Outline

100% Independent UD Annotation for Tweets

100% Independent UD Annotation for Tweets

MoNoise treebank (van der Goot and van Noord, 2018)

- 632 weets, 10,015 words
- Train on EWT (domain-adaptation)
- 1 annotator
- Paper: effect of normalization

100% Independent UD Annotation for Tweets

Tweebank 2.0 (Liu et al, 2018)

- 3550 tweets, 111,214 words
- Train on tweets (+EWT)
- 18 annotators
- Paper: Build ensemble, and make this more efficient

100% Independent UD Annotation for Tweets

Both contain data from Owoputi et al. (2013)!

100% Independent UD Annotation for Tweets

Did this happen before?

100% Independent UD Annotation for Tweets

Did this happen before?

– Bamman, David, Francesco Mambrini & Gregory Crane (2009), An ownership model of annotation: The Ancient Greek dependency treebank. In: *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8)*. Groningen, 5–15. Available at: http://www.perseus.tufts.edu/~ababeu/tlt8.pdf.
– Berzak, Yevgeni, Yan Huang, Andrei Barbu, Anna Korhonen & Boris Katz (2016), Anchoring and Agreement in Syntactic Annotations. In: *Proceedings of EMNLP 2016*. Austin, TX, 2215–2224.
– Berzak, Yevgeni, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza & Boris Katz (2016), Universal Dependencies for Learner English. In: *Proceedings of ACL 2016*. Berlin, Germany, 737–746. Available at: http://www.aclweb.org/anthology/P16-1070.
– Liu, Yijia, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider & Noah A. Smith (2018), Parsing Tweets into Universal Dependencies. In: *Proceedings of NAACL 2018*. New Orleans, LA, 965–975. Available at: http://aclweb.org/anthology/N18-1088.
– Nguyen, Kiem-Hieu (2018), BKTreebank: Building a Vietnamese Dependency Treebank. In: *Proceedings of LREC 2018*. Miyazaki, Japan, 2164–2168. Available at: http://www.lrec-conf.org/proceedings/lrec2018/pdf/69.pdf.
– Seddah, Djamé, Eric De La Clergerie, Benoît Sagot, Héctor Martínez Alonso & Marie Candito (2018), Cheating a Parser to Death: Data-driven Cross-Treebank Annotation Transfer. In: *Proceedings of LREC 2018*. Miyazaki, Japan, 4535–4539. Available at: http://www.lrec-conf.org/proceedings/lrec2018/pdf/1101.pdf.
– Seyoum, Binyam Ephrem, Yusuke Miyao & Baye Yimam Mekonnen (2018), Universal Dependencies for Amharic. In: *Proceedings of LREC 2018*. Miyazaki, Japan, 2216–2222. Available at: http://www.lrec-conf.org/proceedings/lrec2018/pdf/565.pdf.
– Skjærholt, Arne (2014), A Chance-corrected Measure of Inter-annotator Agreement for Syntax. In: *Proceedings of ACL 2014*. Baltimore, MD, 934–944. Available at: http://www.aclweb.org/anthology/P14-1088.

Thanks to Amir Zeldes and the corpora-list

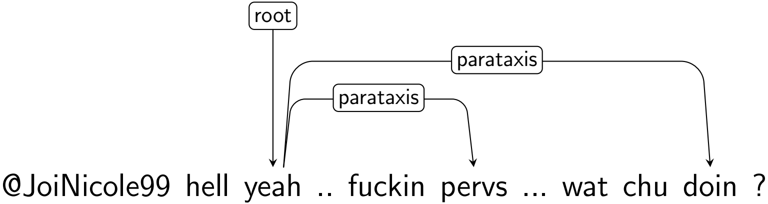
100% Independent UD Annotation for Tweets

For tweets (inter-annotator agreement in 1 paper):

