A Dependency Parser for Tweets

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta,
Archna Bhatia, Chris Dyer, and Noah A. Smith

NLP for Social Media

- Boom! Ya ur website suxx bro

  —*@SarahKSilverman*

- michelle obama great. job. and. whit all my. respect she. look. great. congrats. to. her.

  —*@OzzieGuillen*

(Eisenstein, 2013)

NLP for Social Media

(Gimpel et al., 2011; Owoputi et al., 2013)

Boom ! Ya ur website suxx bro (Ritter et al., 2011)

! , ! D N N N

NER

michelle obama great . job . and . whit all my .

^ ^ A , N , & , V X D ,

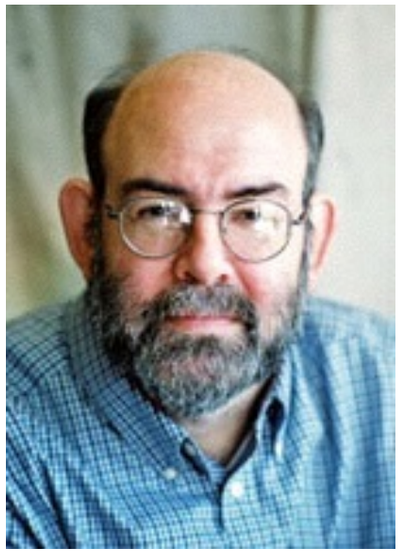
respect she . look . great . congrats . to . her .

V O , V , A , N , P , O ,

The English Web Treebank (Bies et al., 2012) that was sufficient to support a shared task (Petrov and McDonald, 2012) on parsing the web.

NLP for Social Media

Influential members of the House Ways and Means Committee introduced legislation that would restrict how the new savings-and-loan bailout agency can raise capital, creating another potential obstacle to the government's sale of sick thrifts.



— *@MitchellMarcus*

How is Twitter syntax different?

	Twitter-1	Twitter-2	Comments	Forums	Blogs	Wikipedia
Twitter-2	4.0	—	—	—	—	—
Comments	63.7	62.4	—	—	—	—
Forums	91.8	90.6	62.3	—	—	—
Blogs	115.8	119.1	128.4	61.7	—	—
Wikipedia	347.8	360.0	351.4	280.2	157.7	—
BNC	251.8	258.8	245.2	164.1	78.7	92.5

Pairwise corpus similarity ($\times 10^3$) using χ^2 (Baldwin et al., 2013)

How is Twitter syntax different?

	Twitter-1	Twitter-2	Comments	Forums	Blogs	Wikipedia
Twitter-2	4.0	—	—	—	—	—
Comments	63.7	62.4	—	—	—	—
Forums	91.8	90.6	62.3	—	—	—
Blogs	115.8	119.1	128.4	61.7	—	—
Wikipedia	347.8	360.0	351.4	280.2	157.7	—
BNC	251.8	258.8	245.2	164.1	78.7	92.5

Pairwise corpus similarity ($\times 10^3$) using χ^2 (Baldwin et al., 2013)

A Parser?

Frustratingly Hard Domain Adaptation for Dependency Parsing
(Dredze et al., 2011)

#hardtoparse: POS Tagging and Parsing the Twitterverse
(Foster et al., 2011)

Fitting Twitter data to the PTB annotation guideline?

Fitting the parsing task to Twitter data.

Building A Parser — Road Map

- Annotation guidelines
- An annotated corpus
- Parser adaptation
- Useful features

Building A Parser — Road Map

- Annotation guidelines
- An annotated corpus
- Parser adaptation
- Useful features

Not All Tokens Are Syntax

RT @justinbieber : now Hailee get a twitter

Got #college admissions questions ? Ask them tonight during #CampusChat I'm looking forward to advice from @collegevisit http://bit.ly/cchOTk

michelle obama great. job. and. whit all my. respect she. look. great. congrats. to. her.

Token Selection

RT @justinbieber : now Hailee get a twitter

Got #college admissions questions ? Ask them tonight during #CampusChat I'm looking forward to advice from @collegevisit http://bit.ly/cchOTk

michelle obama great. job. and. whit all my. respect she. look. great. congrats. to. her.

Token Selection

RT @justinbieber : now Hailee get a twitter

Got #college admissions questions ? Ask them tonight during #CampusChat I'm looking forward to advice from @collegevisit http://bit.ly/cchOTk

michelle obama great. job. and. whit all my. respect she. look. great. congrats. to. her.

Token Selection

- Pre-processing step
- A first-order sequence model trained using the structured perceptron (Collins, 2002)
- It achieves 97.4% accuracy (ten-fold cross-validated)

Multiword Expressions (MWEs)

Multiword expression should be a single node in the dependency parse from an annotator's perspective.

