Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments

Kevin Gimpel, Nathan Schneider, Brendan O'Connor,
Dipanjan Das, Daniel Mills, Jacob Eisenstein,
Michael Heilman, Dani Yogatama, Jeffrey Flanigan,
and Noah A. Smith

Why does this paper have so many authors?

Why does this paper have so many authors?

Our goal:

Build a Twitter part-of-speech tagger in one day

- Plan:
  - Large team of annotators
  - Simple, carefully-designed annotation scheme
  - Features leveraging existing resources (treebanks) and unannotated data

- Plan:
  - Large team of annotators
  - Simple, carefully-designed annotation scheme
  - Features leveraging existing resources (treebanks) and unannotated data

- Outcome:
  - Tag set for Twitter
  - 1,827 annotated English tweets
  - POS tagger with ~90% accuracy
  - Didn't finish in a day, but took < 250 person-hours

Available to download!

The Data

Noah Smith
@nlpnoah Pittsburgh, PA
http://www.cs.cmu.edu/~nasmith

omg, first tweet evar! I'm in the green room at #SXSW getting ready for my panel, #textworld

13 Mar via web

Favorite Retweet Reply

Also: at-mentions, URLs, emoticons, symbols, typos, etc.

ARK lti Carnegie Mellon

Tag Set

- Start with **coarse** set of Penn Treebank tags
- Add Twitter-specific tags

- Coarse treebank tags:

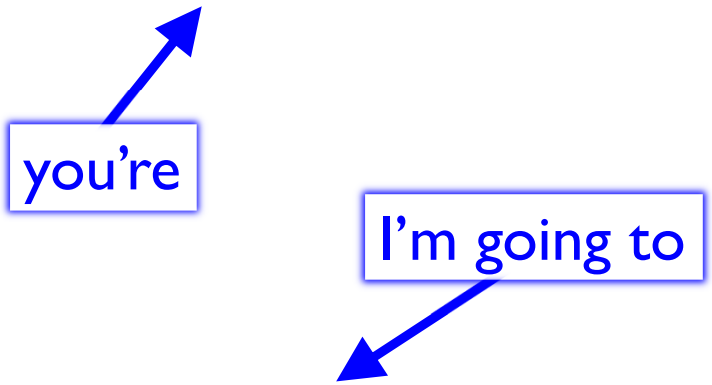
common noun	determiner
proper noun	preposition
pronoun	verb particle
verb	coordinating conjunction
adjective	numeral
adverb	interjection
punctuation	predeterminer / existential *there*

- Coarse treebank tags:

common noun	determiner
proper noun	preposition
pronoun	verb particle
verb	coordinating conjunction
adjective	numeral
adverb	interjection
punctuation	predeterminer / existential *there*

Penn Treebank tokenization is unsuitable for Twitter:

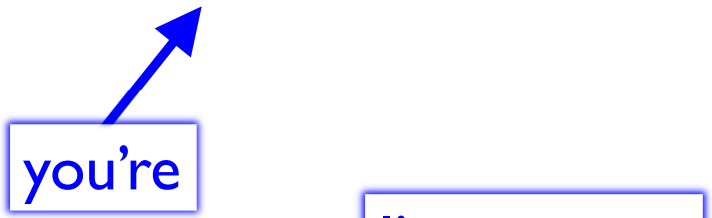
```
@user1 OMG ur from PA ? i am too (: where abouts ?
```




```
@user2 ima get me a flip phone for real
```

Penn Treebank tokenization is unsuitable for Twitter:

```
@user1 OMG ur from PA ? i am too (: where abouts ?
```