POS	96.6%
unlabeled dependencies	88.8%
labeled dependencies	84.3%

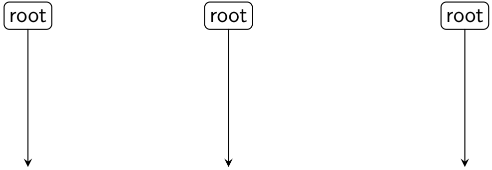
100% Independent UD Annotation for Tweets

Different guidelines (me):



100% Independent UD Annotation for Tweets

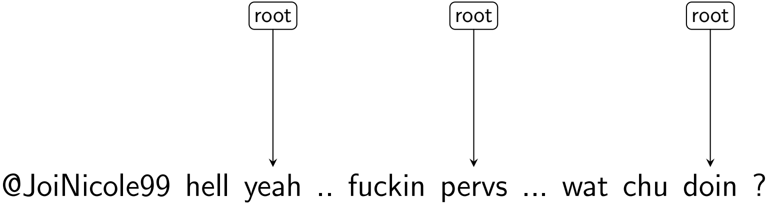
Different guidelines (them):



@JoiNicole99 hell yeah .. fuckin pervs ... wat chu doin ?

100% Independent UD Annotation for Tweets

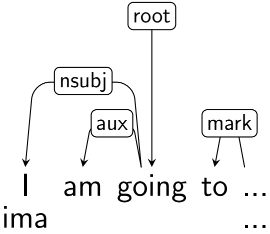
Different guidelines (them):



Easy to converge (rule-based)

100% Independent UD Annotation for Tweets

Different guidelines (me):



100% Independent UD Annotation for Tweets

Different guidelines (them):

root

ima
ima

100% Independent UD Annotation for Tweets

Different guidelines (them):

root

ima
ima

Bit harder to converge (not done yet)

100% Independent UD Annotation for Tweets

Other things:

- I leave phrasal abbreviations as is (they acronyms)

100% Independent UD Annotation for Tweets

Other things:

- I leave phrasal abbreviations as is (they acronyms)
- emoticon & emoji: `SYMB`, `appos`

100% Independent UD Annotation for Tweets

Other things:

- I leave phrasal abbreviations as is (they acronyms)
- emoticon & emoji: `SYMB, appos`
- urls: `X, appos` versus `X, list`

100% Independent UD Annotation for Tweets

Other things:

- I leave phrasal abbreviations as is (they acronyms)
- emoticon & emoji: `SYMB, appos`
- urls: `X, appos` versus `X, list`
- username mentions: `PROPN, vocative`

100% Independent UD Annotation for Tweets

Other things:

- I leave phrasal abbreviations as is (they acronyms)
- emoticon & emoji: `SYMB, appos`
- urls: `X, appos` versus `X, list`
- username mentions: `PROPN, vocative`
- RT: `X, discourse`

100% Independent UD Annotation for Tweets

Other things:

- I leave phrasal abbreviations as is (they acronyms)
- emoticon & emoji: `SYMB, appos`
- urls: `X, appos` versus `X, list`
- username mentions: `PROPN, vocative`
- RT: `X, discourse`
- Annotate accordingly when above things are used in syntactic context

100% Independent UD Annotation for Tweets

first try:

- ID match
- 126 found

100% Independent UD Annotation for Tweets

second try:

- character edit distance
- Ignore whitespace, username and allow for 20% variation
- 142 found

100% Independent UD Annotation for Tweets

why 20% variation?

```
rt@userwho'seversmokedbeforetheytookatestatschool?/*raise
rt@userwho'seversmokedbeforetheytookatestatschool?/*raise
```

```
imhome:)
imhome:-)
```

```
@user601blueroommay19thfemsfreeanddrinkfreetil11:30$5alln
@iamyungsmilezblueroommay19thfemsfreeanddrinkfreetil11:30
```