Annotator's freedom to group words as explicit MWEs:

- proper names: Justin Bieber, World Series
- noncompositional or entrenched nominal compounds: belly button, grilled cheese
- connectives: as well as
- prepositions: out of
- adverbials: so far
- idioms: giving up, make sure

(Baldwin and Kim, 2010; Finkel and Manning, 2009; Constant and Sigogne, 2011; Schneider et al., 2014; Constant et al.,2012; Green et al., 2012; Candito and Constant, 2014; Le Roux et al., 2014)

Multiple Roots

Single root is assumed in PTB — parse one sentence at one time

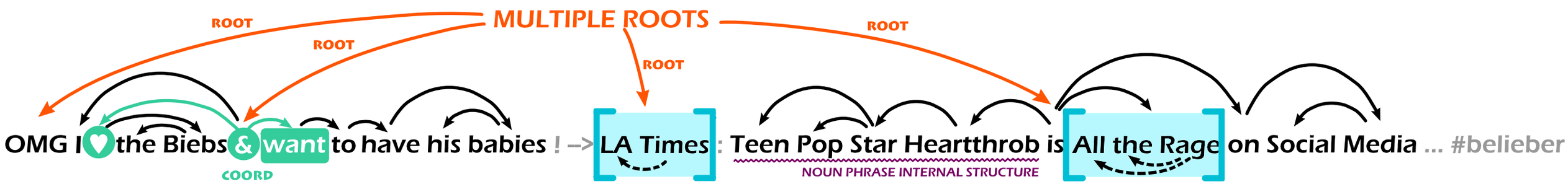
Tweets — often contain multiple sentences or fragments (i.e. "utterances")

We allow multiple attachments to the "wall" symbol (i.e. multi-rooted)

Full Analysis of a Tweet

OMG I ♥ the Biebs & want to have his babies ! –> LA Times : Teen Pop Star Heartthrob is All the Rage on Social Media ... #belieber

Full Analysis of a Tweet

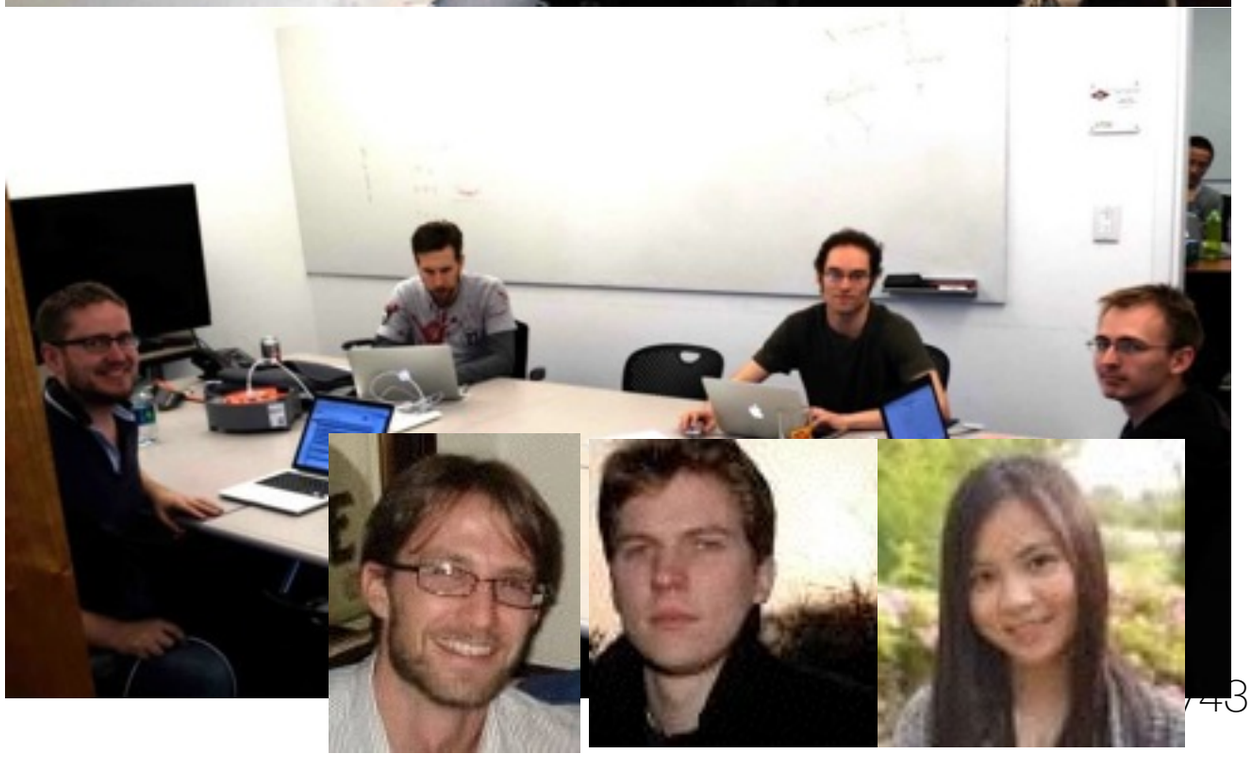


Building A Parser — Road Map

- Annotation guidelines
- An annotated corpus
- Parser adaptation
- Useful features

