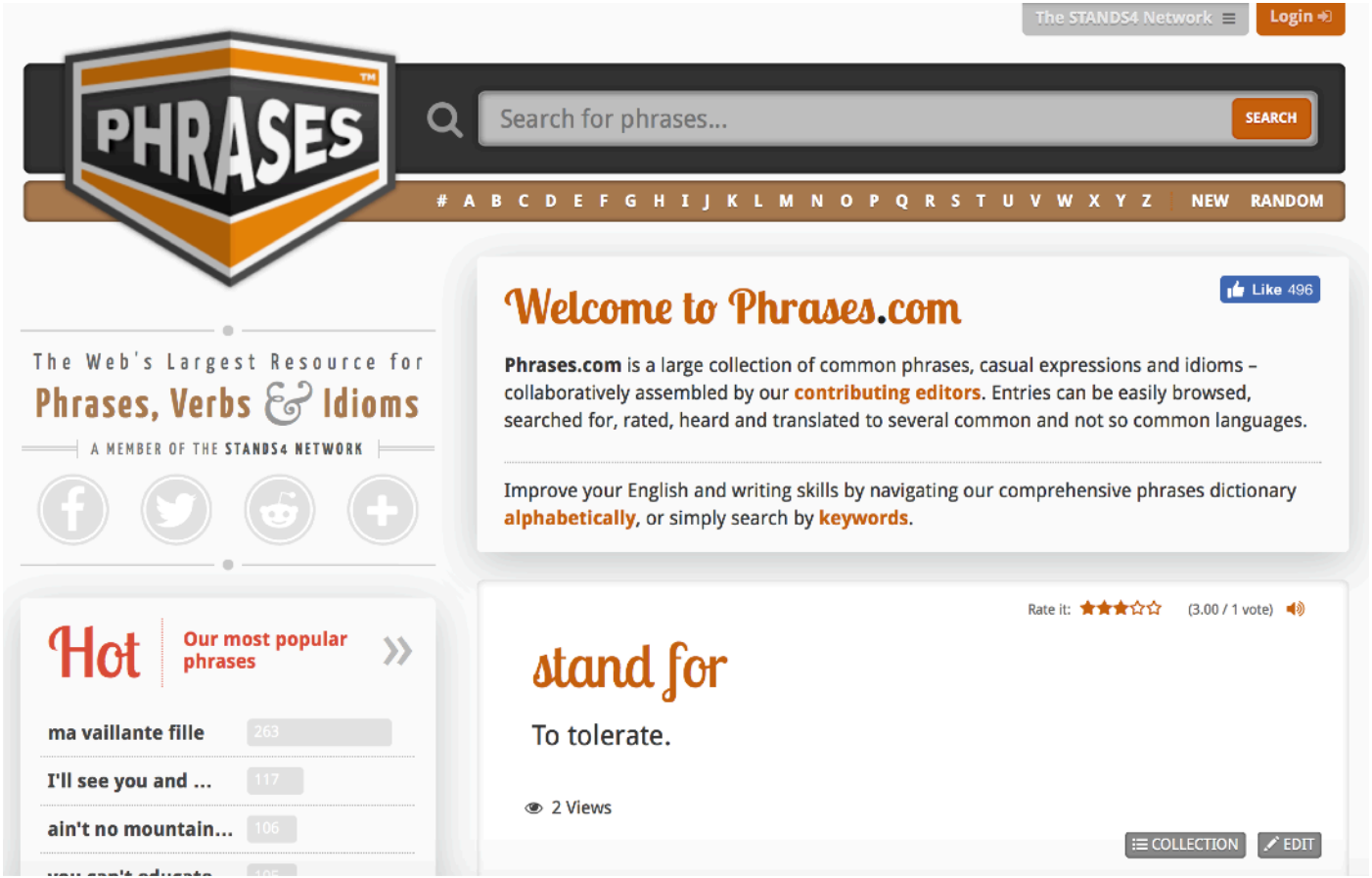
WikiMWE

- 350,000 MWEs (2-4grams) of technical terminology and collocations

- Anchor Text (Internal Links): [[target|**term_candidate**]]
- Section Headers: ==== * **term_candidate** ==== *
- Phrases in Boldface: '''**term_candidate**'''
- Phrases in Italics: ''**term_candidate**''

```
His administration filed briefs that urged the [[Supreme Court of the United States|Supreme Court]] to strike down [[Same-sex marriage in the United States|same-sex marriage]] bans as unconstitutional (''[[United States v. Windsor]]'' and ''[[Obergefell v. Hodges]]'');
```

https://www.informatik.tu-darmstadt.de/ukp/research_6/data/lexical_resources/wikimwe/index.en.jsp