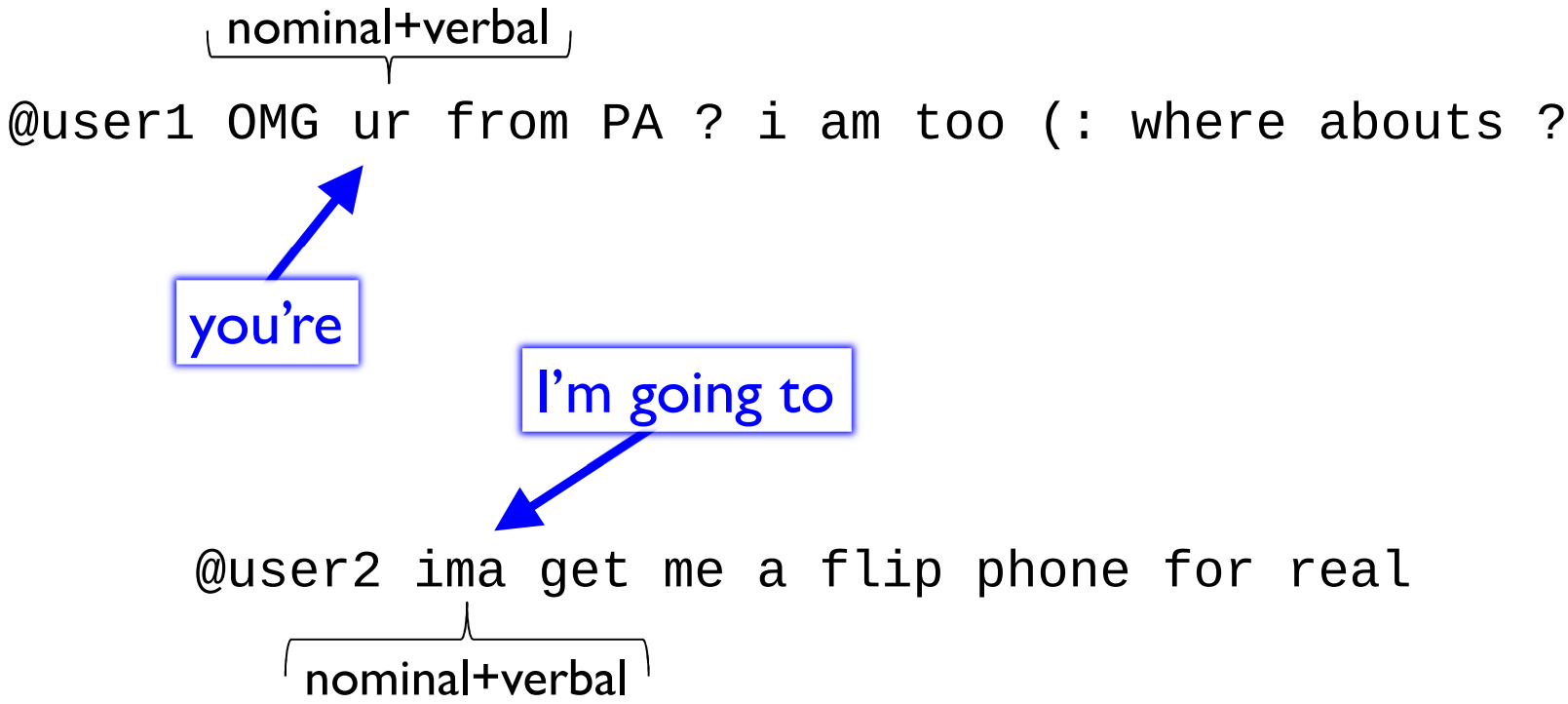

```
@user2 ima get me a flip phone for real
```

Solution:

Don't try to tokenize these

Instead, introduce **compound tags**

Penn Treebank tokenization is unsuitable for Twitter:



Solution:

Don't try to tokenize these

Instead, introduce **compound tags**

ARK lti Carnegie Mellon

- Twitter-specific tags:

  hashtag

  at-mention

  URL / email address

  emoticon

  Twitter discourse marker

  other (multi-word abbreviations, symbols, garbage)

- Twitter-specific tags:

  hashtag

  at-mention

  URL / email address

  emoticon

  Twitter discourse marker

  other (multi-word abbreviations, symbols, garbage)