Building the Tweebank

- Penn Treebank Annotation:
  - take years, involve thousands of person-hours of work by linguists
- Tweebank Annotation:
  - mostly built in a day by two dozen annotators with only cursory training in the annotation scheme



Graph Fragment Language

- A text-based notation that facilitates keyboard entry of parses (Schneider et al., 2013)

bieber is an alien ! :O he went down to earth .

bieber > is** < alien < an
he > [went down]** < to < earth

Number 10 of 10

Sentence:

```
The child ran quickly .
```

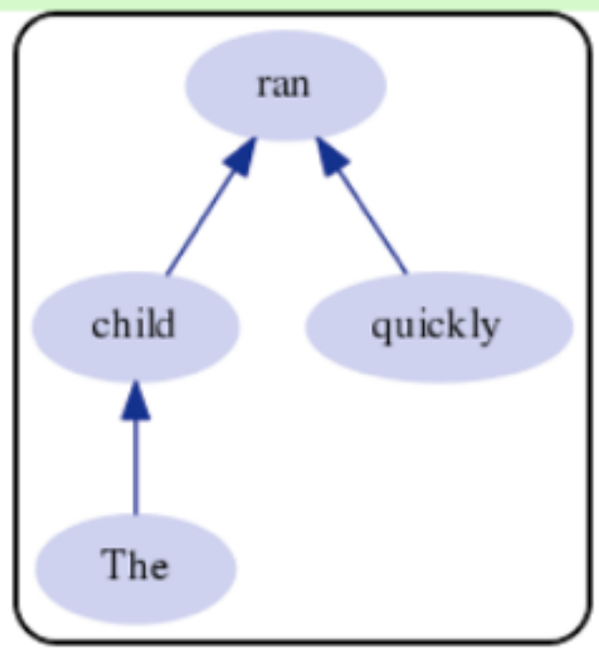
Annotation:

```
1  The > child > ran < quickly
2
3
```

Comments:

< Analyze Submit >

Home



(Mordowanec et al., 2014)

Tweebank

- Tweebank contains 929 tweets (12,318 tokens) with manual dependency parses.
- Tweets drawn from the POS-tagged Twitter corpus of Owoputi et al. (2013), which are tokenized and contain manually annotated POS tags.
- 170 of the tweets were annotated by multiple users — Inter-annotator agreement > 90%

Statistics of our datasets

	Train	Test
tweets	717	201
utterances	1,473	429
tokens	9,310	2,839
selected tokens	7,105	2,158

Building A Parser — Road Map

- Annotation guidelines
- An annotated corpus
- Parser adaptations
- Useful features

Parser Adaptation — Baseline

Out-of-the-Box Parser + Remove all the unselected tokens

OMG I ♥ the Biebs & want to have his babies ! —> LA Times : Teen Pop Star Heartthrob is All the Rage on Social Media … #belieber

Parser Adaptation — Baseline

Out-of-the-Box Parser + Remove all the unselected tokens

OMG I ♥ the Biebs & want to have his babies LA Times
Teen Pop Star Heartthrob is All the Rage on Social Media

lose information (Ma et al. 2014)

“visible” to feature functions, but excluded from the parse tree

Parser Adaptation —TurboParser

A graph-based dependency parser (Martins et al., 2009; Martins et al., 2014)

$$\text{parse}^*(x) = \arg\max_{y \in \mathcal{Y}_x} \mathbf{w}^\top \mathbf{g}(x, y)$$

Decoding using AD^3 (Martins et al., 2014). Many overlapping parts (tree, head-automata etc.) can be handled making use of separate combinatorial algorithms for efficiently handling subsets of constraints.

