

MWE-UD2024 - May 24, 2024

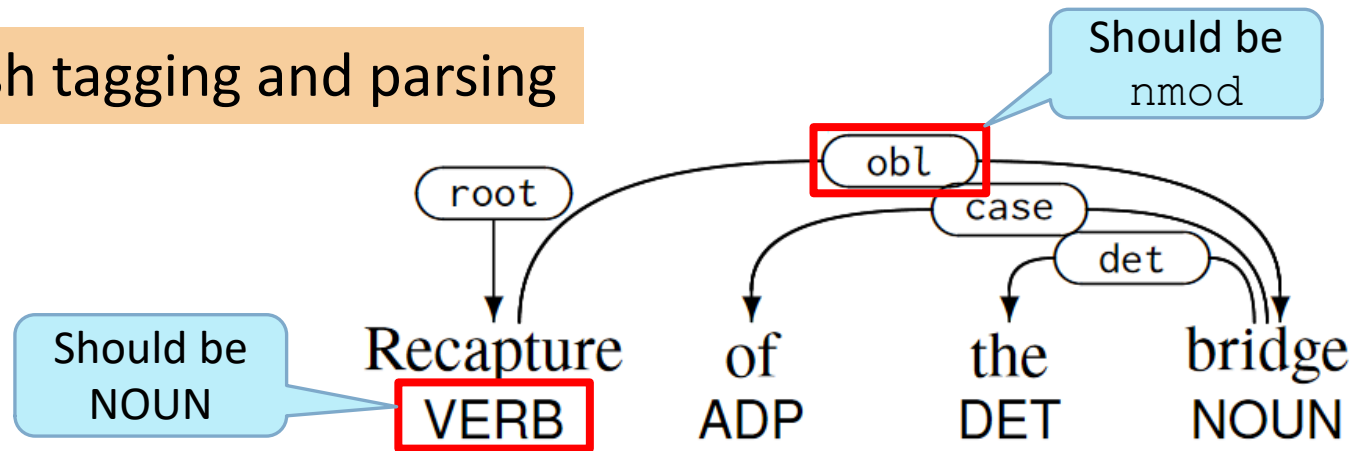
Automatic Manipulation of Training Corpora to Make Parsers Accept Real-world Text