The STANDS4 Network
Login
PHRASES
Search for phrases...
SEARCH
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
NEW
RANDOM
The Web's Largest Resource for
Phrases, Verbs & Idioms
A MEMBER OF THE STANDS4 NETWORK
Welcome to Phrases.com
Like 496
Phrases.com is a large collection of common phrases, casual expressions and idioms – collaboratively assembled by our contributing editors. Entries can be easily browsed, searched for, rated, heard and translated to several common and not so common languages.
Improve your English and writing skills by navigating our comprehensive phrases dictionary alphabetically, or simply search by keywords.
Hot
Our most popular phrases
ma vaillante fille 263
I'll see you and ... 117
ain't no mountain... 106
Rate it: (3.00 / 1 vote)
stand for
To tolerate.
2 Views
COLLECTION
EDIT

MWE Extraction

- In many cases, existing MWE lexica don't cover the specific MWE present in a new domain.
- Several methods for extracting MWE from a corpus.

Collocations

“An arbitrary and recurrent word combination”

[Benson 1990; Baldwin and Kim 2010]

χ^2

- χ^2 (chi-square) is a statistical test of dependence—-here, dependence between the two variables of word 1 identity and word2 identity.

- For assessing the difference in two datasets, this test assumes a 2x2 contingency table:

	word 1	¬word 1
word 2	7	104023
¬word 2	104	251093

X^2

To test whether “white house” is a meaningful collocation, we can ask: does the word *house* occur significantly more frequently after *white*?

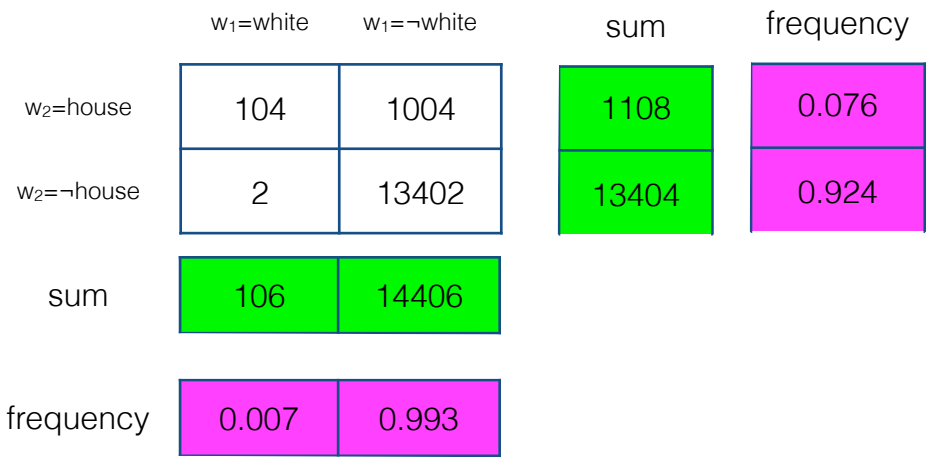
	w_1=white	w_1=¬white	
w_2=house	104	1004	“red house”, “my house”
w_2=¬house	2	13402	“red car”, “my dog”

“white dog”, “white truck”

X^2

For each cell in contingency table, sum the squared difference between observed value in cell and the expected value assuming independence.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$


w1=white
w1=¬white
sum
frequency
w2=house
104
1004
1108
0.076
w2=¬house
2
13402
13404
0.924
sum
106
14406
frequency
0.007
0.993

Assuming independence:

$$P(w_1 = \text{white}, w_2 = \text{house}) = P(w_1 = \text{white}) \times P(w_2 = \text{house})$$
$$= 0.007 \times 0.076 = 0.00053$$

Among 14512 words, we would expect to see 7.69 occurrences of *white house*.

	w_1=white	w_1=¬white	
w_2=house	7.69	1095.2	$P(w_2 = \text{house})$ 0.076
w_2=¬house	93.9	13315.2	$P(w_2 = \neg\text{house})$ 0.924