100% Independent UD Annotation for Tweets

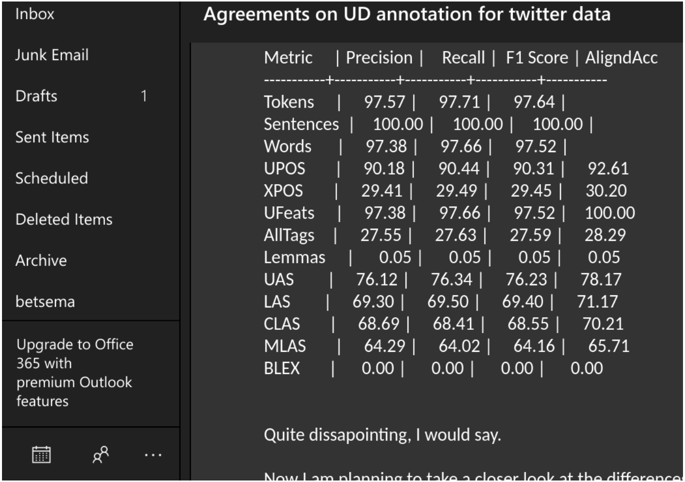
```
==> outputRob <==
# sent_id = owoputi.406.28857809439
# text = Yall sholl is quiet!! SPEAK UP lol RT @MzCHinezeEyez: @McQSpeaks We Still Here!!....lol
1       Y       _       PRON    _       _       5       nsubj   _       Norm=you|SpaceAfter=No
2       all     _       DET     _       _       1       det     _       Norm=all
3       sholl   _       AUX     _       _       5       aux     _       Norm=should
4       is      _       AUX     _       _       5       cop     _       Norm=be
5       quiet   _       ADJ     _       _       0       root    _       Norm=quiet|SpaceAfter=No
6       !!      _       PUNCT   _       _       5       punct   _       Norm=!!
7       SPEAK   _       VERB    _       _       5       parataxis       _       Norm=SPEAK
8       UP      _       ADP     _       _       7       compound:prt    _       Norm=UP

==> outputTweebank.fixed.fixed <==
# tweet_id = oct27.28857809439
# text = Yall sholl is quiet!! SPEAK UP lol RT @MzCHinezeEyez: @McQSpeaks We Still Here!!....lol
1       Yall    yall    PRON    O       _       4       nsubj   NormType=contraction|NormWord=you_all
        _
2       sholl   sholl   ADV     R       _       4       advmod  _       _
3       is      be      AUX     V       _       4       cop     _       _
4       quiet   quiet   ADJ     A       _       0       root    _       SpaceAfter=No
5       !!      !       PUNCT   ,       _       4       punct   _       _
6       SPEAK   speak   VERB    V       _       4       parataxis       _       _
7       UP      up      ADP     T       _       6       compound:prt    _       _
8       lol     lol     INTJ    !       _       6       discourse       _       _
```

100% Independent UD Annotation for Tweets

First test, `conll18_ud_eval.py`:

Metric	Precision	Recall	F1 Score	AligndAc
Tokens	97.57	97.71	97.64	
Sentences	100.00	100.00	100.00	
Words	97.38	97.66	97.52	
UPOS	90.18	90.44	90.31	92.6
UAS	76.12	76.34	76.23	78.1
LAS	69.30	69.50	69.40	71.1
CLAS	68.69	68.41	68.55	70.2

100% Independent UD Annotation for Tweets



Agreements on UD annotation for twitter data

Metric	Precision	Recall	F1 Score	AligndAcc
Tokens	97.57	97.71	97.64	
Sentences	100.00	100.00	100.00	
Words	97.38	97.66	97.52	
UPOS	90.18	90.44	90.31	92.61
XPOS	29.41	29.49	29.45	30.20
UFeats	97.38	97.66	97.52	100.00
AllTags	27.55	27.63	27.59	28.29
Lemmas	0.05	0.05	0.05	0.05
UAS	76.12	76.34	76.23	78.17
LAS	69.30	69.50	69.40	71.17
CLAS	68.69	68.41	68.55	70.21
MLAS	64.29	64.02	64.16	65.71
BLEX	0.00	0.00	0.00	0.00

Quite dissapointing, I would say.

Now I am planning to take a closer look at the difference

100% Independent UD Annotation for Tweets

Answer:

`Thanks for the experiments. The number seemed OK to me ..`

100% Independent UD Annotation for Tweets

Answer:

`Thanks for the experiments. The number seemed OK to me ..`

Conclusion: we do not agree...