Hiroshi Kanayama, Ran Iwamoto, Masayasu Muraoka,
Takuya Ohko, Kohtaroh Miyamoto

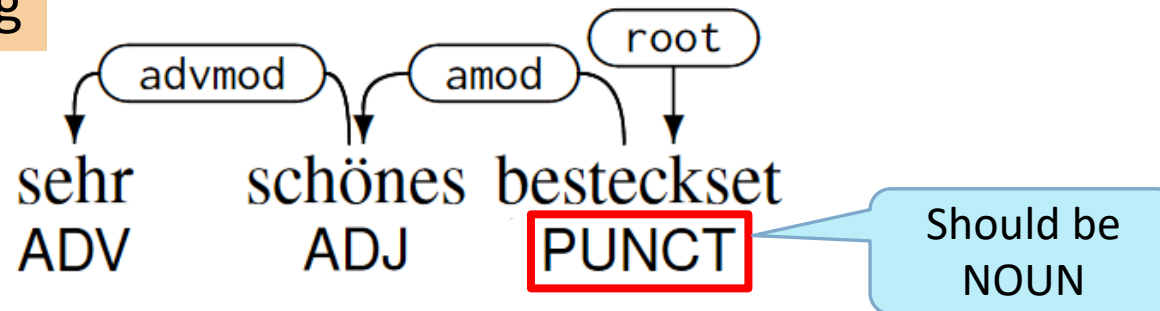
IBM Research - Tokyo

Frustrating examples: tagging and parsing errors in noun phrases

English tagging and parsing



German tagging

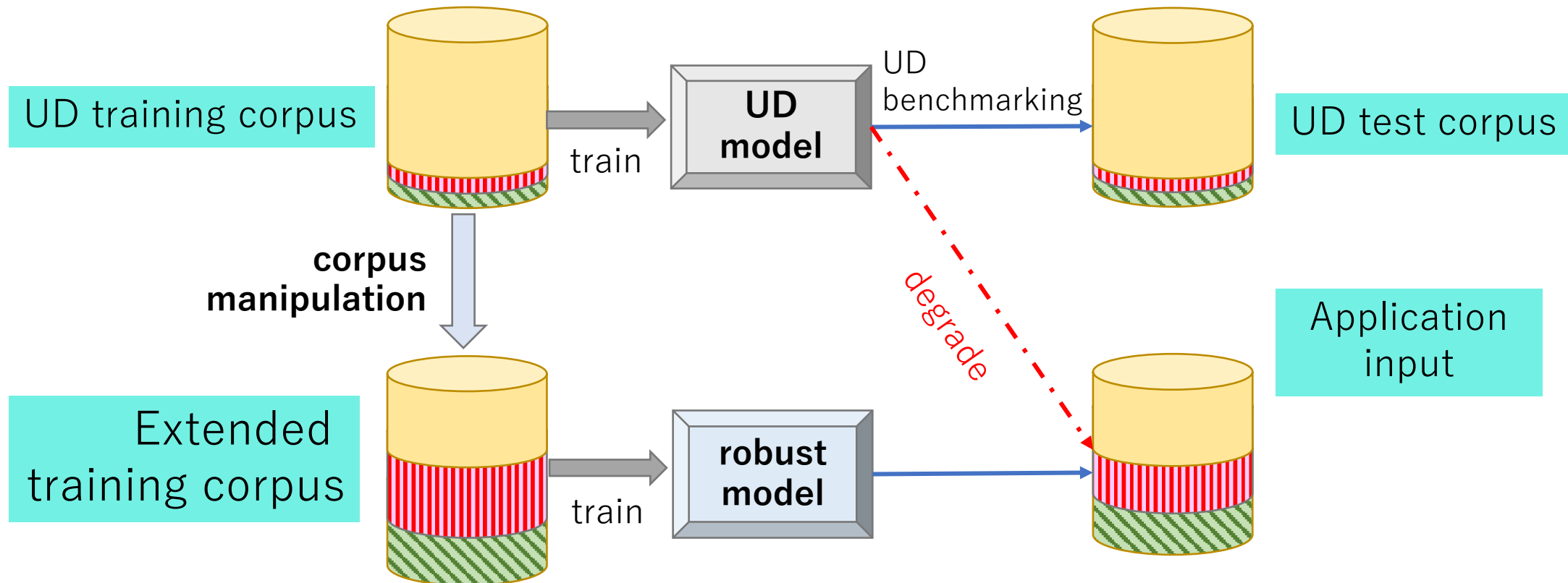


Noun phrases and omitted punctuation – Appear in real input, but not in UD corpora



- Frequently appear in real-word text
 - Title of documents and sections
 - Informal data, Review comments, ...
- Not appear in UD corpora
 - Not in training data → Existing parsers cannot handle
 - Not in test data → This problem was overlooked

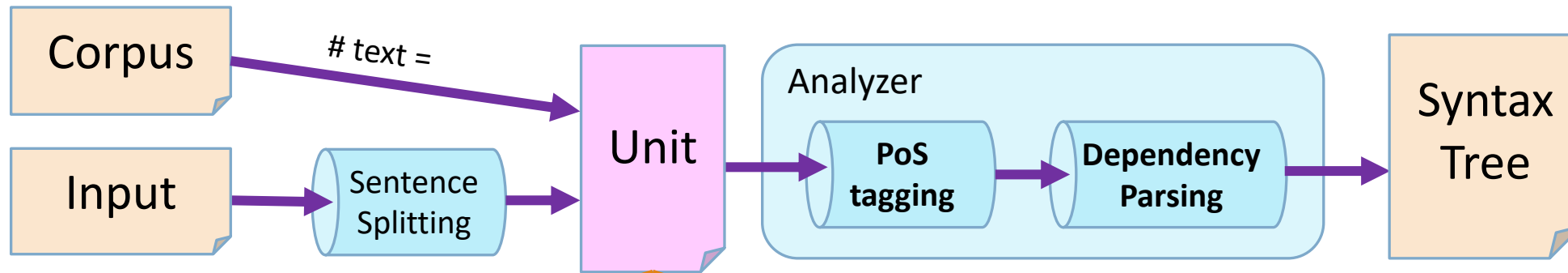
Discrepancy between UD and real-word → Automatic manipulation of training corpora



