Discriminative Lexical Semantic Segmentation with Gaps:

Running the MWE Gamut

Nathan Schneider • August 27, 2013

Opiliones

daddy longlegs

harvestman

Kevin Knight →

Weberknechte

Schuster

Kanker

Opa Langbein

Zimmermann

Schneider

Opiliones

daddy longlegs

harvestman

Kevin Knight →

Weberknechte

Schuster

Kanker

Opa Langbein

Zimmermann

Schneider

The aliens will destroy Earth unless we

accept

agree to
accede to
yield to
give in to

comply with
cooperate with
go along with

their demands.


give_in_to
daddy_longlegs
Kevin_Knight
Jonathan Huang

Kevin Knight refused to give in to the vicious daddy longlegs .

Kevin_Knight refused to give in to the vicious daddy longlegs .

Kevin_Knight refused to give_in_to the vicious daddy longlegs .

Kevin_Knight refused to give_in_to the vicious daddy_longlegs .

Lexical segmentation

Kevin_Knight

refused

to

give_in_to

the

vicious

daddy_longlegs

.

Roadmap

- MWEs in NLP
  - What are they?
  - Why are they important?
  - Why are they challenging?
  - How are they handled?
- Corpus annotation
- Sequence tagging formulation & experiments

Definition

- **Multiword expression** (MWE): 2 or more orthographic words/lexemes that function together as an **idiomatic whole**
- *idiomatic* = not fully predictable in form, function, and/or frequency
  - *unusual morphosyntax*: **Me/*Him neither**; **by and large**; plural of **daddy longlegs**?
  - *non-* or *semi-compositional*: **ice cream**, **daddy longlegs**, **pay attention**
  - statistically *collocated*: *p*(**highly unlikely**) > *p*(**strongly unlikely**)

Definition

- **Multiword expression** (MWE): 2 or more orthographic words/lexemes that function together as an **idiomatic whole**
- *idiomatic* = not fully predictable in form, function, and/or frequency
  - *unusual morphosyntax*: **Me/*Him neither**; **by and large**; plural of **daddy longlegs**?
  - *non- or semi-compositional*: **ice cream**, **daddy longlegs**, **pay attention**
  - statistically *collocated*: *p*(**highly unlikely**) > *p*(**strongly unlikely**)

SPECIALLY LEARNED

Applications

- **semantic analysis:** minimal meaning-bearing units (e.g., predicates)
    - **named entity recognition**, **supersense tagging** already target some kinds of MWEs
    - **sentiment analysis:** MW opinion expressions & opinion targets
- **IR:** keyphrase extraction, query segmentation
- **MT:** decomposing MWEs in translation often incorrect or more ambiguous
- **language acquisition:** many MWEs are difficult for learners

Challenges

- Not superficially apparent in text
- Number/frequency
  - Too many expressions to list all of them
  - Individually rare, but frequent in aggregate
- Diversity
  - Many different construction types
  - Semantically unrestricted
  - Can be **gappy**

Kevin Knight

daddy longlegs, hot dog

dry out the clothes

depend on

no attention was paid (to)

pay attention (to)

put up with, give in (to)

under the weather

cut and dry

in spite of

pick up where they left off

easy as pie

You’re welcome.

To each his own.

Current state of affairs

Resource-building

- lexicons (e.g., WordNet, WikiMwe), grammars
- corpora: treebanks (French Treebank, Prague Czech-English Dependency Treebank)

Explicit

- **Corpus → List:** collocation extraction by word association measures
- **List → Corpus:** matching, classification
- **Corpus (+ List):** sequence modeling, parsing

Implicit

- language modeling, phrase-based MT

Current state of affairs

Resource-building

- lexicons (e.g., WordNet, WikiMwe), grammars
- corpora: treebanks (French Treebank, Prague Czech-English Dependency Treebank)

Explicit

- **Corpus → List:** collocation extraction by word association measures
- **List → Corpus:** matching, classification
- **Corpus (+ List):** sequence modeling, parsing