** AD^3 — Alternating Directions Dual Decomposition

Parser Adaptation —TurboParser

Do NOT change the feature function + Do NOT remove the unselected tokens

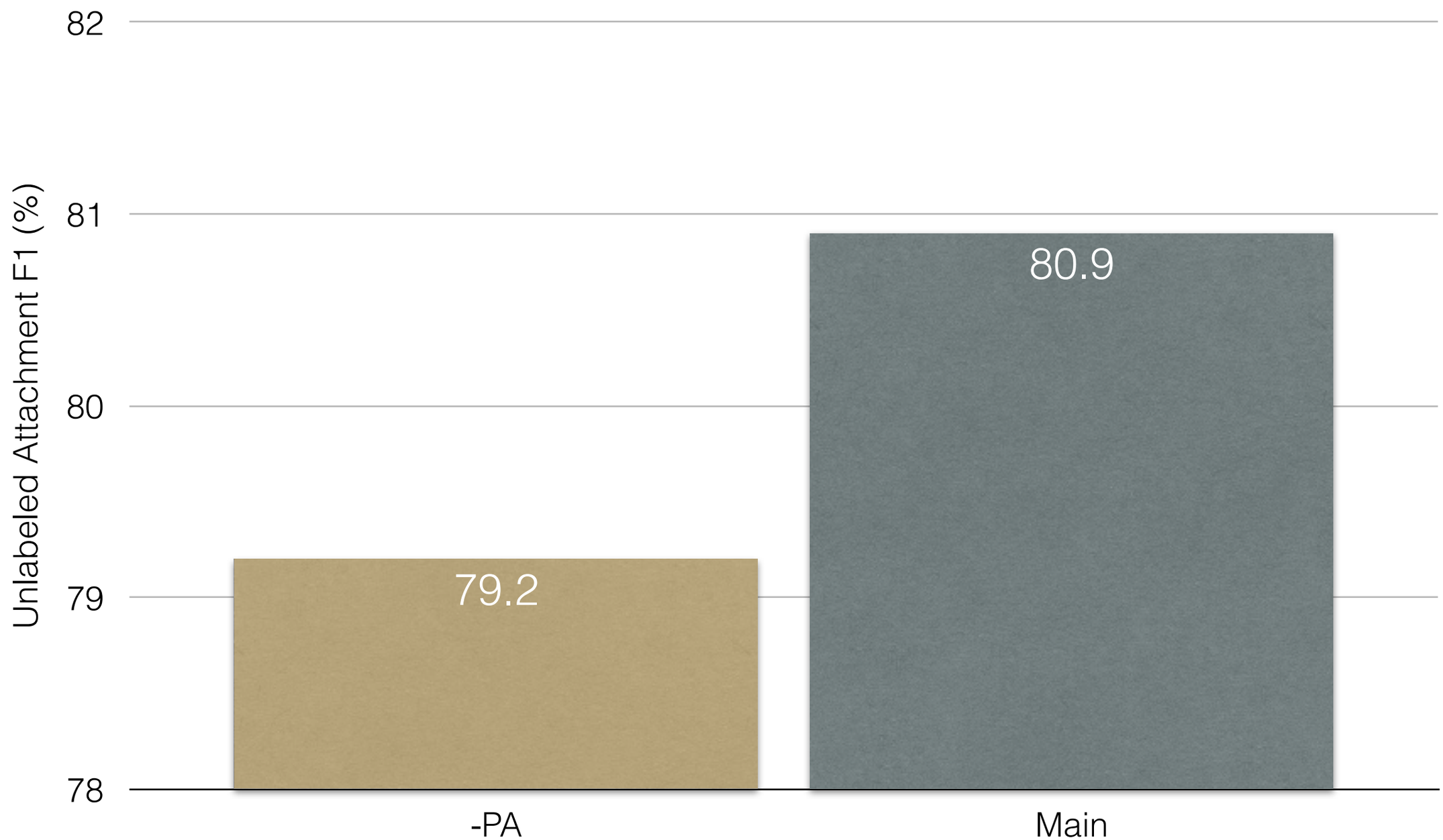
+ Adapt the decoding algorithm to excluded unselected tokens from the tree

Constrain $z_{arc}(i, j) = 0$ whenever x_i or x_j is excluded

For second order factorization (i.e. sibling [p,c,c'] & grandparent [p,c,g]) (McDonald and Satta, 2007; Carreras, 2007)

Grand-sibling head automata (Koo et al., 2010; Martins et al, 2014) for an unselected x_p or x_g, and transitions that consider unselected tokens as children, are eliminated.