100% Independent UD Annotation for Tweets

`eval.pl` by Yuval Krymolowski

100% Independent UD Annotation for Tweets

```
p270396@vesta1:udNew$ perl eval.pl -g outputTweebank.fixe
Word/pos mismatch, line 1:
 gold: # tweet_id = oct27.28857809439
 sys : # sent_id = owoputi.406.28857809439
Word/pos mismatch, line 3:
 gold: 1     Yall    yall    PRON    O    _    4    nsubj    N
 sys : 1     Y    _    PRON    _    _    5    nsubj    _    Norm=
Word/pos mismatch, line 4:
 gold: 2     sholl    sholl    ADV R    _    4    advmod    _    _
 sys : 2     all _    DET _    _    1    det _    Norm=all
Word/pos mismatch, line 5:
 gold: 3     is   be   AUX V    _    4    cop _    _
 sys : 3     sholl    _    AUX _    _    5    aux _    Norm=shou
Word/pos mismatch, line 6:
 gold: 4     quiet    quiet    ADJ A    _    0    root    _    S
 sys : 4     is  _    AUX _    _    5    cop _    Norm=be
```

100% Independent UD Annotation for Tweets

For now:

- Filtered, only tweets with same tokenization
- 114 tweets left

100% Independent UD Annotation for Tweets

5 focus words where most of the errors occur:

	any	head	dep	both
lol / _	4	4	1	1
it / _	4	1	4	1
RT / _	4	4	0	0
that / _	4	1	3	0
me / _	4	3	4	3

100% Independent UD Annotation for Tweets

1. head one word after the correct head (after the focus word), correct dependency : 11 times
2. dependency "root" instead of "parataxis" : 11 times
3. head one word before the correct head (after the focus word), correct dependency : 11 times
4. dependency "aux" instead of "cop" : 5 times
5. dependency "discourse" instead of "parataxis" : 5 times
6. dependency "advcl" instead of "parataxis" : 5 times

100% Independent UD Annotation for Tweets

Incoming labels I used where they did not:

parataxis	46
discourse	24
root	22
obj	18
nsubj	16
xcomp	12
advcl	9
compound	9
obl	9
advmod	8

100% Independent UD Annotation for Tweets

Incoming labels they used where I did not:

vocative	23
discourse	22
root	22
advcl	20
advmod	17
nsubj	17
obl	15
compound	13
ccomp	12
aux	9

100% Independent UD Annotation for Tweets

(Preliminary) conclusion: Most mistakes made for:

- vocative
- discourse
- root
- parataxis

100% Independent UD Annotation for Tweets

(Preliminary) conclusion: Most mistakes made for:

- vocative
- discourse
- root
- parataxis
- LAS might sketch a too negative image

100% Independent UD Annotation for Tweets

Next:

- Get parser performance for both (train on EWT)
- MaltEval
- More manual analysis
- Merge styles
- ...

Outline

1. 100% Independent UD Annotation for Tweets
2. Effect of Normalization Categories on Parsing
3. Master theses

Effect of Normalization Categories on Parsing

Effect of Normalization Categories on Parsing

Thesis:

- Evaluating normalization per category
- Effect of normalization on parsing

Effect of Normalization Categories on Parsing

Thesis:

- Evaluating normalization per category
- Effect of normalization on parsing
- Logical follow up: evaluating effect normalization categories for parsing

Effect of Normalization Categories on Parsing

Thesis:

- Evaluating Normalization per category
- Effect of normalization on parsing
- Logical follow up: Evaluating effect normalization categories for parsing

Effect of Normalization Categories on Parsing

Tyler Baldwin, Yunyao Li. 2015. An In-depth Analysis of the Effect of Text Normalization in Social Media. In *Proceedings of NAACL*.

Effect of Normalization Categories on Parsing

So why do it again?