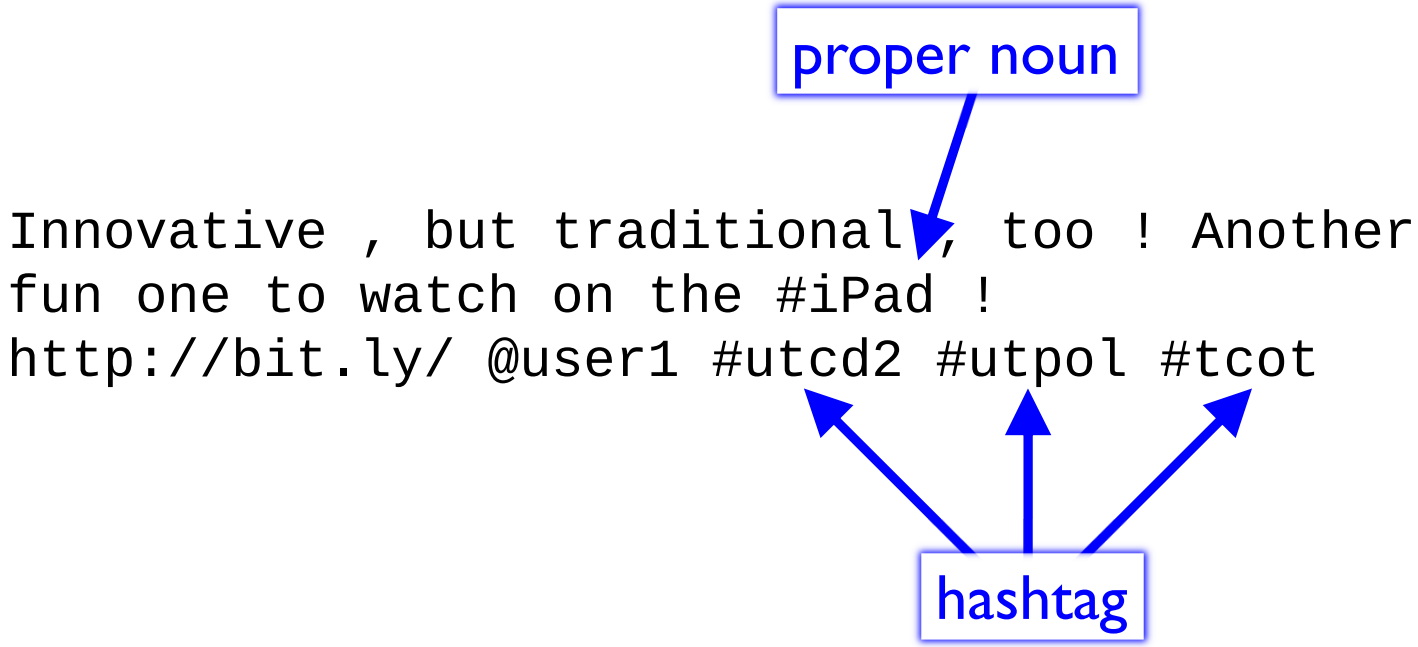
Hashtags

Twitter hashtags are sometimes used as ordinary words (35% of the time) and other times as topic markers

```
Innovative , but traditional , too ! Another
fun one to watch on the #iPad !
http://bit.ly/ @user1 #utcd2 #utpol #tcot
```

Hashtags

Twitter hashtags are sometimes used as ordinary words (35% of the time) and other times as topic markers



We only use "hashtag" for topic markers

ARK lti Carnegie Mellon

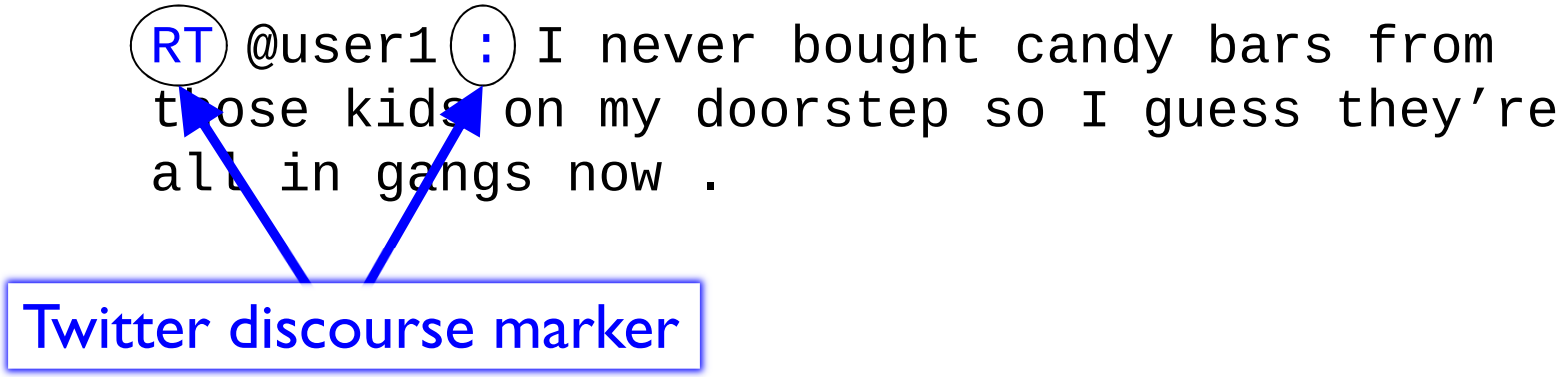
Twitter Discourse Marker

Retweet construction:

```
RT @user1 : I never bought candy bars from
those kids on my doorstep so I guess they’re
all in gangs now .
```