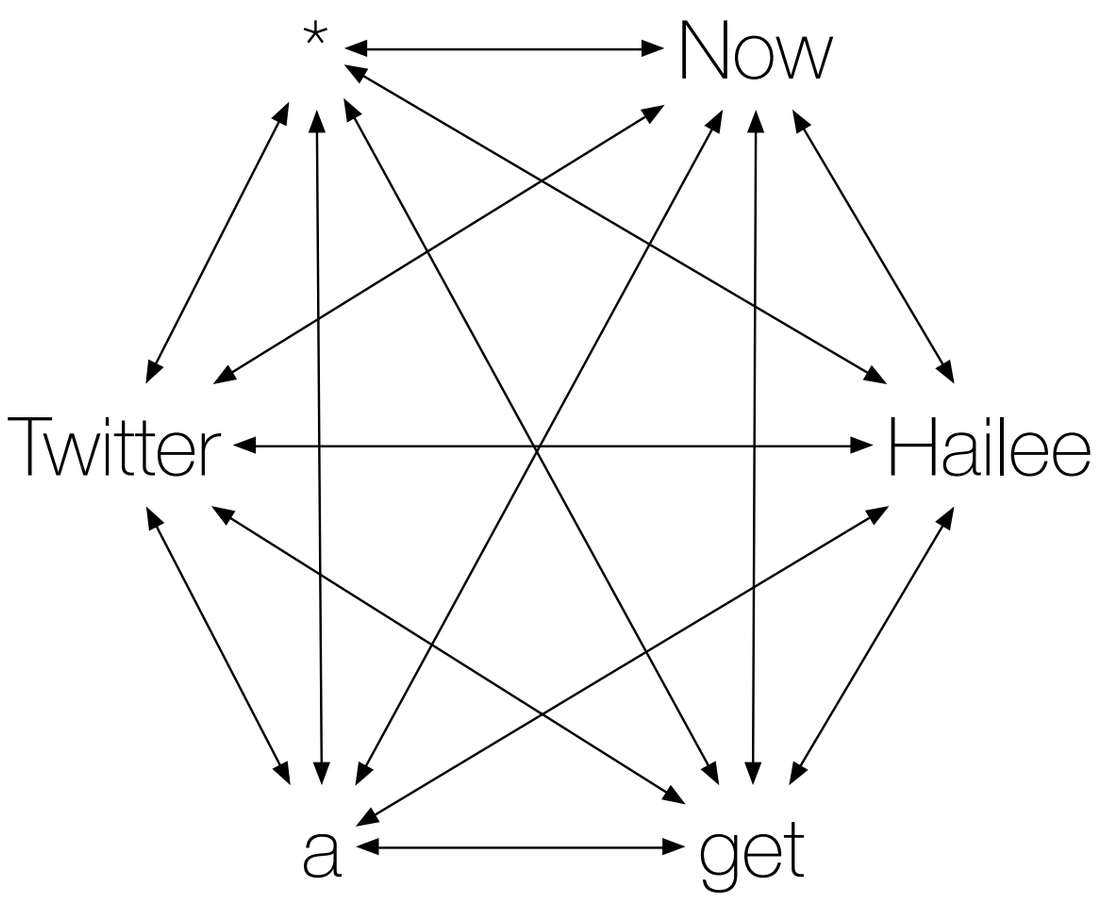
Parser Adaptation

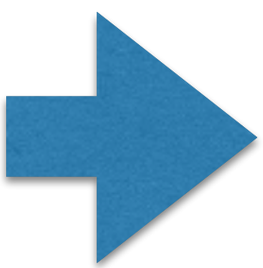


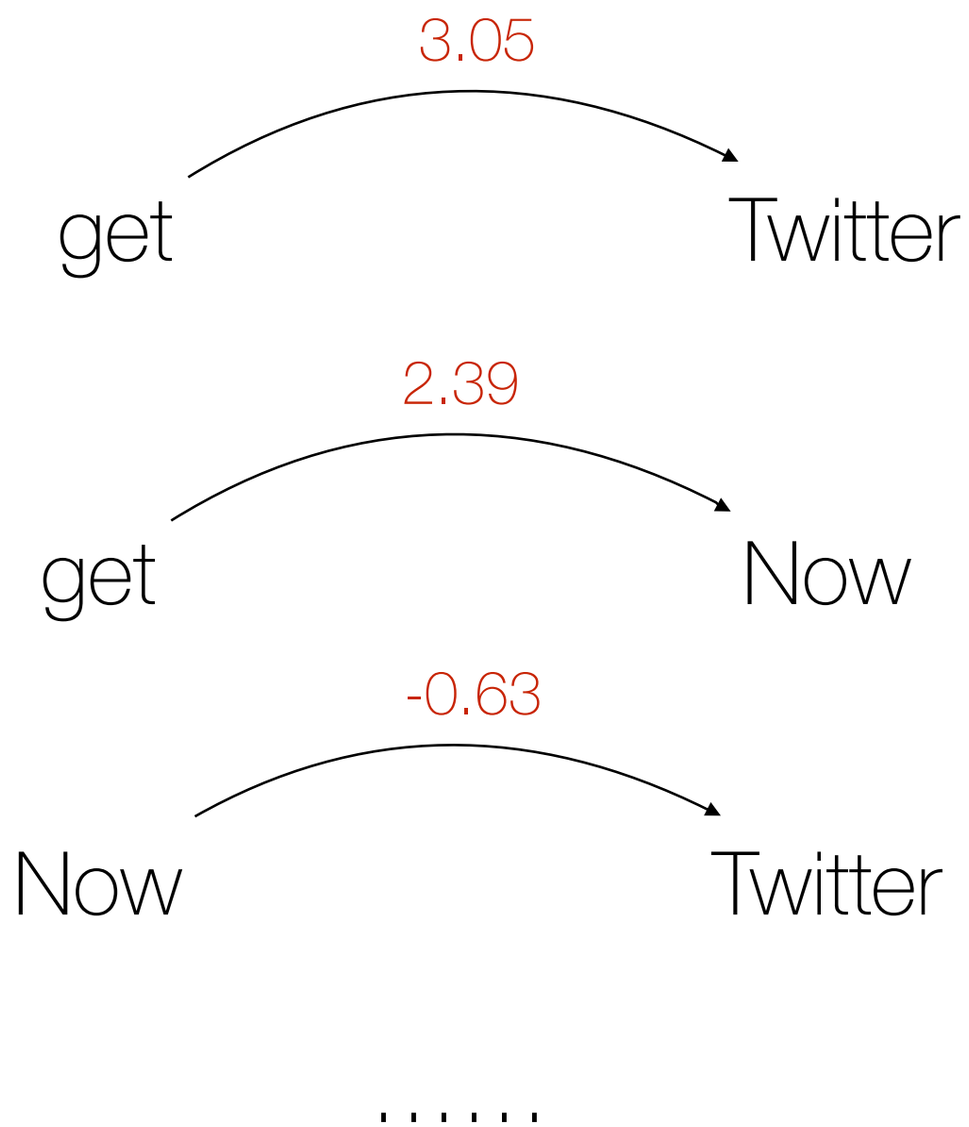
Building A Parser — Road Map

- An annotation guideline
- An annotated corpus
- Parser adaptations