Effect of Normalization Categories on Parsing

Their taxonomy:

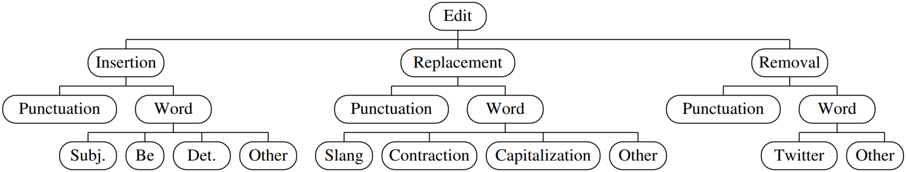


Figure 1: Taxonomy of normalization edits

Effect of Normalization Categories on Parsing

Their taxonomy:

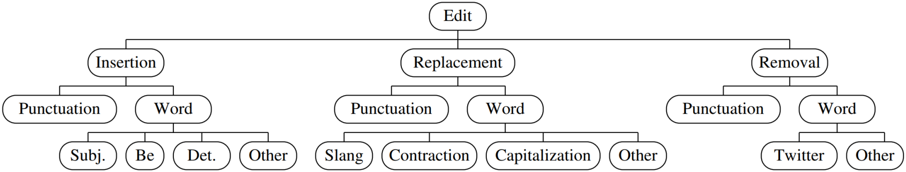


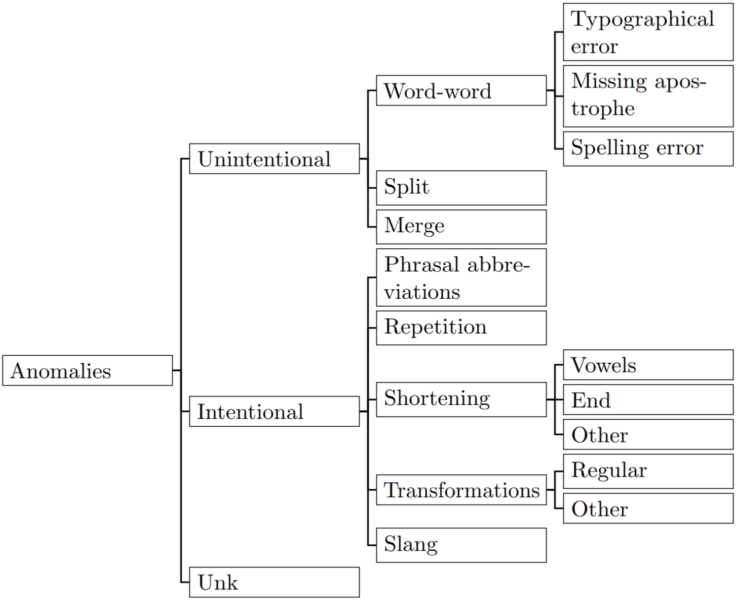
Figure 1: Taxonomy of normalization edits

For automated normalization, the scope is often different!

Effect of Normalization Categories on Parsing

Rob van der Goot, Rik van Noord and Gertjan van Noord. 2018. A Taxonomy for In-depth Evaluation of Normalization for User Generated Content. In *Proceedings of LREC*

Effect of Normalization Categories on Parsing



Effect of Normalization Categories on Parsing

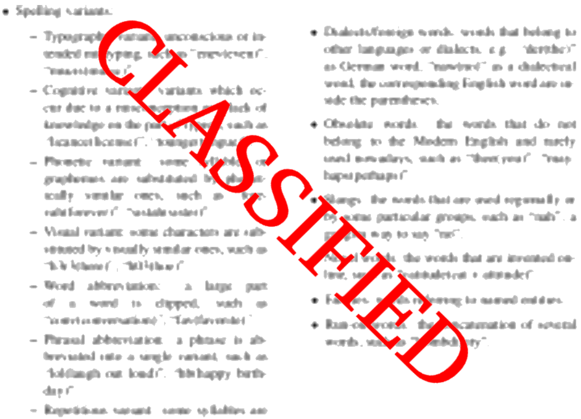
But how do you classify 'lolllll'?