Implicit

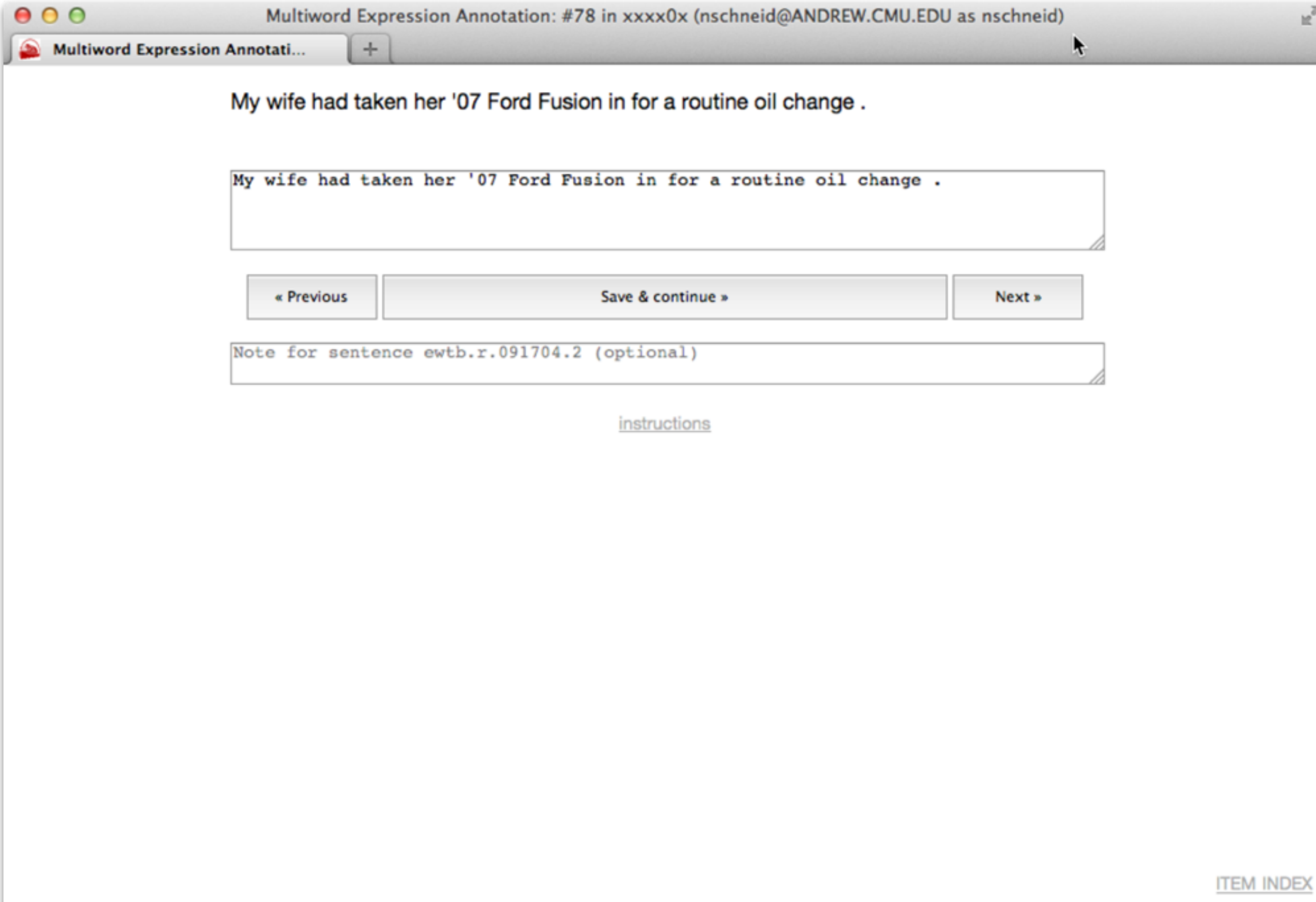
- language modeling, phrase-based MT

Contributions

- **Our goal:** general-purpose, shallow, automatic identification of MWEs in context
- Existing **resources** are not satisfactory.
  - New corpus—first freely annotated for MWEs, without a preexisting lexicon.
- Existing discriminative sequence modeling techniques do not handle **gaps**.
  - New gappy tagging scheme + model trained and evaluated on our annotated corpus.

Roadmap

- ✓ MWEs in NLP
  - What are they?
  - Why are they important?
  - Why are they challenging?
  - How are they handled?
- Corpus annotation
- Sequence tagging formulation & experiments


Multiword Expression Annotation: #78 in xxxx0x (nschneid@ANDREW.CMU.EDU as nschneid)
Multiword Expression Annotati...
My wife had taken her '07 Ford Fusion in for a routine oil change .
My wife had taken her '07 Ford Fusion in for a routine oil change .
« Previous
Save & continue »
Next »
Note for sentence ewtb.r.091704.2 (optional)
instructions
ITEM INDEX

Examples

My wife had taken_ her '07_Ford_Fusion _in for a routine oil_change .

The corpus

- The entire **Reviews** subsection of the English Web Treebank (Bies et al. 2012), fully annotated for MWEs
  - 723 reviews
  - 3,800 sentences
  - 55,000 words
- Every sentence: negotiated consensus between at least 2 annotators
  - IAA between *pairs*: ~77%

Examples

Among the animals that were available to touch were pony's , camels and EVEN AN OSTRICH !!!

No MWEs here. (This sentence is in the minority: 57% of all sentences/72% >10 words contain an MWE.)

Examples

They gave me the run around and missing paperwork only to call back to tell me someone else wanted her and I would need to come in and put down a deposit .