- Useful features

PTB Features

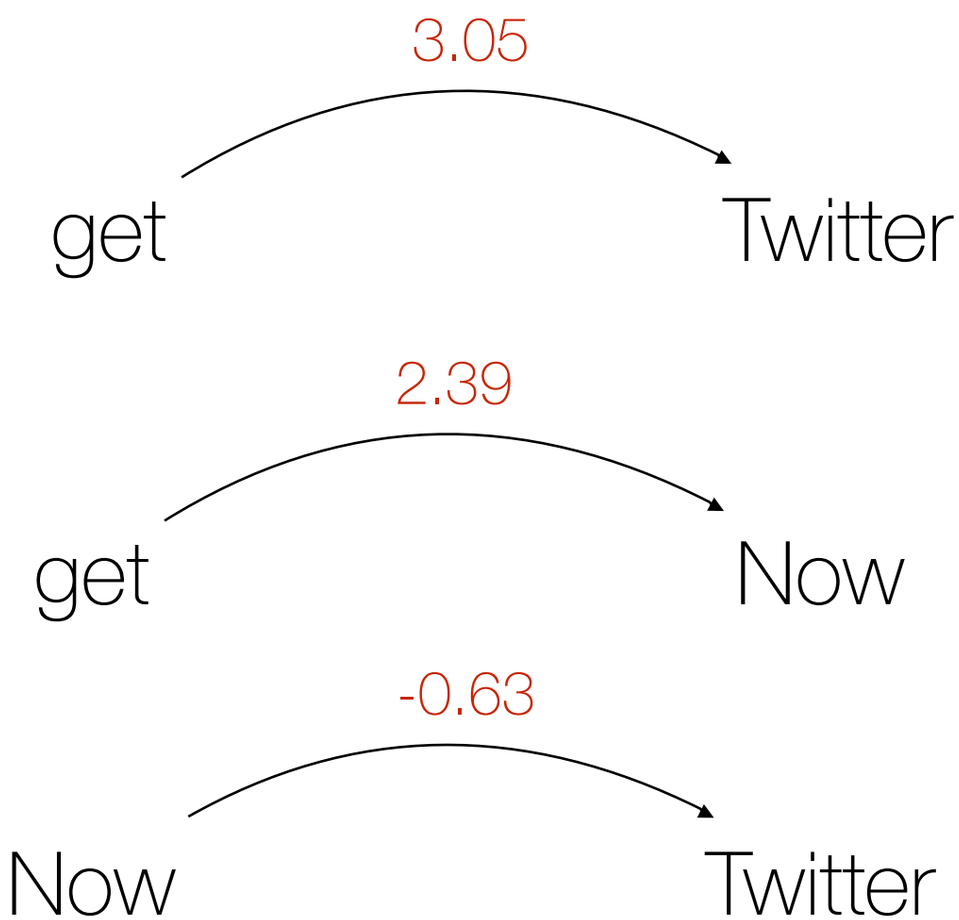






Getting the scores from a first-order model trained on the PTB

PTB Features



w_h = "get" & w_m="Twitter"

p_h = "V" & p_m="^"

direction = "right"