$P(w_1 = \text{white})$	$P(w_1 = \neg\text{white})$
0.007	0.993

χ^2

- What χ^2 is asking is: how different are the observed counts different from the counts we would expect given complete independence?

	w_1=white	w_1=¬white
w_2=house	104	1004
w_2=¬house	2	13402

	w_1=white	w_1=¬white
w_2=house	7.69	1095.2
w_2=¬house	93.9	13315.2

X^2

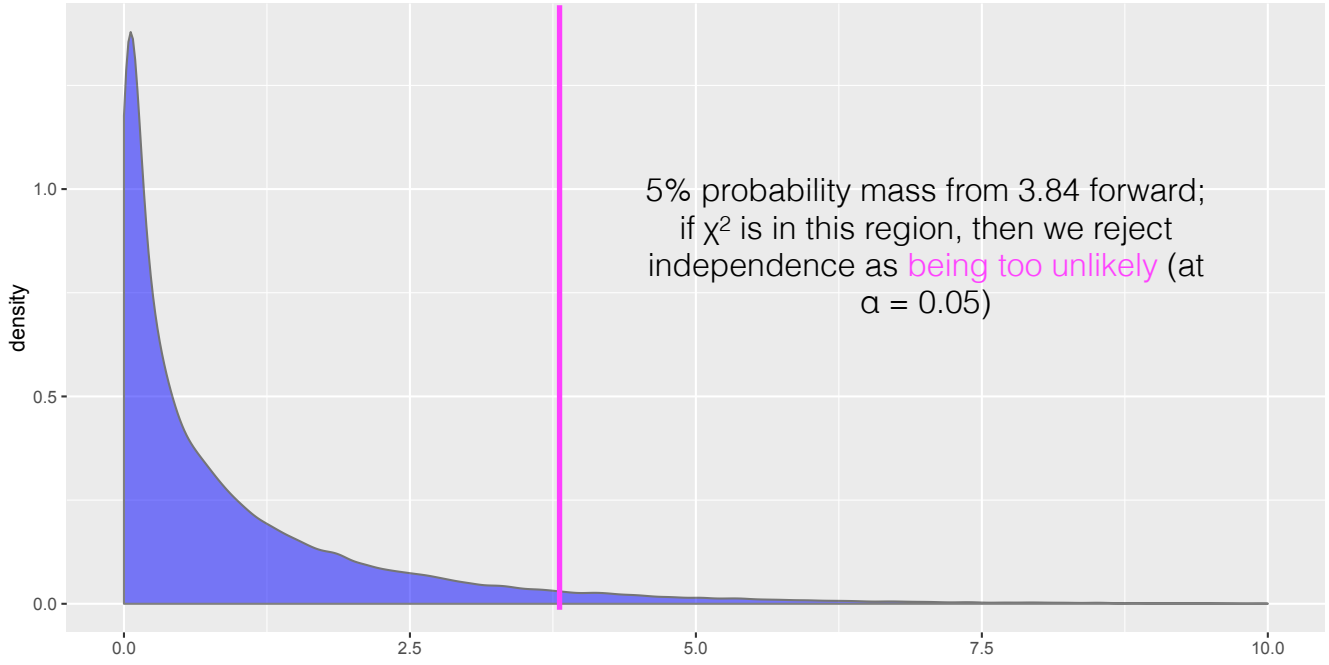
- With algebraic manipulation, simpler form for 2x2 table O (cf. Manning and Schütze 1999)

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

χ^2

- The χ^2 value is a statistic of dependence with a probability governed by a χ^2 distribution; if this value has low enough probability in that measure, we can reject the null hypothesis of the independence between the two variables.

χ^2


density
1.0
0.5
0.0
0.0
2.5
5.0
7.5
10.0
5% probability mass from 3.84 forward;
if χ^2 is in this region, then we reject
independence as being too unlikely (at
$\alpha = 0.05$)

X^2

- Chi-square is ubiquitous in corpus linguistics (and in NLP as a measure of collocations).
- A few caveats for its use:
    - Each cell should have an *expected* count of at least 5
    - Each observation is independent

Why is part of speech tagging useful?