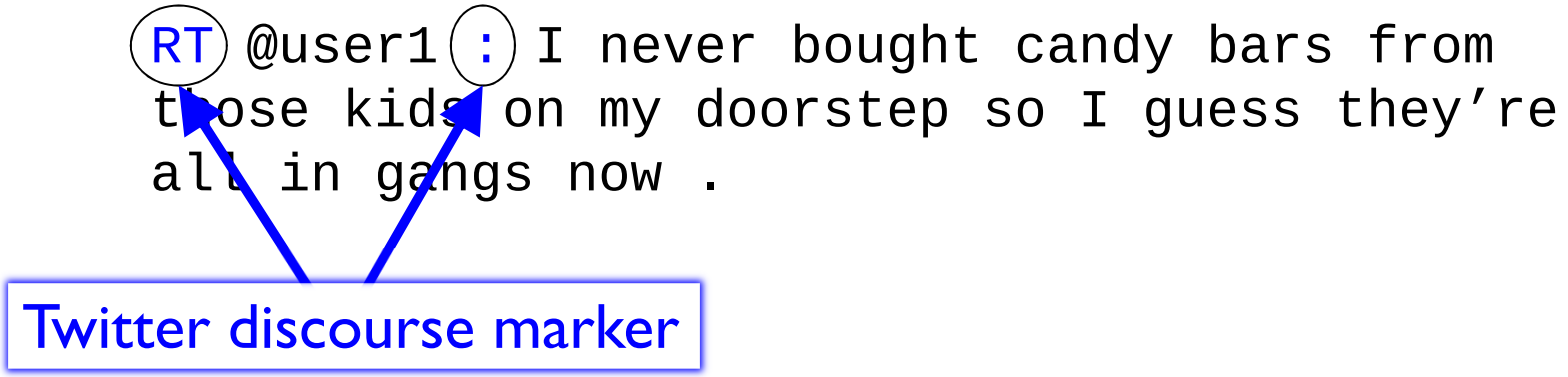
Twitter Discourse Marker

Retweet construction:



Twitter Discourse Marker

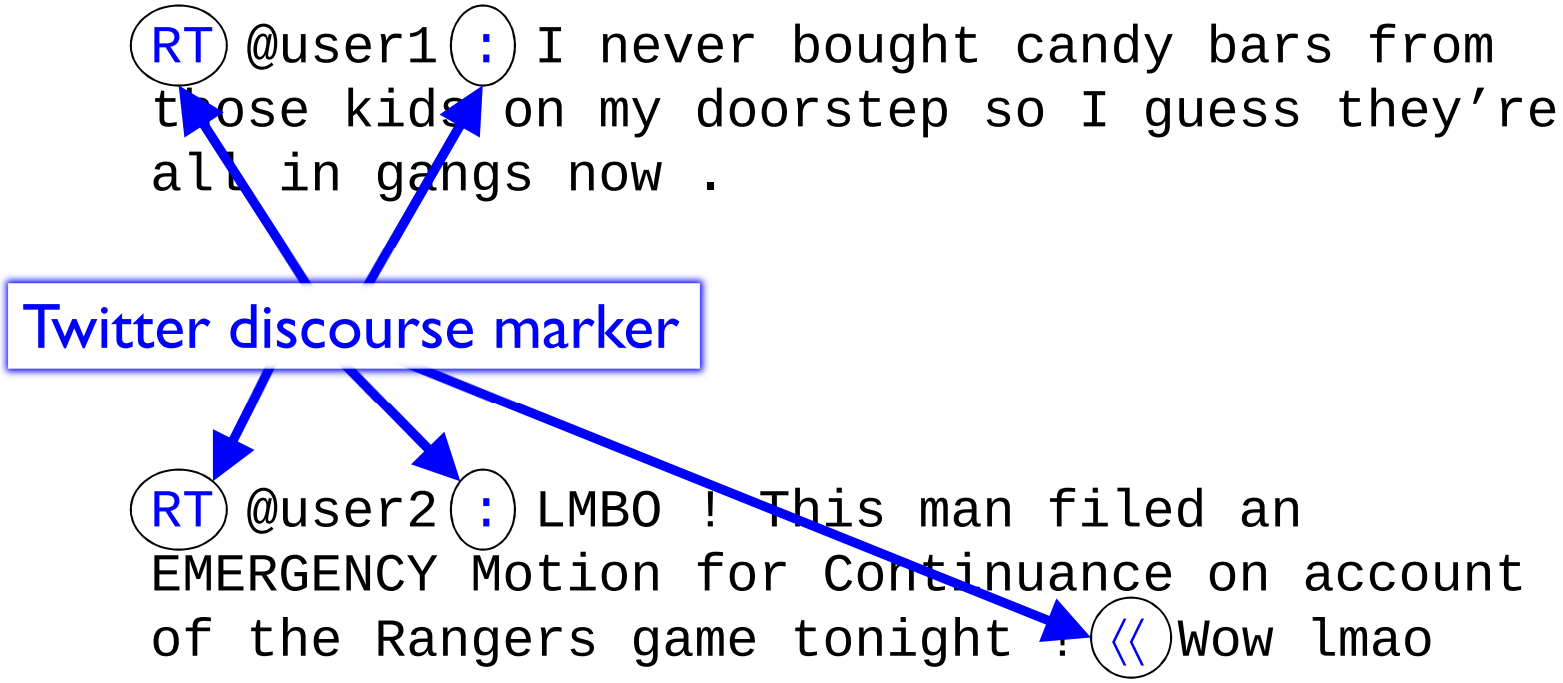
Retweet construction:



RT @user2 : LMBO ! This man filed an EMERGENCY Motion for Continuance on account of the Rangers game tonight ! ⟨⟨ Wow lmao

Twitter Discourse Marker

Retweet construction:



- Resulting tag set: 25 tags

Annotation

- 17 researchers from Carnegie Mellon

- Each spent 2-20 hours annotating
- Annotators corrected output of Stanford tagger
- Penn Treebank consulted for difficult cases

ARK lti Carnegie Mellon

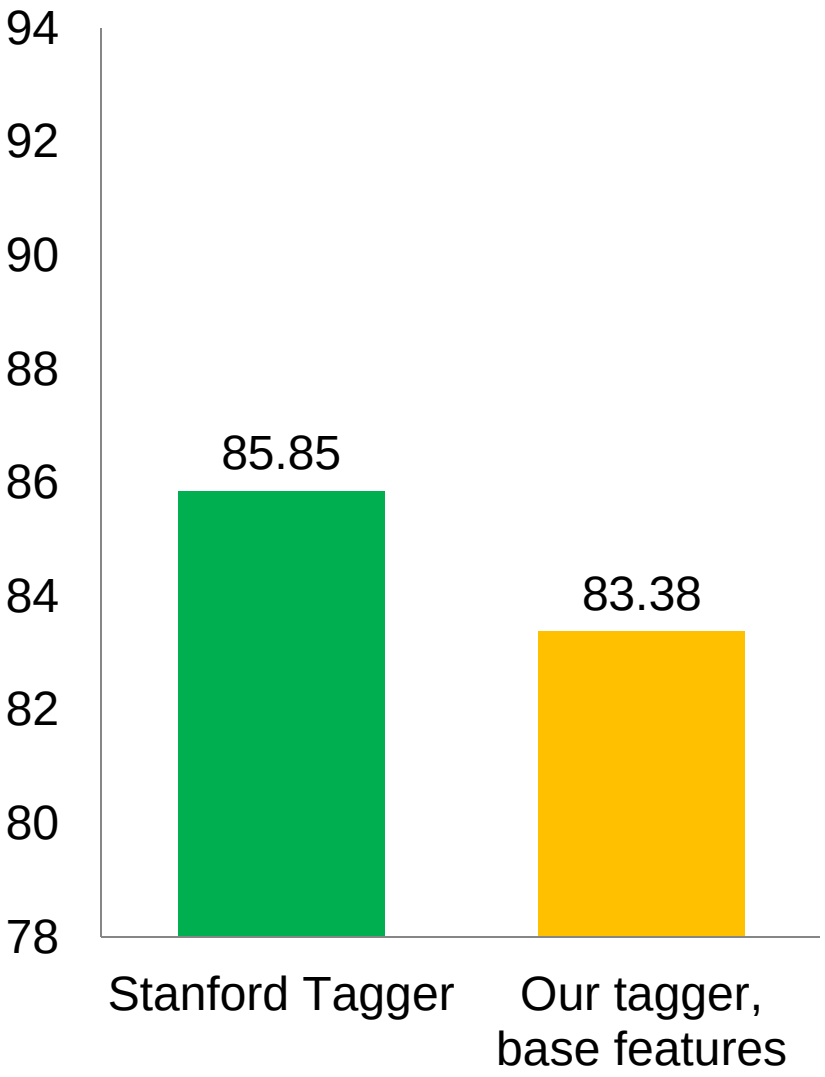
- Two annotators corrected and standardized annotations from the original 17 annotators
- A third annotator tagged a sample of the tweets from scratch
  - Inter-annotator agreement: 92.2%
  - Cohen's kappa: 0.914
- One annotator made a single final pass through the data, correcting errors and improving consistency