1. Problem setting – corpus discrepancy
2. Corpus Manipulation
3. Evaluation
 - Unit test with noun phrase data
 - Intrinsic and extrinsic evaluation

1. Problem setting – corpus discrepancy
2. Corpus Manipulation
3. Evaluation
 - Unit test with noun phrase data
 - Intrinsic and extrinsic evaluation

Sentence vs. noun phrase in a unit to apply tagging and parsing



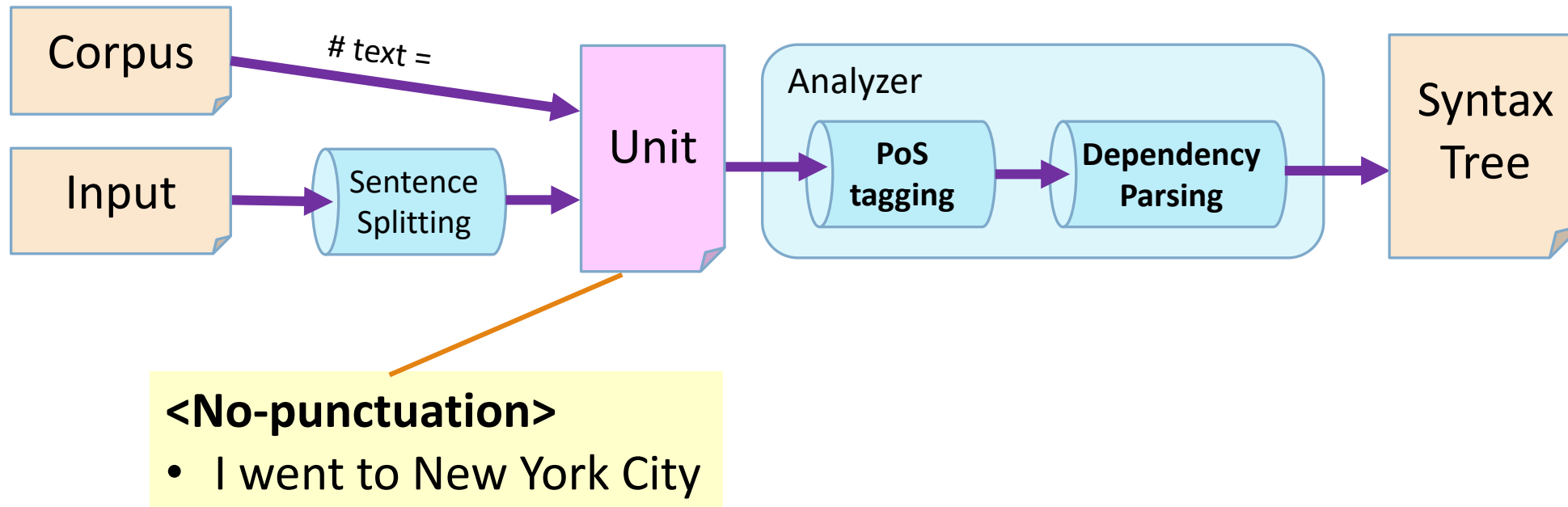
<Sentence>

- This hotel was excellent.

<Noun Phrase>

- An excellent hotel on top of a hill.

No-punctuation: omitted period at the end of a unit



Ratios of noun phrases and no-punctuation are very different between UD and Review data

	Noun phrase (%)		No-punctuation (%)	
	UD	Review	UD	Review
German	2.4	28.0	0.4	12.0
French	2.6	36.0	1.9	3.0
Spanish	2.6	25.0	0.2	7.0
English	6.5	3.0	14.0	1.0

UD_English-EWT

UD:

Wir hatten wunderschöne Spaziergänge und die Städte der Region mit Ihren Gründerzeithäusern sind sehenswert.