POS indicative of MWE

at least one adjective/noun or noun phrase	and definitely one noun

((A | N)+ | ((A | N)*(NP))(A | N)*)N

AN: linear function; lexical ambiguity; mobile phase
NN: regression coefficients; word sense; surface area
AAN: Gaussian random variable; lexical conceptual paradigm; aqueous mobile phase
ANN: cumulative distribution function; lexical ambiguity resolution; accessible surface area
NAN: mean squared error; domain independent set; silica based packing
NNN: class probability function; text analysis system; gradient elution chromatography
NPN: degrees of freedom; [*no example*]; energy of adsorption

Justeson and Katz 1995

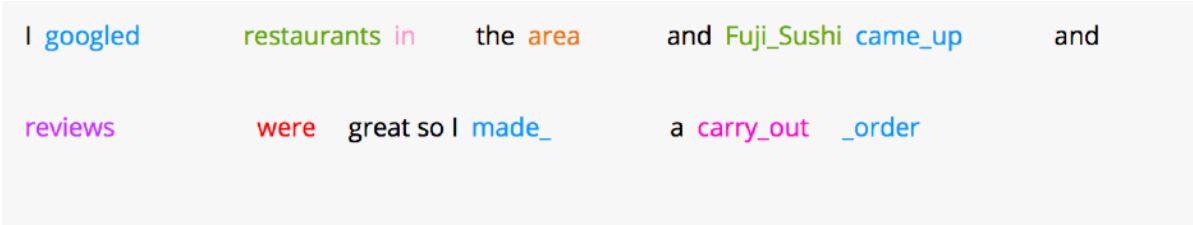
MWE prediction

- Many phrases are ambiguous about whether they are a MWE in context.
  - the white house pledged to reduce red tape
  - he lives in the white house on the corner
  - Kim made a face at the policeman.
  - Kim made a face in pottery class.

Baldwin and Kim (2010)

MWE prediction

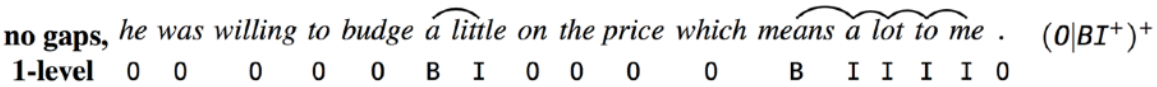
- Data: 55,000 tokens of web reviews annotated for MWE in context.



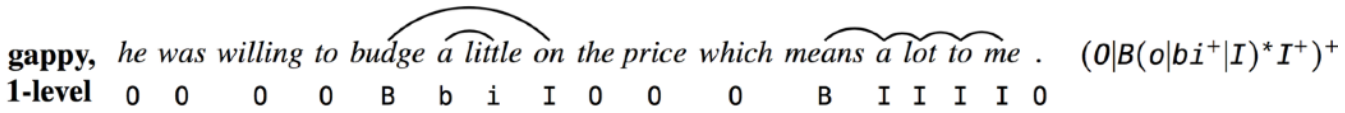
https://github.com/nert-nlp/streusle/

BIO notation

Standard BIO entity notation



Expanded BIO to accomodate one layer of nesting



MWE prediction

			LOOKUP				SUPERVISED MODEL			
preexising lexicons	entries	max gap length	$\bar{P}$	$\bar{R}$	$\overline{F_1}$	σ	$\bar{P}$	$\bar{R}$	$\overline{F_1}$	σ
none	0						74.39	44.43	55.57	2.19
WordNet + SemCor	71k	0	46.15	28.41	35.10	2.44	74.51	45.79	56.64	1.90
6 lexicons	420k	0	35.05	46.76	40.00	2.88	76.08	**52.39**	**61.95**	1.67
10 lexicons	437k	0	33.98	47.29	39.48	2.88	75.95	51.39	61.17	2.30
best configuration with in-domain lexicon		1	**46.66**	**47.90**	**47.18**	2.31	**76.64**	51.91	61.84	1.65
			2 lexicons + *MWtypes*(train)$_{\geq 1}$				6 lexicons + *MWtypes*(train)$_{\geq 2}$			

Schneider et al. (2014), "Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut" (TACL)

Activity

`12.mwe/JustesonKatz95_Topics.ipynb`

- Explore using POS regexes to find multiword expressions in a collection of Wikipedia articles. What happens when you use multiword expressions in a topic model?