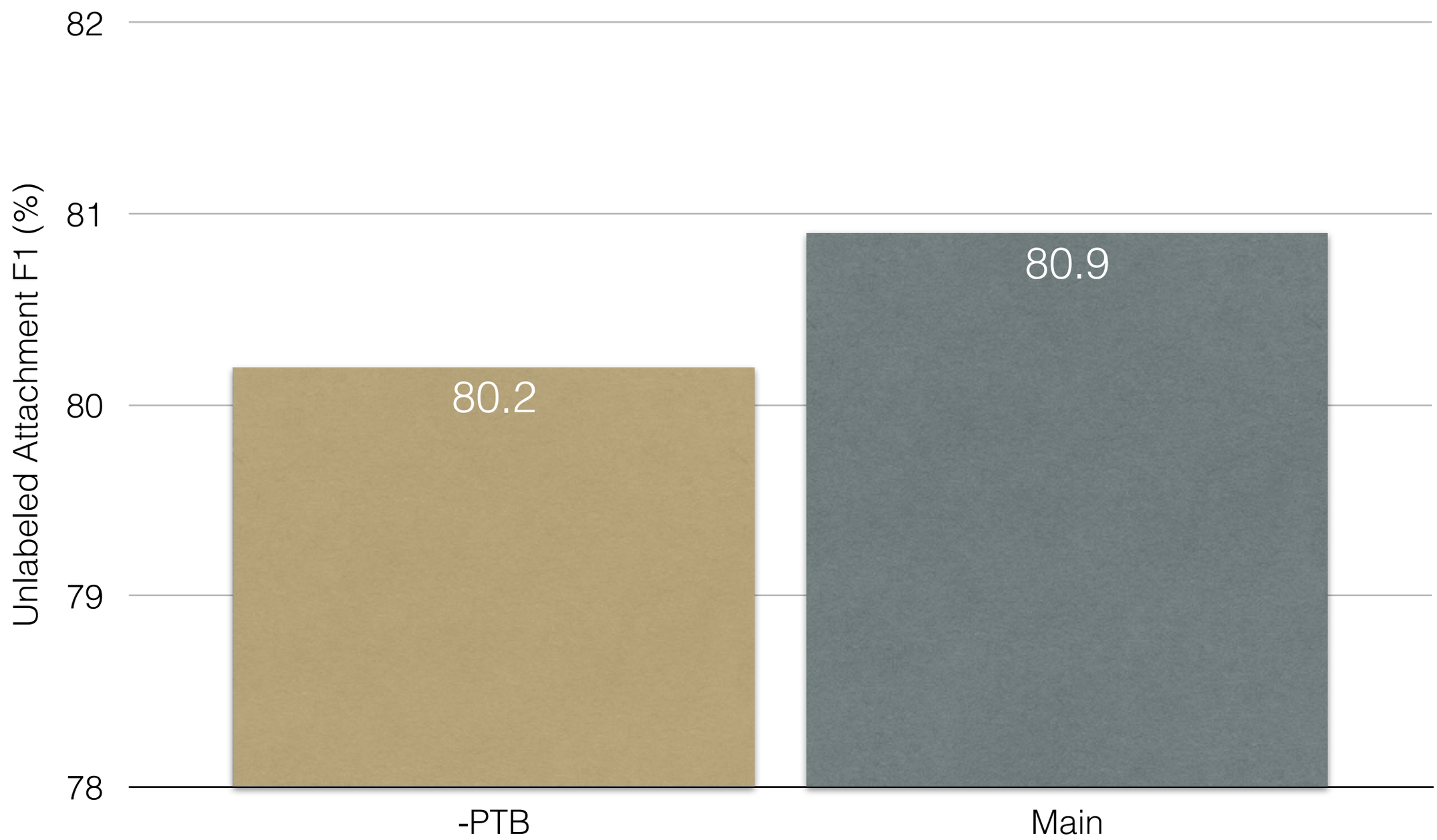
PTB model score = 3.05

.....

.....