Review: (SemEval, etc.)

Ein gutes Besteck für jeden Tag.

1. Problem setting – corpus discrepancy
2. Corpus Manipulation
3. Evaluation
 - Unit test with noun phrase data
 - Intrinsic and extrinsic evaluation

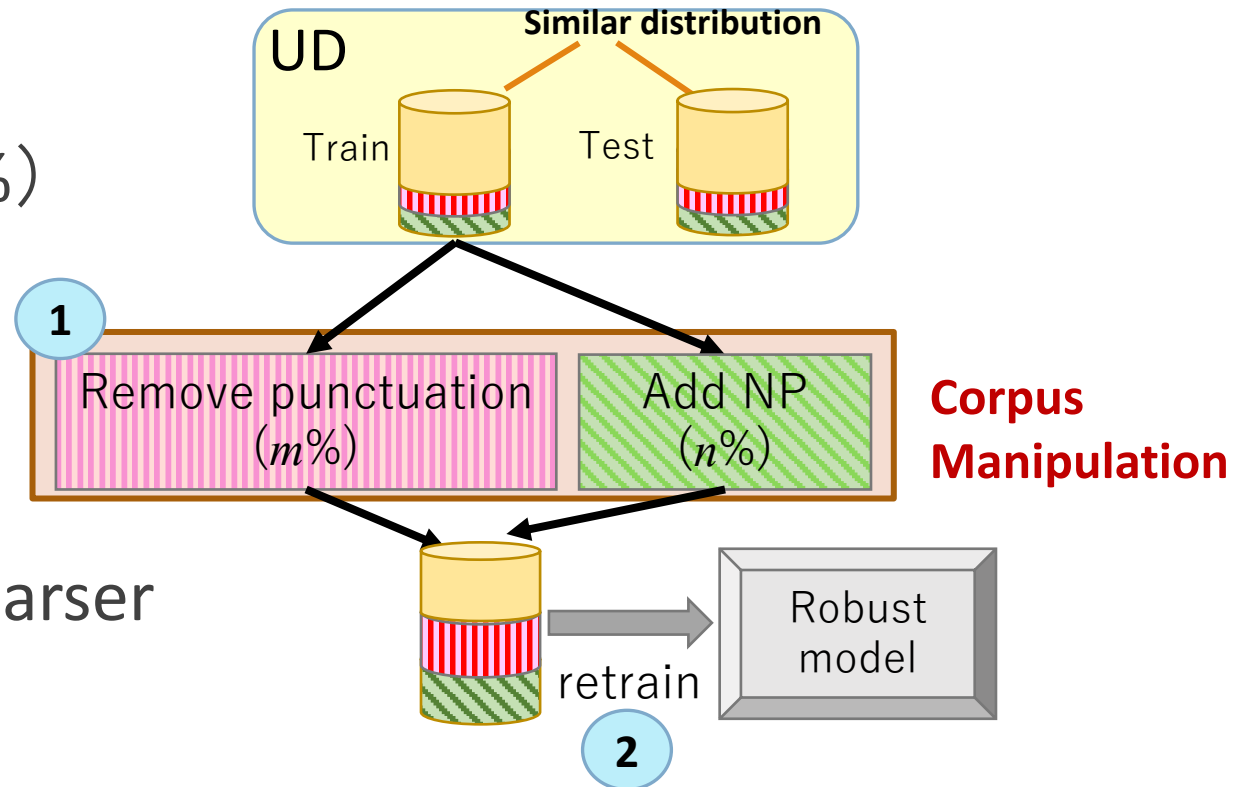
To overcome the corpus discrepancy manipulate training corpus and retrain the model

1. Corpus manipulation

- Removing punctuation ($m\%$)
- Adding noun phrases ($n\%$)

2. Model retraining

- PoS tagger + dependency parser
- Tested on Stanza parser



Remove punctuation:

Just remove sentence-end periods in $m\%$ of units

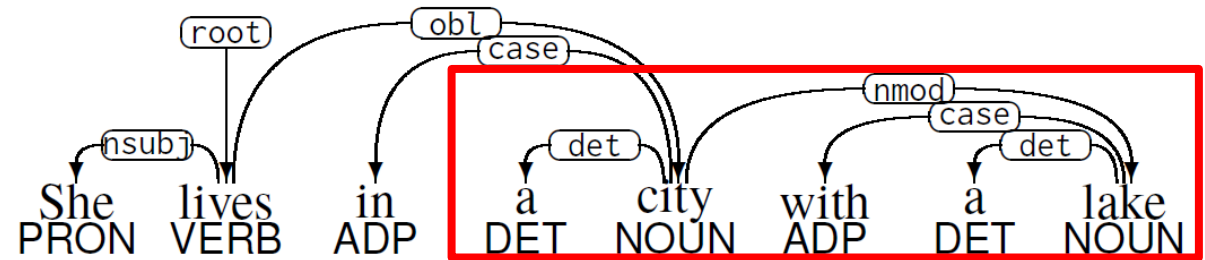
1	His	he	PRON	PRP\$	Gender=Masc Number=Sing Person=3	3 nmod:poss	—	—
2	superior	superior	ADJ	JJ	Degree=Pos	3 amod	—	—
3	officers	officer	NOUN	NNS	Number=Plur	4 nsubj	—	—
4	said	say	VERB	VBD	Mood=Ind Tense=Past VerbForm=Fin	0 root	—	—
5	OK	ok	INTJ	UH	—	4 obj	—	SpaceAfter=No
6	.	.	PUNCT	.	—	4 punct	—	—



1	His	he	PRON	PRP\$	Gender=Masc Number=Sing Person=3	3 nmod:poss	—	—
2	superior	superior	ADJ	JJ	Degree=Pos	3 amod	—	—
3	officers	officer	NOUN	NNS	Number=Plur	4 nsubj	—	—
4	said	say	VERB	VBD	Mood=Ind Tense=Past VerbForm=Fin	0 root	—	—
5	OK	ok	INTJ	UH	—	4 obj	—	SpaceAfter=No

Increase noun phrases by $n\%$, by extracting noun phrases in sentences

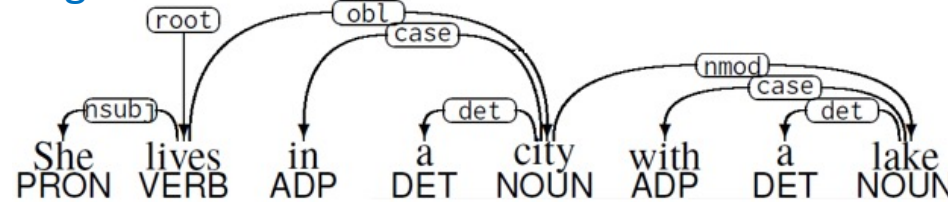
- Detect noun phrases (non-root)
 - Subtree consists of ≥ 4 words headed by "NOUN"
 - Exclude words of case and punct



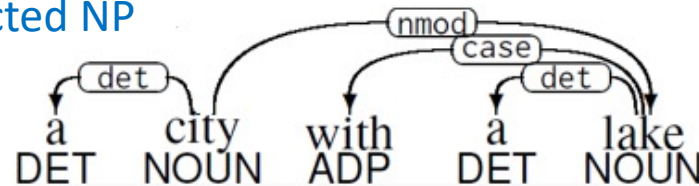
$n\%$

- Randomly pick up NPs to increase the training corpus to $(100+n)\%$

Original sentence



Extracted NP



1. Problem setting – corpus discrepancy
2. Corpus Manipulation
3. Evaluation
 - Unit test with noun phrase data
 - Intrinsic and extrinsic evaluation