Examples

It put_hair_on_ my _chest and thanks_to the owner s advice I invested vanguard , got myself a woman like Jerry , and became a republican .

Examples

They gave_ me _the_run_around and missing paperwork only to call_back to tell me someone else wanted her and I would need to come_in and put_down a deposit .

Simplified a bit for presentational purposes
(we also made a strong/weak distinction)

Examples

I highly~recommend Debi , she does~ an amazing ~job , I " love " the way she cuts_ my _hair , extremely thorough and cross_checks her work to make_sure my hair is perfect .

Weak expressions: highly~recommend, does~job

Examples

I recently threw~ a surprise ~birthday_party for my wife at Fraiser_'s .

Weak expressions can contain strong MWEs.

Overlap: Ideally we'd have threw~party, birthday_party, surprise_party

Annotation guidelines

https://github.com/nschneid/nanni/wiki/MWE-Annotation-Guidelines

Roadmap

✓ MWEs in NLP

- What are they?
- Why are they important?
- Why are they challenging?
- How are they handled?

✓ Corpus annotation

- Sequence tagging formulation & experiments

Gappy sequence tagging

- Simplest tagger (our **baseline**):
  1. obtain MWE candidates from lexicons
  2. predict the segmentation with fewest total expressions
- We extract lexicons from 10 existing sources of MWEs
  - WordNet, SemCor, Prague Czech-English Treebank, SAID, WikiMwe, Wiktionary, and other lists

Gappy sequence tagging

- *Contiguous* MWE identification resembles chunking, so we can use the familiar `BIO` scheme (Ramshaw & Marcus 1995):

O	O	B	I	O
a	routine	oil_	change	.

- We add 3 new tags for gaps:

O	O	O	B	o	b	i	i	I
My	wife	had	taken_	her	'07_	Ford_	Fusion	_in

  - Assumption: no more than 1 level of nesting
- **Evaluation**: MWE precision/recall
  - MUC criterion: partial credit for partial overlap

Pathological examples

On August 3 , two massive headlands reared out_of the mists -- great gateways never~before~ , so_far_as~ Hudson ~knew , ~seen by Europeans .

Pathological examples

All you have to do to make it authentic Jamaican food , is add a_~whole~_lot of pepper .

Gappy sequence tagging

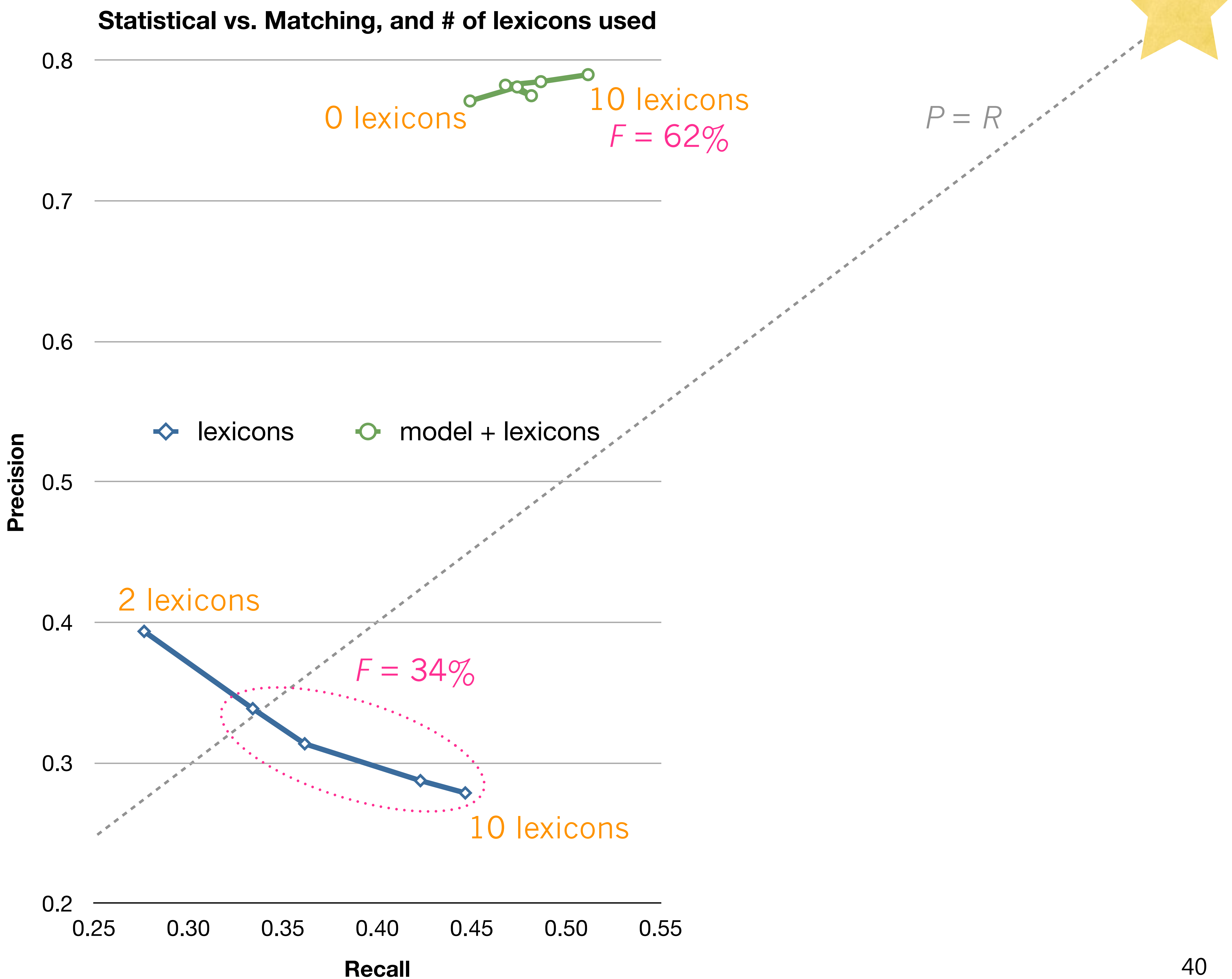
- Standard supervised learning with the enriched tagging scheme
- We use the **structured perceptron** (Collins 2002)
  - **Discriminative**
  - 1st-order Markov assumption
  - Averaging
  - Fast to train