Experiments

Experimental Setup

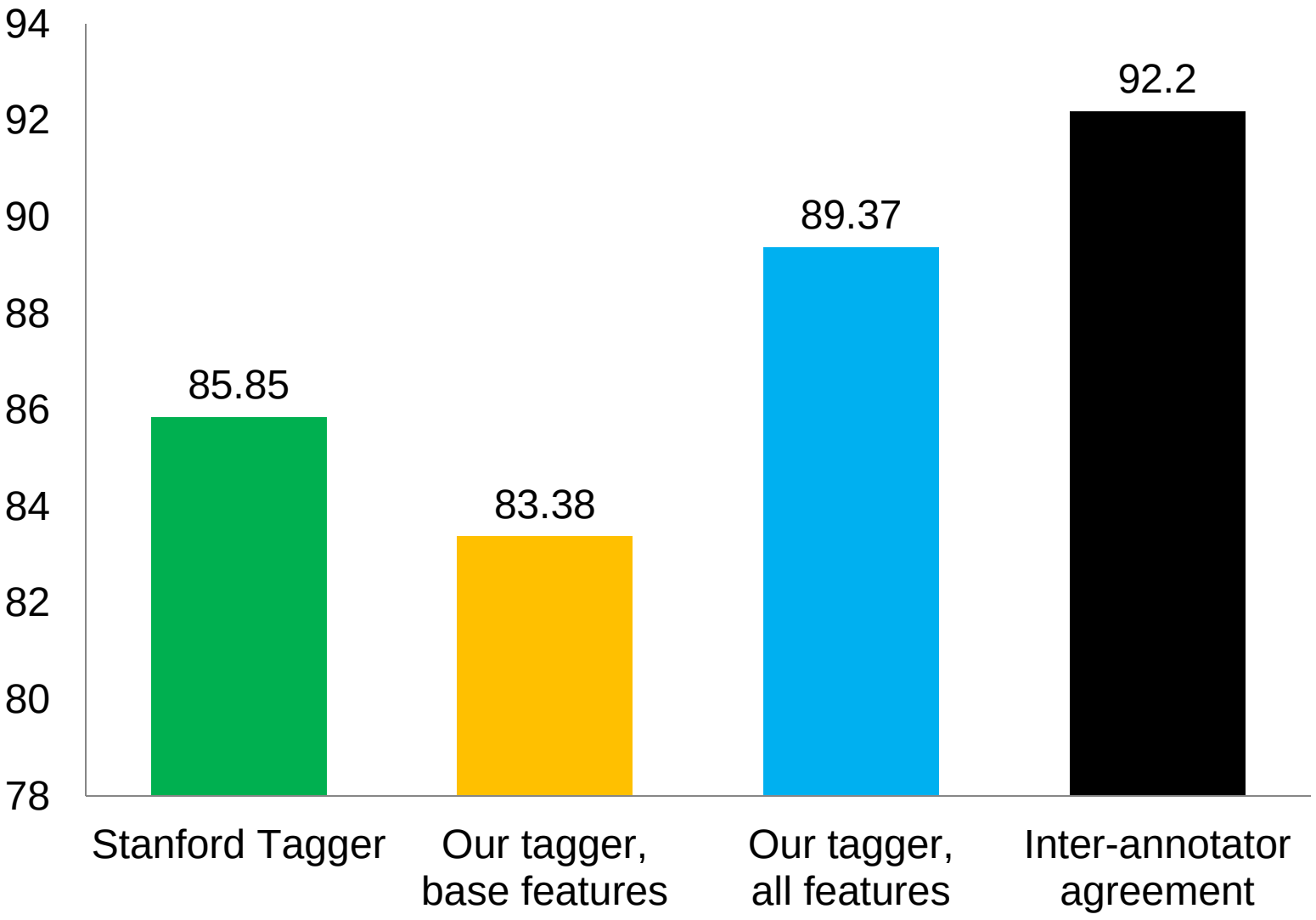
- 1,827 annotated tweets
  - 1,000 for training
  - 327 for development
  - 500 for testing (OOV rate: 30%)
- Systems:
  - Stanford tagger (retrained on our data)
  - Our own baseline CRF tagger
  - Our tagger augmented with Twitter-specific features