Conduct three types of evaluation with the retrained model

1. Unit test on Noun Phrase Data

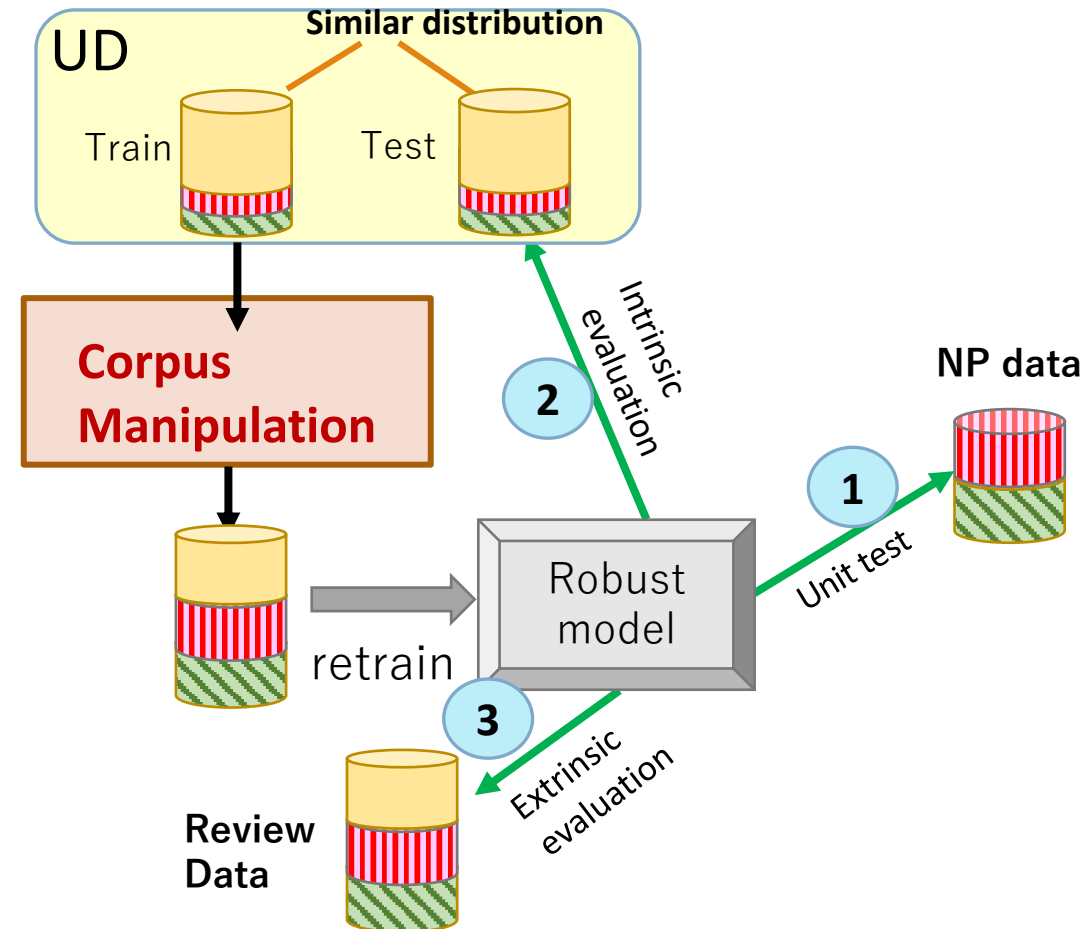
- The root word is tagged as NOUN?
- Isn't there terrible PUNCT errors?

2. Intrinsic evaluation on UD

No degrade on the parsing score?

3. Extrinsic evaluation on Review Data

Sentiment extraction is improved?