Gappy sequence tagging

- **Basic features**
  adapted from Constant et al. (2012):
  - **word:** current & context, unigrams & bigrams
  - **gold POS:** current & context, unigrams & bigrams
  - capitalization; word shape
  - prefixes, suffixes up to 4 characters
  - has digit; non-alphanumeric characters
  - lemma + context lemma if one is a V and the other is ∈ {N, V, Adj., Adv., Prep., Part.}
- **Lexicon features:** WordNet & other lexicons

Gappy sequence tagging

- Experimental setup
  - Regularization by early stopping
  - 8-fold cross-validation; results are 8-way averages
  - Jon Clark's `ducttape`


Statistical vs. Matching, and # of lexicons used
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Precision
0.25
0.30
0.35
0.40
0.45
0.50
0.55
Recall
lexicons
model + lexicons
0 lexicons
10 lexicons
F = 62%
P = R
2 lexicons
F = 34%
10 lexicons

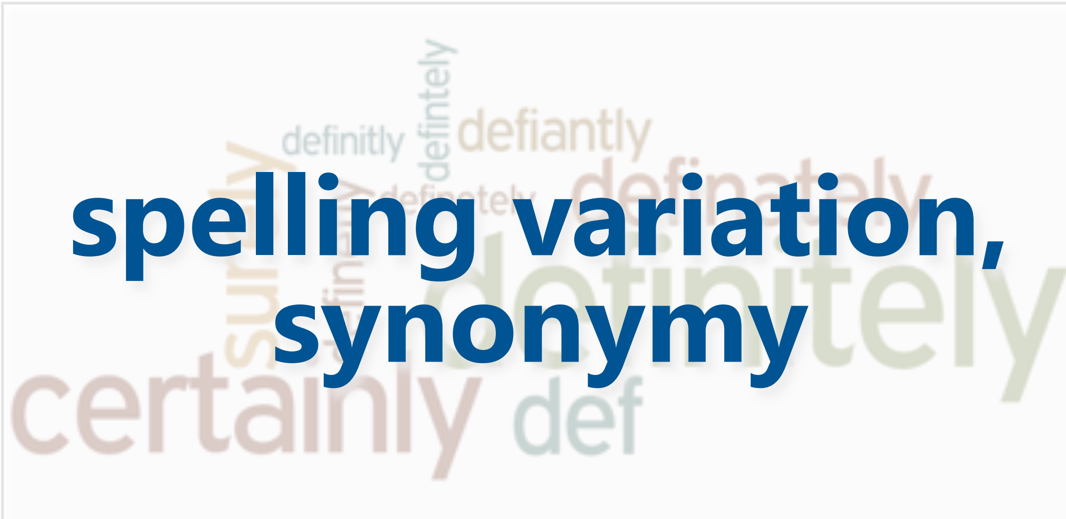
(all use 10 lexicons)	**P**	**R**	**F**
Baseline: lexicon matching	0.279	0.446	0.342
Sequence model	0.790	0.511	0.618