Results


94
92
90
88
86
84
82
80
78
85.85
83.38
Stanford Tagger
Our tagger, base features

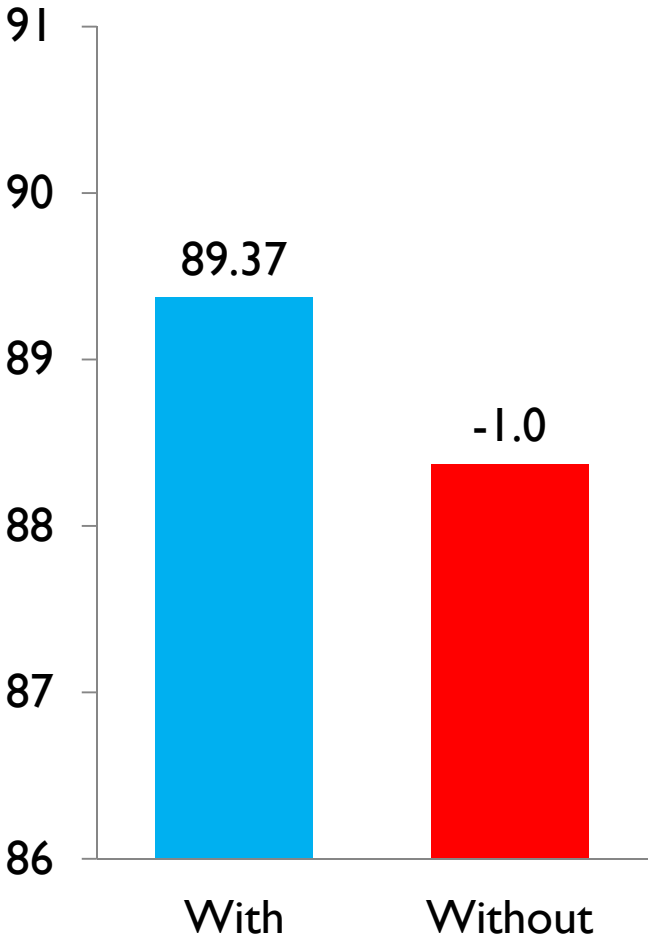

ARK
lti
Carnegie Mellon

Results



ARK lti Carnegie Mellon

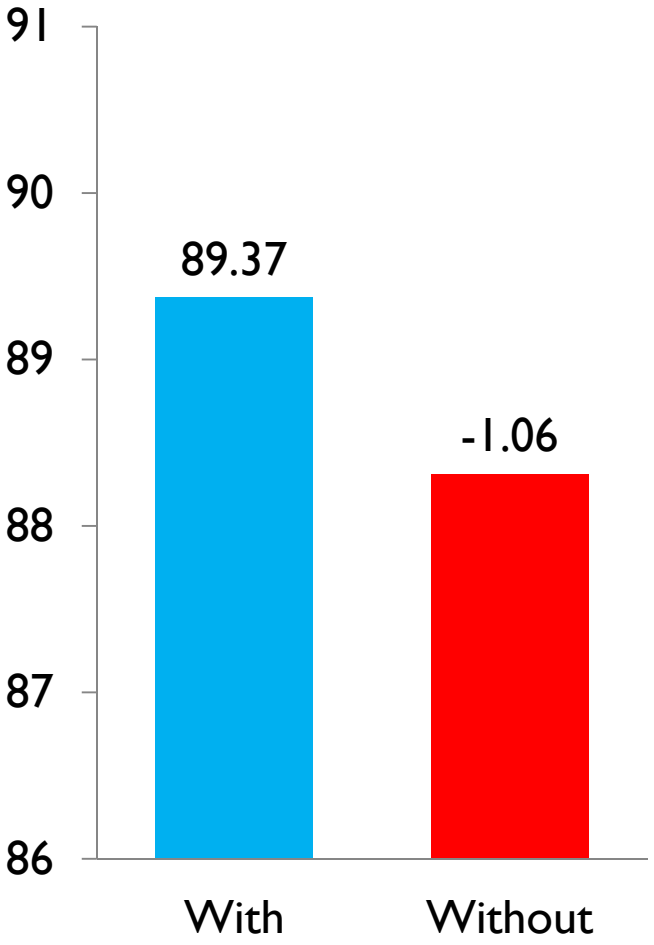
Twitter Orthographic Features



- Regular expressions to detect at-mentions, hashtags, and URLs

ARK lti Carnegie Mellon

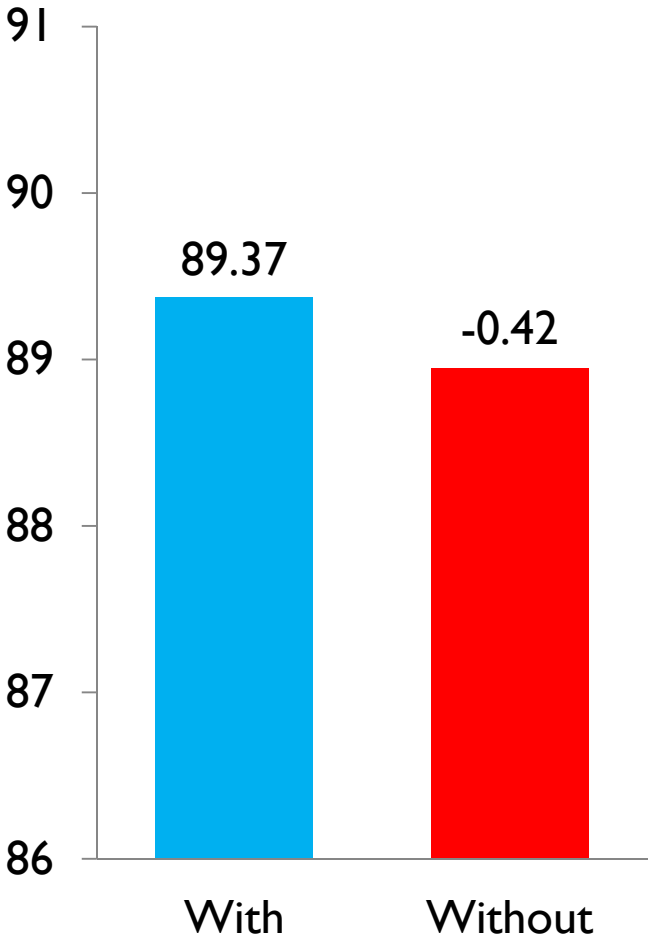
Distributional Similarity Features



- Embeddings in a low-dimensional space based on neighboring words
- Computed using 134k unannotated tweets

ARK lti Carnegie Mellon

Phonetic Normalization Features



- Metaphone algorithm (Philips, 1990) maps tokens to equivalence classes based on phonetics
- Examples:

tomarrow tommorow tomorr
tomorrow tomorrowwww