Generated Noun Phrase Data using Wikipedia's section titles

- Wikipedia section titles, 3 words or more
- Exclude ones with special characters
- Subsampling to diversify

→ Obtained 1,500 instances per language

English :

boundary extension and different brains
ties with groups marked as terror organizations
evidence for innate language capacities

French :

panthéon de la musique canadienne
commandeurs avec plaque
viroïde de la maladie des tubercules en fuseau



https://en.wikipedia.org/wiki/Morning_Musume

Successfully avoid NP and no-punct problems without degrading general scores

German

Varying m and n

Baseline: UD as it is

The better, the more NPs added

Perfectly removed stupid PUNCT errors

No degrade; Sometimes improved

Good balance of m and n

Removing all punctuation is not good

Remove Punct	Add NP	Unit Test		Intrinsic	Extrinsic
$m\%$	$n\%$	NOUN (↑)	Wrong Punct (↓)	LAS on UD	Sentiment F2
0	0	97.4	3.2	79.68	81.2
0	20	97.7	0	79.64	81.7
0	50	98.1	0	79.23	80.2
0	100	98.4	0	79.60	80.5
10	10	97.8	0	79.87	82.8
20	0	97.1	0	78.98	80.9
20	10	97.5	0	80.20	82.2
50	0	97.3	0	79.73	79.7
100	0	97.4	0	76.78	80.3

Good results in 4 languages in all aspects though the optimal m and n are different

German	m	n	UD	Sentiment	NP	wrong PUNCT
	0	0	79.68	81.2	97.4	3.2
	0	20	79.64	81.7	97.7	0
	0	50	79.23	80.2	98.1	0
	0	100	79.60	80.5	98.4	0
	10	10	79.87	82.8	97.8	0
	20	0	78.98	80.9	97.1	0
	20	10	80.20	82.2	97.5	0
	50	0	79.73	79.7	97.3	0
	100	0	76.78	80.3	97.4	0

French	m	n	UD	Sentiment	NP	wrong PUNCT
	0	0	87.14	73.9	91.4	4.2
	0	20	87.25	74.9	93.2	0
	0	50	86.76	74.0	94.4	0
	0	100	86.37	74.5	95.5	0
	10	10	87.09	74.3	93.2	0
	20	0	87.31	73.5	90.6	0
	20	10	87.19	73.5	92.9	0
	50	0	87.57	74.7	91.1	0
	100	0	84.57	73.2	92.3	0

Spanish	m	n	UD	Sentiment	NP	wrong PUNCT
	0	0	87.58	69.8	91.5	4.1
	0	10	88.21	69.4	93.1	1
	0	50	87.37	71.1	94.2	1
	0	100	87.59	70.0	94.7	0
	10	10	87.67	68.4	92.7	0
	20	0	87.28	69.2	93.5	0
	20	10	87.28	69.7	93.1	0
	50	0	87.52	70.1	91.0	0
	100	0	86.83	70.4	91.9	0

English	m	n	UD	Sentiment	NP	wrong PUNCT
	0	0	83.84	78.7	91.6	0.7
	0	10	84.09	78.8	93.9	1
	0	50	83.88	78.4	95.3	0
	0	100	84.00	78.5	95.3	0
	10	0	83.81	78.2	91.7	0
	10	10	83.71	78.8	94.2	0
	50	0	84.03	79.5	91.4	2
	50	50	83.75	77.6	95.4	0
	100	0	83.46	78.7	90.1	0

Conclusion

Corpus manipulation made parsers more robust for real-world input


- Handling of Noun Phrases and no-punctuation
 - Confirmed discrepancy between UD corpora and real text
 - Proposed algorithms of automatic conversion
- Showed results in 4 languages
 - Improvements in unit test, intrinsic (UD) and extrinsic (SA) evaluation
 - Worked in English as well, even with different trends
 - Future work: cover other languages (e.g. Japanese)

Should UD corpora be modified?

<https://github.com/stanfordnlp/stanza/issues/471>


Hebrew parser ends with punct if there is no "." #471

🔒 Closed KoichiYasuoka opened this issue on Sep 19, 2020 · 8 comments



KoichiYasuoka commented on Sep 19, 2020

```
>>> import stanza
>>> nlp=stanza.Pipeline("he")
>>> doc=nlp("אֵל לֹא הֵתוּכַח" לֵאמֹר)
>>> print(doc)
[
```



AngledLuffa commented on Oct 23, 2020 via email

Good to hear. What I did was added "data augmentation" to the training of two models for two languages - a fancy way of saying I removed the final punctuation from 10% of the sentences.

😊 🎉 1 ❤️ 1

