

Lets agree to disagree about syntactic annotations

Rob van der Goot

40 years GroNLP

21 oktober 2026

Background

- I'm Rob
- Studied Informatiekunde from 2009-2014 in Groningen
- PhD on "Normalization and Parsing for Uncertain Input"
- Supervised by Gertjan



Background

- I'm Rob
- Studied Informatiekunde from 2009-2014 in Groningen
- PhD on "Normalization and Parsing for Uncertain Input"
- Supervised by Gertjan
- Since then: IT University of Copenhagen

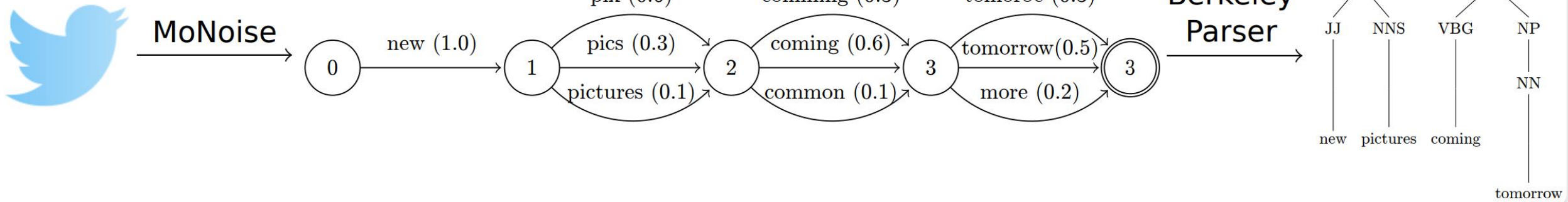


Story

- Present my slowest, most interesting (?!) project
- Not published
- Based on actual events

Story

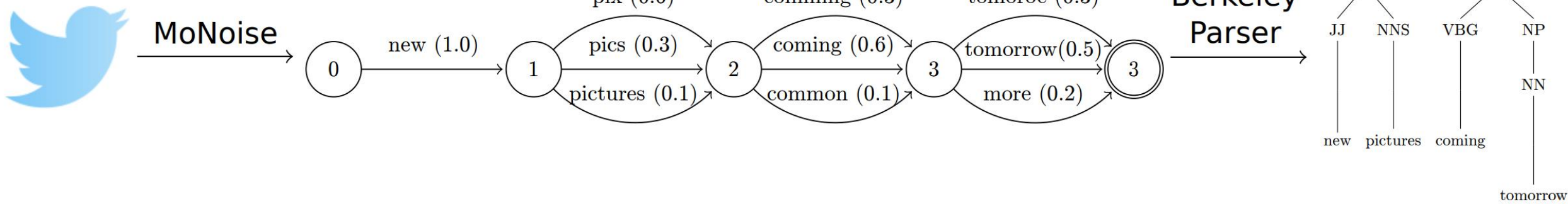
- Parsing Uncertain Input
- 2018:
 - SOTA normalization
 - Top-n constituency parser



Story



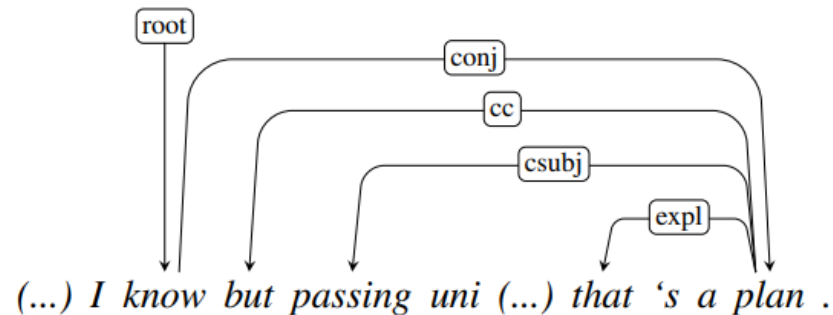
- Parsing Uncertain Input
- 2018:
 - SOTA normalization
 - Top-n constituency parser



- Dependency Parsing became more popular

So: "Modeling Input Uncertainty in Neural Network Dependency Parsing"

- Created a small UD treebank based on social media data. (and build a parser)



- Annotated by me

- Submitted to EMNLP 2018.
- Scores: 2.5 3 3 (out of 5)

 You forwarded this message on Fri 8/10/2018 20:58

Dear Rob van der Goot:

On behalf of the EMNLP 2018 Program Committee, we are delighted to inform you that the following submission has been accepted to appear at the conference:

Modeling Input Uncertainty in Neural Network Dependency Parsing

E-mail

Wow, really! Congratulations!