Effect of Normalization Categories on Parsing

$$\kappa = 0.807$$

Effect of Normalization Categories on Parsing

EMNLP 2018 submission:



Effect of Normalization Categories on Parsing

But I annotated lexnorm2015 with categories, and Owoputi and Lexnorm with UD...

Effect of Normalization Categories on Parsing

But I annotated lexnorm2015 with categories, and Owoputi and Lexnorm with UD...
So I added category annotation to Owoputi treebank

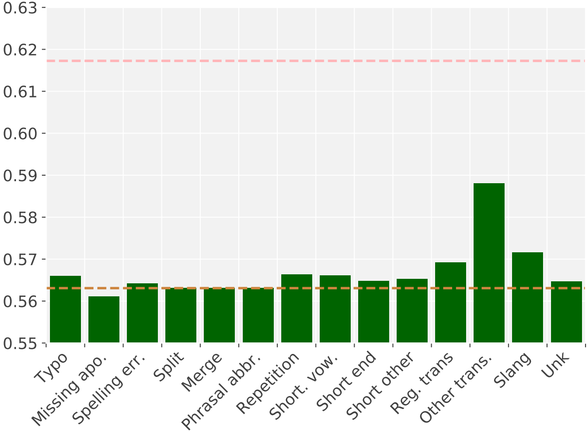
Effect of Normalization Categories on Parsing

Setup:

- UUParser 2.0
- Use gold normalization only for specific categories:
  - in isolation
  - ablation

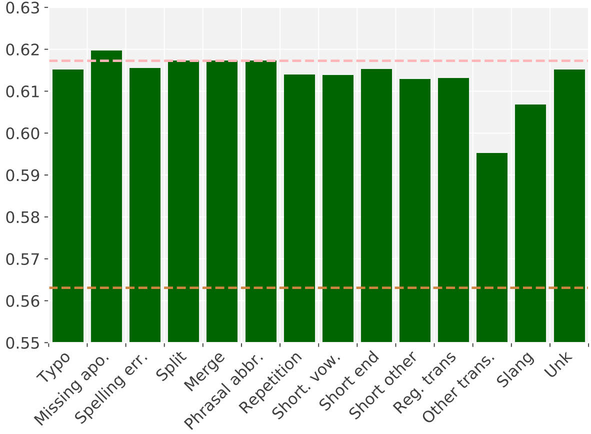
Effect of Normalization Categories on Parsing

Results: (isolation)



Effect of Normalization Categories on Parsing

Results: (ablation)



Effect of Normalization Categories on Parsing

Next:

- Use automatic normalization
- Test for other tasks?
- ...

Outline

Master theses

- Distant supervision for normalization (* 2)
- Automatic prediction of taxonomy categories
- The effect of lexical normalization on POS tagging for Dutch

Master theses

Distant supervision for normalization:

- Ian Matroos: 14:45
- Kelly Dekker: +Human evaluation

Master theses

Automatic prediction of taxonomy categories (Wessel Reijngoud):

- in corpus
- cross-corpus
- cross-language

Master theses

Why?

- Compare corpora (languages?)
- Evaluate normalization models in more detail for multiple languages

Master theses

The effect of lexical normalization on POS tagging for Dutch (youri schuur):

- van der Goot et al. (2017). English. BiLSTM with pre-trained embeds: small gain
- Schulz et al. (2016). Dutch. Treetagger: huge gain

Master theses

The effect of lexical normalization on POS tagging for Dutch (youri schuur):

- van der Goot et al. (2017). English. BiLSTM with pre-trained embeds: small gain
- Schulz et al. (2016). Dutch. Treetagger: huge gain
- Is this an effect of language? or setup?

Master theses

Additional benefits:

- First work to annotate tokenization and normalization as separate layer
- Correct capitalization
- Publicly available evaluation set for Dutch UGC normalization and POS tagging
- Improve MoNoise for Dutch
- Can be used for all the other master theses

Master theses

Thanks, Questions? (you may leave the easy ones for tomorrow)