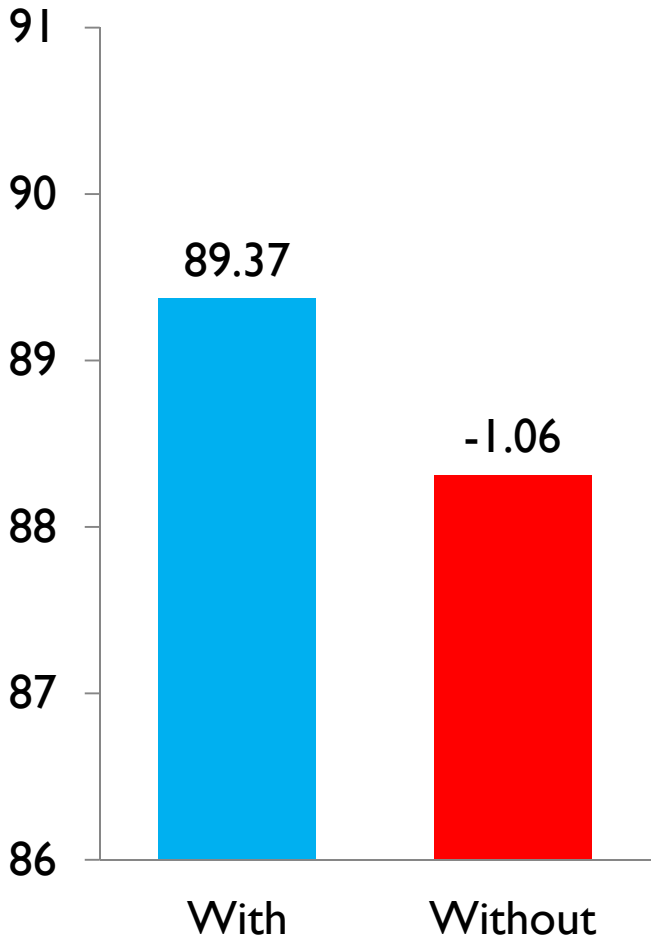
hahaaha hahaha hahahah
hahahahhaa hehehe hehehee

thangs thanks thanksss thanx
things thinks thnx

knew kno know knw n nah naw
new no noo nooooooo now

ARK lti Carnegie Mellon

Tag Dictionary Features



- One feature for each tag a word occurs with in the Penn Treebank, with its frequency rank
- A similar feature for Metaphone classes of Penn Treebank words

ARK lti Carnegie Mellon

Conclusions

- We developed a tag set, annotated data, designed features, and trained models
- Case study in rapidly porting a fundamental NLP task to a social media domain
- Data may be useful for domain adaptation or semi-supervised learning

Thanks!

- Tagger, tokenizer, and annotations are available (50+ downloads already!):

  www.ark.cs.cmu.edu/TweetNLP/