* Now Hailee get a Twitter

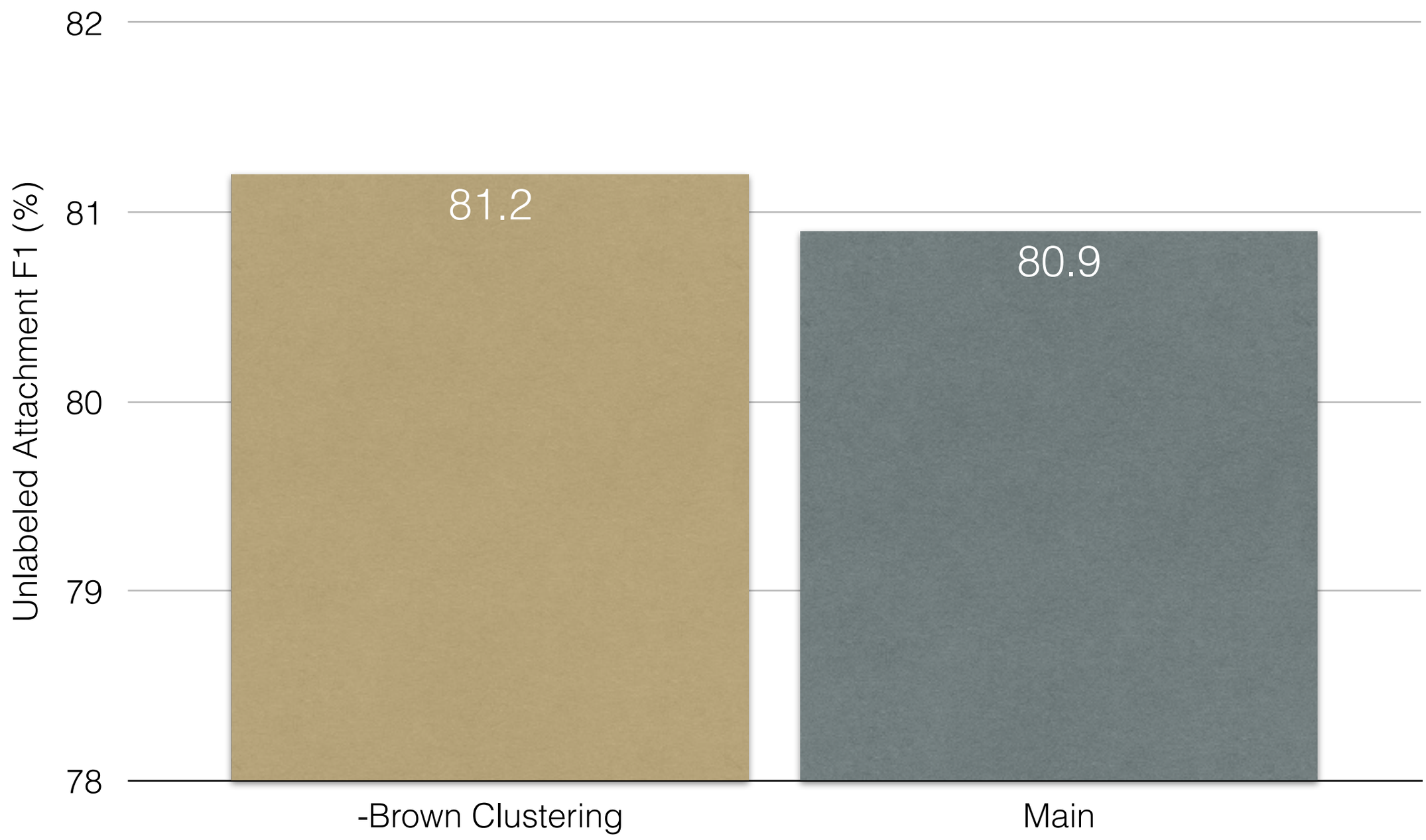
PTB Features



Brown Clustering

- Found very useful in dependency parsing and Twitter POS tagging (Brown et al.,1992; Koo et al., 2008; Owoputi et al. 2013)
- We use clusters trained on 56,345,753 tweets from Owoputi et al. (2012)
- We implement the Brown clustering features following Koo et al. (2008)

Brown Clustering



Building A Parser — Road Map

- Annotation guidelines
- An annotated corpus
- Parser adaptations
- Useful features

Experiments — Setup

	Train	Test-New	Test-Foster
tweets	717	201	< 250
utterances	1,473	429	337
tokens	9,310	2,839	2,841
selected tokens	7,105	2,158	2,366

Experiments

	Unlabeled Attachment F	
	Test-New	**Test-Foster**
Main Parser	80.9	76.1

On par with state-of-the-art reported results for news text in Turkish (77.6%; Koo et al., 2010) and Arabic (81.1%; Martins et al., 2011).

Experiments — Dataset

	Test-New	Test-Foster
sample	50% — random sampled from tweets in 10/27/2010 50% — random sampled from 1/2011 through 6/2012	selected tweets from Bermingham and Smeaton's (2010) corpus, which uses fifty predefined topics
OOV **the Penn Treebank Training Set**	45.2%	21.6% (PTB Test Set —13.2%)

Experiments — Preprocessing

	Test-New
Main Parser	80.9
(++) Gold POS and TS	83.2
(+) Gold POS, automatic TS	82.0
(+) Automatic POS, gold TS	82.0

Experiments — Which Training Set?

	Unlabeled Attachment F	
	mod. POS**	**POS as-is**
Baseline	73.0	73.5
Main Parser	80.9	

** mod. POS — maps at-mentions to pronoun, and hashtags and URLs to noun at test time

Conclusion

- TweeboParser — a dependency parser for English tweets that achieves over 80% unlabeled attachment score on a new, high-quality test set.
- Tweebank — a corpus of 929 tweets (12,318 tokens) with manual dependency parses
- Adaptations to a statistical parsing algorithm
- New approach to exploiting data in a better-resourced domain (PTB)

Thanks!

The dataset and parser are available online!

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