So... not getting married after all?

GJ

E-mail

Wow, really! Congratulations!

So... not getting married after all?

GJ

Yes, I can go. Then I'll actually be doing something for my co-authorship for once...

GJ

One year later

- Tweebank 2.0 was released
- UD annotations for social media data
 - Both contain data from Owoputi et al. (2013)!
 - First totally independent annotations for syntax (?)

One year later

- Tweebank 2.0 was released
- UD annotations for social media data
 - Both contain data from Owoputi et al. (2013)!
 - First totally independent annotations for syntax (?)
- Gertjan recommended me to move to another university "preferably in another country"

One year later

- Tweebank 2.0 was released
- UD annotations for social media data
 - Both contain data from Owoputi et al. (2013)!
 - First totally independent annotations for syntax (?)
- Gertjan recommended me to move to another university "preferably in another country"
- Moved to Denmark to work with Barbara

Cross-annotator F1 score

Task	F1
Words	97.38
UPOS	90.31
Dependency	69.40

Cross-annotator F1 score

Task	F1
Words	97.38
UPOS	90.31
Dependency	69.40

SOTA (news data)

Task	F1
Words	~99.9?
UPOS	~97
Dependency	~95

Hello, I am Rob van der Goot (we shortly met at ACL). I have finally found time to look at the differences between our corpora. I created some character edit distance based method to align the sentences, and found 142 overlapping trees (2234 and 2241 nodes/words). The results of the default conll script (by using your treebank as gold, and mine as prediction) are as follows:

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.

Reply:

Hi Rob,

Thanks for the experiments. The number seemed OK to me

Reply:

Hi Rob,

Thanks for the experiments. The number seemed OK to me

Lets agree to disagree

Interesting!, but what to do?

- Not sure... was also moving

Interesting!, but what to do?

- Not sure...
- Until I met an excellent master student in Copenhagen (2023).
- We published some papers, and she wanted a challenging, linguistically focused project for thesis.

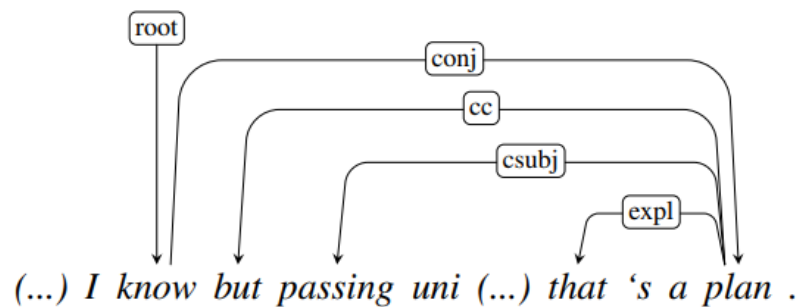
Comparing different annotations (LAS)

	TB	2018
TB	100.0	
2018	74.0	100.0

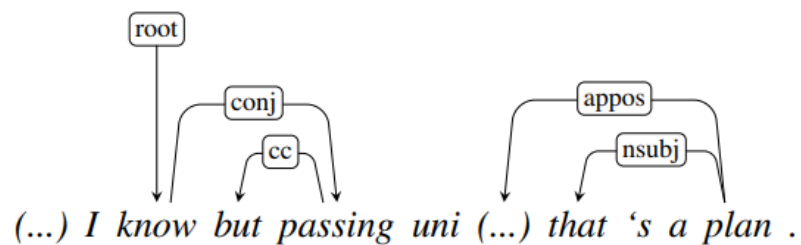
Comparing different annotations (LAS)

	TB	2018	2024
TB	100.0		
2018	74.0	100.0	
2024	72.1	70.3	100.0

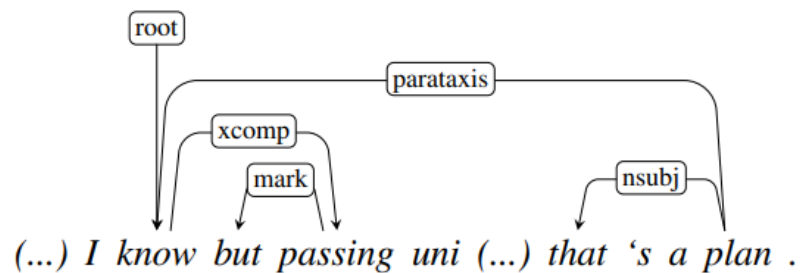
Example



(a) KKH



(b) MoNoise: *plan* is apppos to *go*, see (g).