Word clusters

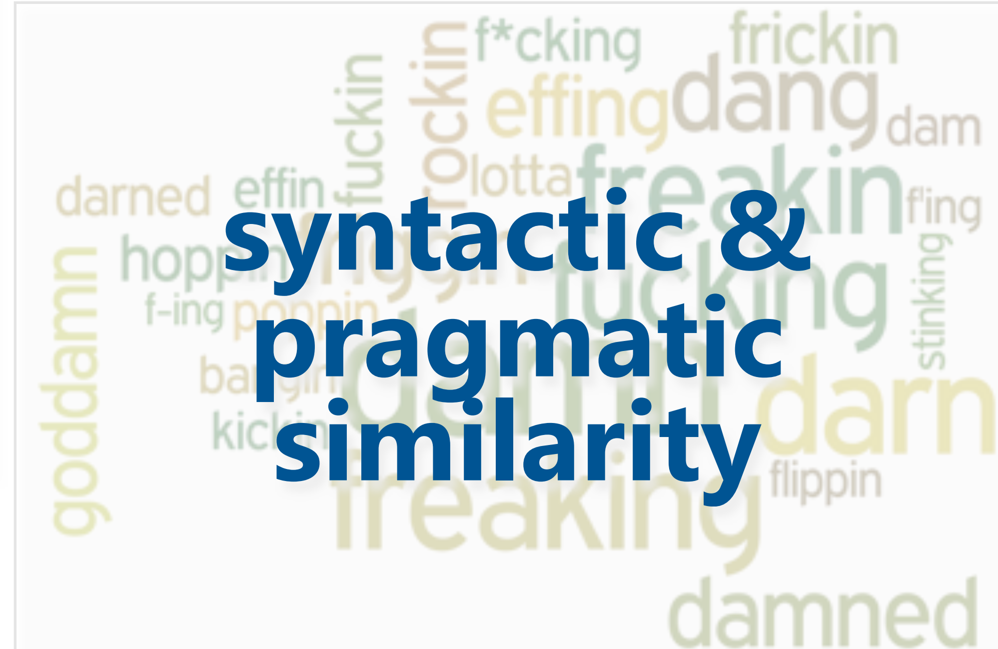
- **Brown** clusters (Brown et al. 1992)
  - latent word categories explaining observed sequences
  - hard assignment: each word goes in 1 cluster
  - agglomerative, greedy, scalable algorithm
- 1000 from reviews in the Yelp Academic Dataset (20.7M words)
  - words occurring ≥25 times
  - Percy Liang's implementation


definitely
certainly
def
definately
surely
definetly
definitly
defintely
defiantly
definetely

