

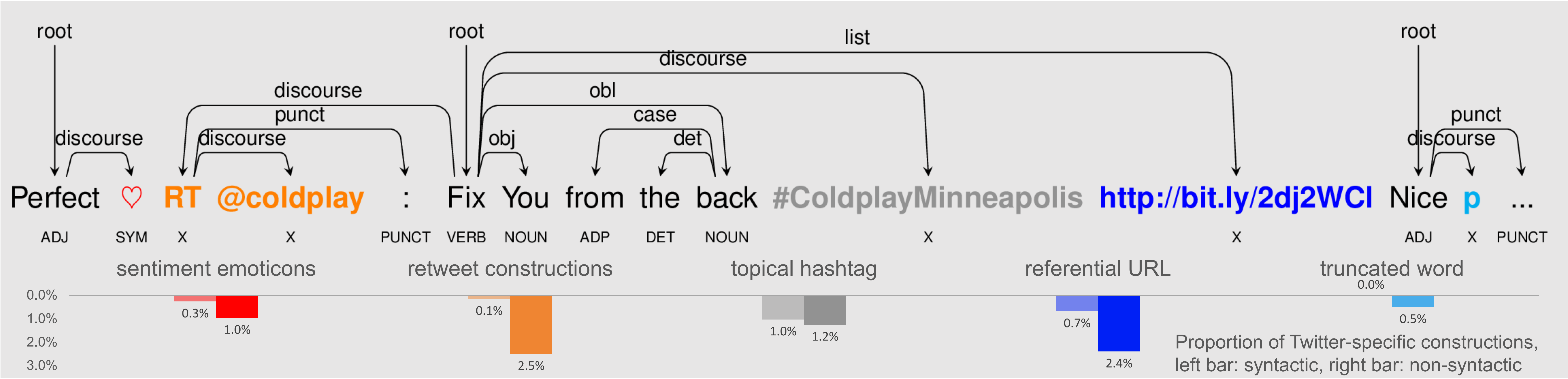
Parsing Tweets **into** Universal Dependencies

Introduction

- Social Media NLP:** domain adaptation and annotated datasets
- Universal Dependencies (UD):** adaptable to different genres and languages
- Our work:** UD v2 on English Social Media
 - Annotation: Tweebank v2 (4x larger than v1)
 - Pipeline: Distillation for fast/accurate parsing

Challenges

- Annotation**
 - Twitter-specific constructions that are not covered by UD guidelines (cf. Sanguinetti et al. 2017 for Italian)
- Pipeline**
 - overcome noise in the annotation
 - accurate parsing without sacrificing speed



Annotation Guidelines

- Tokenization**
Tradeoff between preservation of original tweet content and respecting the UD guidelines.
- Part-of-Speech**
Conform to UD guidelines in most cases. Use syntactic head's POS for abbreviations.
- Dependencies**
Identify non-syntactic tokens (see above Fig.)
 - discourse* for **sentiment emoticon**, **topical hashtag**, and **truncated word**
 - list* for **referential URL** conforming UD
 - Retweet construction** is treated as a whole

Twitter-specific Constructions

	Foster et al. (2010) Stanford Dependencies	Tweebank v1 (Kong et al., 2014) FUDG Dependencies	Tweebank v2 (UD)
• URL	Yes	Yes	Yes
• Ellipsis	Not mentioned	Not mentioned	Yes
• Listing of entities	Not mentioned	Not mentioned	Yes
• Parataxis sentences	Not mentioned	Not mentioned	Yes
• Phrasal abbreviations	Not mentioned	Not mentioned	Our contribution
• Retweet	Yes	Yes	Our contribution
• @-mention (reply)	Yes	Yes	Our contribution
• Hashtag	Yes	Yes	Our contribution
• Truncated words	Not mentioned	Not mentioned	Our contribution

Common in web-text · Common in tweets

Tweebank v2

- Data source:** Tweebank v1 + Feb to Jul 2016 Twitter Stream
- Statistics:**
 - 18 people involved
 - 3,550 annotated tweets
 - 4.5 times larger than v1
 - POS agreement: 96.6
 - Dep. agreement: 88.8 (U) / 84.3 (L)
- Disagreements:**
 - POS for named entities
 - Syntactically ambiguous tweets
 - See our paper for more details

Tokenizer

- Tweet tokenization:** contextual dependent and requires adaption
- Statistical modeling vs rule-based model*
- We propose to use **bi-LSTM** for tokenization and it performs better

System	F1
Stanford CoreNLP	97.3
Ttokenizer	94.6
Ours biLSTM	98.3

POS tagger

- We consider the existing POS taggers
- Rich feature-based (Owoputi et al., 2013) vs neural tagger (Ma and Hovy, 2016)* and careful feature engineering still helps

System	Acc.
Stanford CoreNLP	90.6
Owoputi et al., 2013	94.6
Ma and Hovy, 2016	92.5

Parser

- Annotation:** noisy, complicates the parser training
- Overcome the noise with **ensemble**
- Ensemble is slow.* We do *distillation* and it's fast and accurate

System	LAS	Kt/s
Kong et al. (2014)	76.9	0.3
Dozat et al. (2017)	77.7	1.7
Ballesteros et al. (2015)	75.7	2.3
Ensemble	79.4	0.2
Distillation	77.9	2.3

Pipeline Evaluation

Tokenization: 98.3, POS tagging: 93.3, UD parsing: 74.0

Yijia Liu¹ · Yi Zhu² · Wanxiang Che¹ · Bing Qin¹ · Nathan Schneider³ · Noah A. Smith⁴

¹Harbin Institute of Technology

²University of Cambridge

³Georgetown University

⁴University of Washington