(c) Tweepbank v2

Figure 6: Annotation subsets of the tweet in (g).

Compare to standard

- Experienced linguist converges the different annotations to a gold standard (SyntacTw)
- Addendum to UD guidelines

Comparing different annotations (LAS)

	TB	2018	2024	SYNTACTW
TB	100.0			
2018	74.0	100.0		
2024	72.1	70.3	100.0	
SYNTACTW	80.4	75.5	83.8	100.0

What about automated parsers?

- Train MaChAmp toolkit (deep-biaffine parser)
- RoBERTa-base
- TweepBank 2



Comparing different annotations (LAS)

	TB	2018	2024	SYNTACTw	MAChAMP
TB	100.0				
2018	74.0	100.0			
2024	72.1	70.3	100.0		
SYNTACTw	80.4	75.5	83.8	100.0	
MAChAMP	69.8	63.7	62.5	67.1	100.0

Overall scores (internal agreement/parser) are lower compared to previous studies

- Non-standardness of data
- We do annotations from scratch, avoiding anchoring (Berzak et al. 2016)
- Others had clearer guidelines from the beginning

Are we close to a performance ceiling for parsing social media data?

	TB	2018	2024	SYNTACTW	MAChAMP
TB	100.0				
2018	74.0	100.0			
2024	72.1	70.3	100.0		
SYNTACTW	80.4	75.5	83.8	100.0	
MAChAMP	69.8	63.7	62.5	67.1	100.0

Lets ask another experienced linguistic annotator

- Annotator selects preferred structure across two annotations
- Blind, order=mixed

Lets ask another experienced linguistic annotator

PREFERENCE	COUNT	%
SYNTACTWEETS	85	80.20
MACHAMP	7	6.60
NEITHER	5	4.72
N/A	9	8.49

Table 2: Pairwise comparison of gold trees and predicted trees for 106 tweets by an external annotator. N/A means that the trees were identical.

Lets ask another experienced linguistic annotator

- **Humanly annotated structures are better**

What about aligning parsers?

- Evaluation
- Parser implementation

Use distribution for evaluation: Dependency Total Variation Similarity

- Loop over all words (i)
 - Loop over each label (j)
 - Sum absolute differences

$$\sum_{j=0}^{len(i)} |g_{ij} - p_{ij}|$$

Use distribution for evaluation: Dependency Total Variation Similarity

- Loop over all words (i)
 - Loop over each label (j)
 - Sum absolute differences
- Divide by 2 to make range 0-1
- Add 1- to make higher = better

$$DepTVS_i = 1 - \frac{\sum_{j=0}^{len(i)} |g_{ij} - p_{ij}|}{2}$$

Use distribution for evaluation: Dependency Total Variation Similarity

- Loop over all words (i)
 - Loop over each label (j)
 - Sum absolute differences
- Divide by 2 to make range 0-1
- Add 1- to make higher = better

$$DepTVS_i = 1 - \frac{\sum_{j=0}^{len(i)} |g_{ij} - p_{ij}|}{2}$$

Building a parser

Deep-biaffine parser

- Predictions every word to every word
- Chu-Liu algorithm

Building a parser

Deep-biaffine parser

- Predictions every word to every word
- ~~Chu-Liu algorithm~~

Use distribution for evaluation: Dependency Total Variation Similarity

	HEADS	LABELS	AVG.
DepTVS	75.37	77.79	76.58
top-1 _{biaf}	74.60	77.46	76.03
top-1 _{full}	74.60	77.46	76.03
%found	91.28	82.66	86.97

Table 3: DepTVS when comparing the full distributions for both heads and labels, and the scores when only using a single prediction for each sub-task. Also the % of correctly found labels for each task.

Conclusions

- I learned some things from Gertjan
- People disagree
- Parsing is not solved
- We can evaluate against a distribution of annotators
- Distribution of deep biaffine parser has some relevant signal encoded

Many thanks!

- Gertjan
- Barbara
- Kia
- Sonnik
- Gosse

Thanks!

IT UNIVERSITY OF COPENHAGEN



CARLSBERG
FONDET