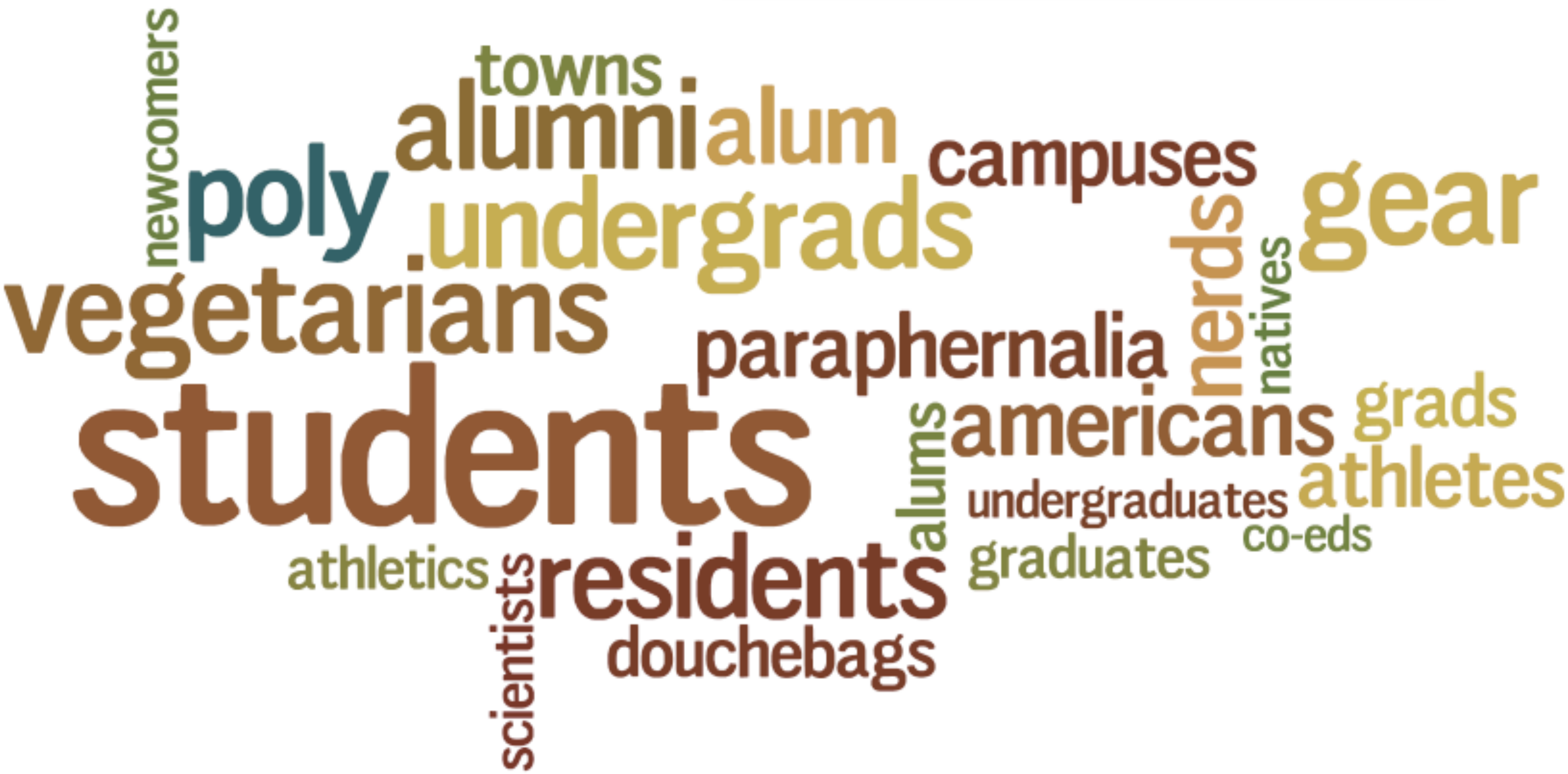
spelling variation, synonymy



spelling variation, synonymy

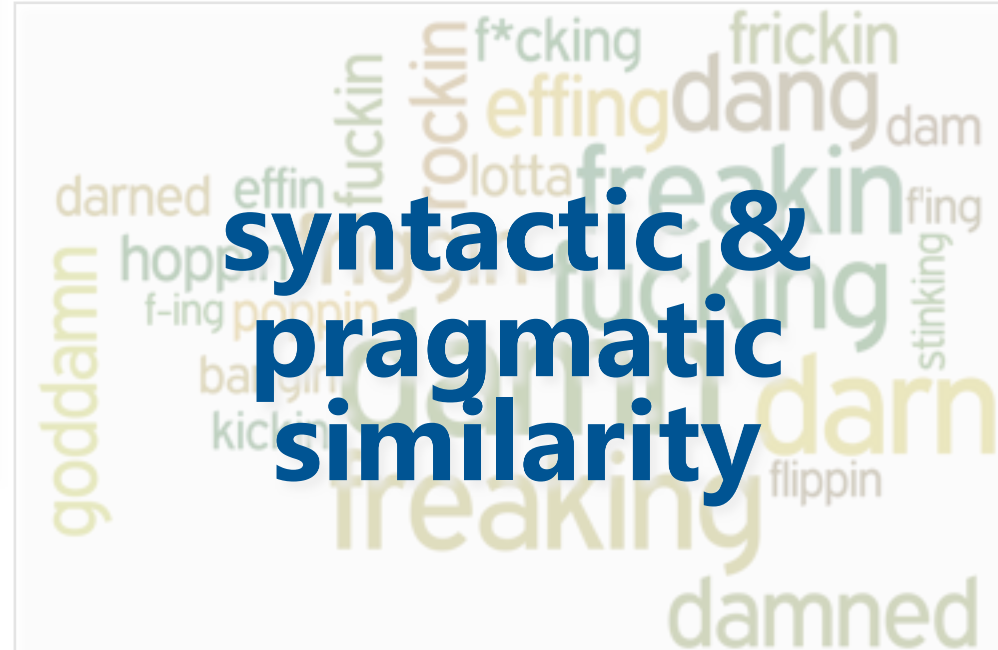


syntactic & pragmatic similarity



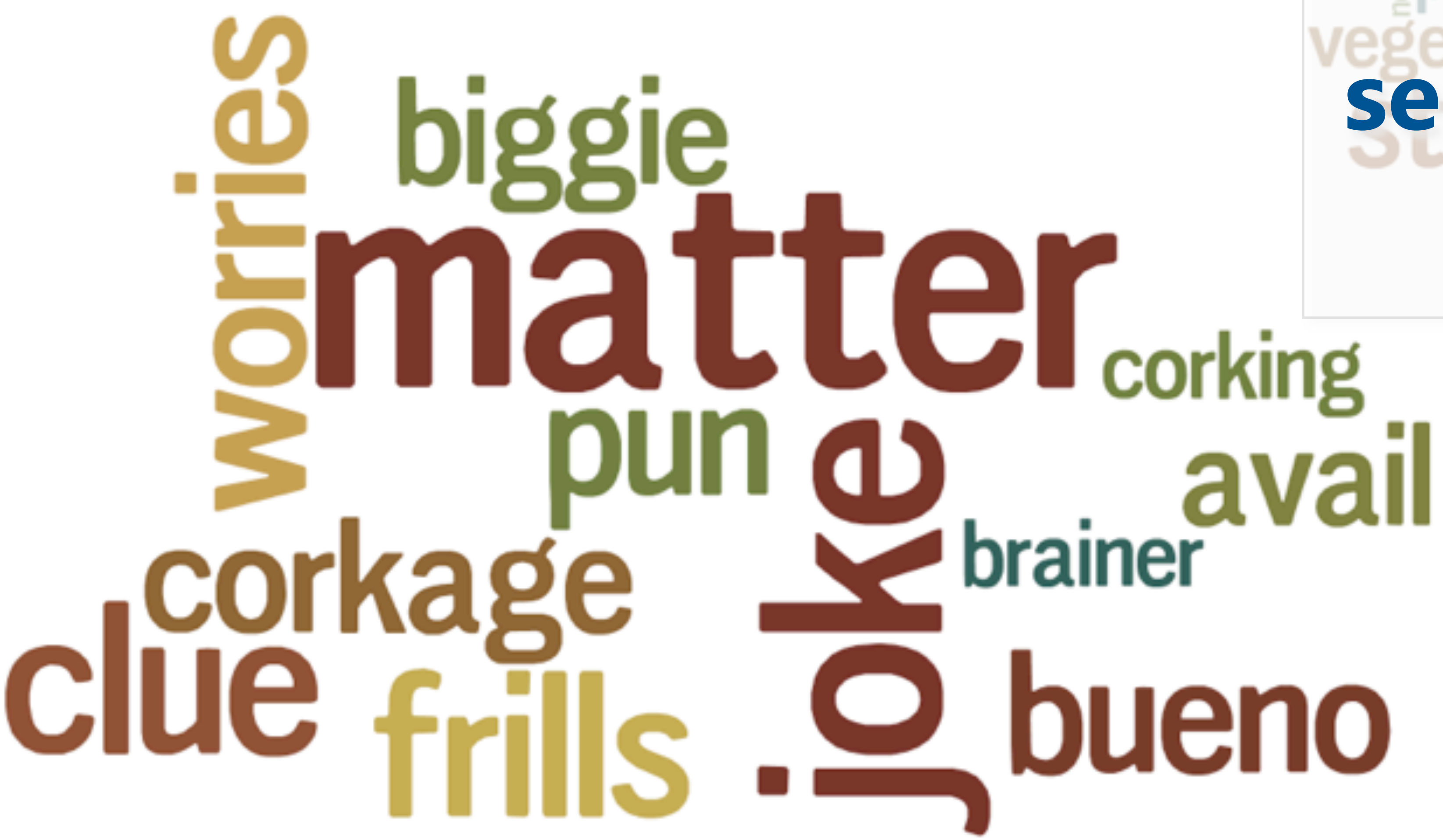
spelling variation, synonymy

syntactic & pragmatic similarity



semantic category

idiosyncratic lexical context



(all use 10 lexicons)	**P**	**R**	**F**
Baseline: lexicon matching	0.279	0.446	0.342
Sequence model	0.790	0.511	0.618
+ Brown clusters	0.790	0.515	0.624

Word association scores

- large literature on statistical measures of **collocation**
  - *information theoretic:* mutual information, ...
  - *frequentist:* t–statistic, χ^2, ...
- scores → rankings → rank threshold features

1. POS tag the Yelp Academic Dataset with the Twitter tagger (Owoputi et al. 2013)
2. Define 2-word patterns of interest:
   Adj. N, N N, Prep. N, V N, V Prep., V Particle
3. Use `mwetoolkit` (Ramisch et al. 2010) to identify, score (t), and rank each group of candidates

(all use 10 lexicons)	**P**	**R**	**F**
Baseline: lexicon matching	0.279	0.446	0.342
Sequence model	0.790	0.511	0.618
+ Brown clusters	0.790	0.515	0.624
+ `mwetoolkit` word associations	0.793	0.511	0.621

Recall-oriented learning

- Our supervised learner is actually optimizing for *tag accuracy*, not *expression precision/recall*
    - This tends to hurt recall, because (short of strong evidence) the safest tag is O
- A **recall-oriented** cost function can compensate by biasing in favor of recall (Mohit et al. 2012), improving the *F* score
    - Tunable hyperparameter controls the strength of this preference

(all use 10 lexicons)	**P**	**R**	**F**
Baseline: lexicon matching	0.279	0.446	0.342
Sequence model	0.790	0.511	0.618
+ Brown clusters	0.790	0.515	0.624
+ `mwetoolkit` word associations	0.793	0.511	0.621
+ recall-oriented learning	0.700	0.596	0.645

Error analysis

- Cross-gap recall: 155/466 = 33%

✓ unseen TPs:

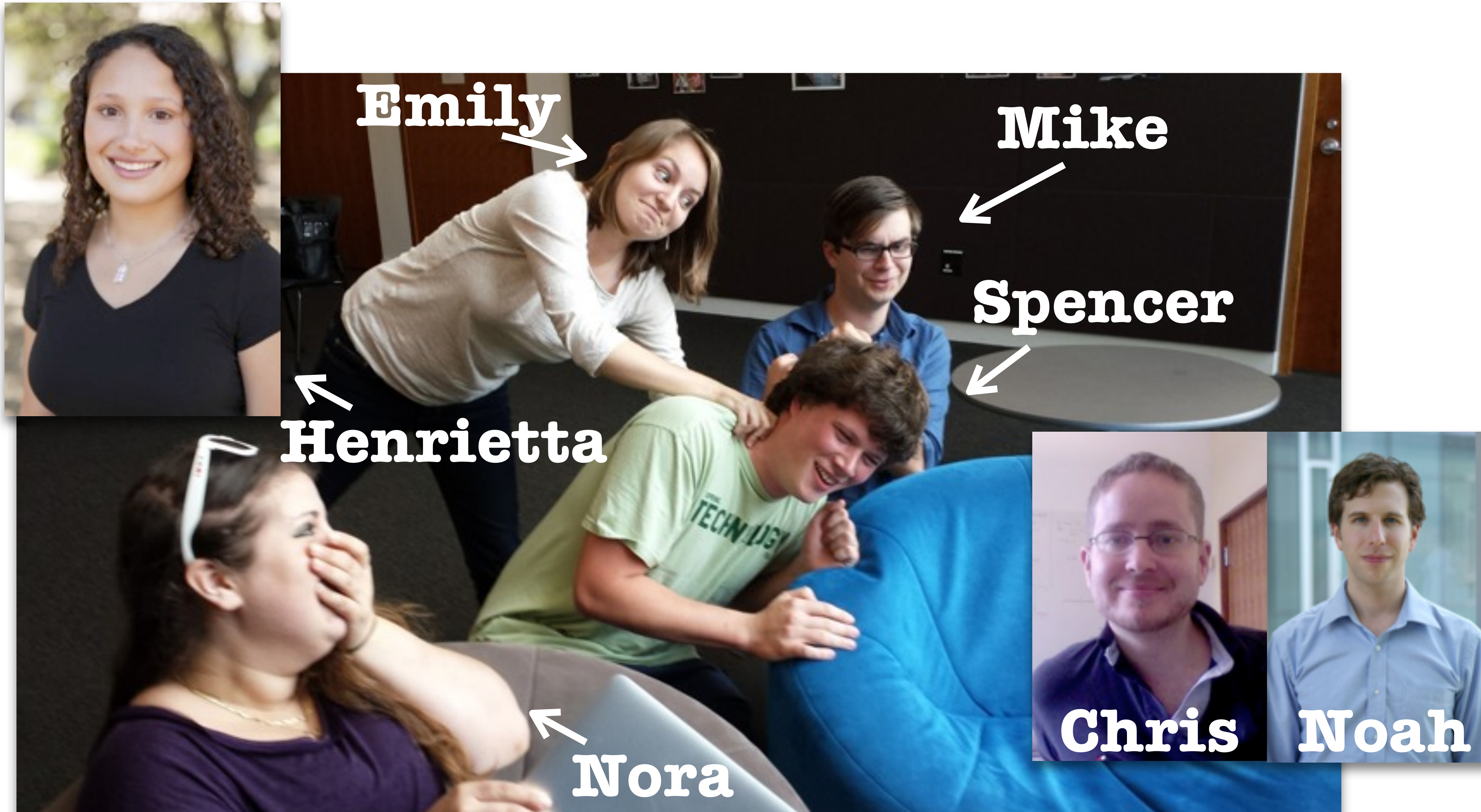
above all	associate with	check - in
allen tire	at peace	cleaning lady
amusement parks	behind the scene	come up
antipasto misto	brand new	cowboy boot
aortic stenosis	carnegie mellon	cup of joe

✗ unseen FPs: a little girl, **bad for**, cigarette smoke, funeral director, get coupon, kitchen sink

✗ unseen FNs: an arm and a leg, **bad for business**, child predator, dfw metro area

Conclusions

- Multiword expressions are important and challenging
- We can shallowly mark them in free text
  - new corpus resource!
- MWE identification can be modeled as sequence tagging
  - even with gaps!
  - statistical learning » lexicon-based segmentation
  - but lexical resources are still useful (features!)



Many_thanks
(*Several thanks)

Thanks_a_million
(*Thanks a thousand)

Thanks_a_lot
(?Lots of thanks)