The Role of Syntactic Span Preference in Post-Hoc Explanation Disagreement



Jonathan Kamp, Lisa Beinborn, Antske Fokkens

Computational Linguistics and Text Mining Lab

Vrije Universiteit Amsterdam

LREC-Coling 2024, May 20-25, Turin

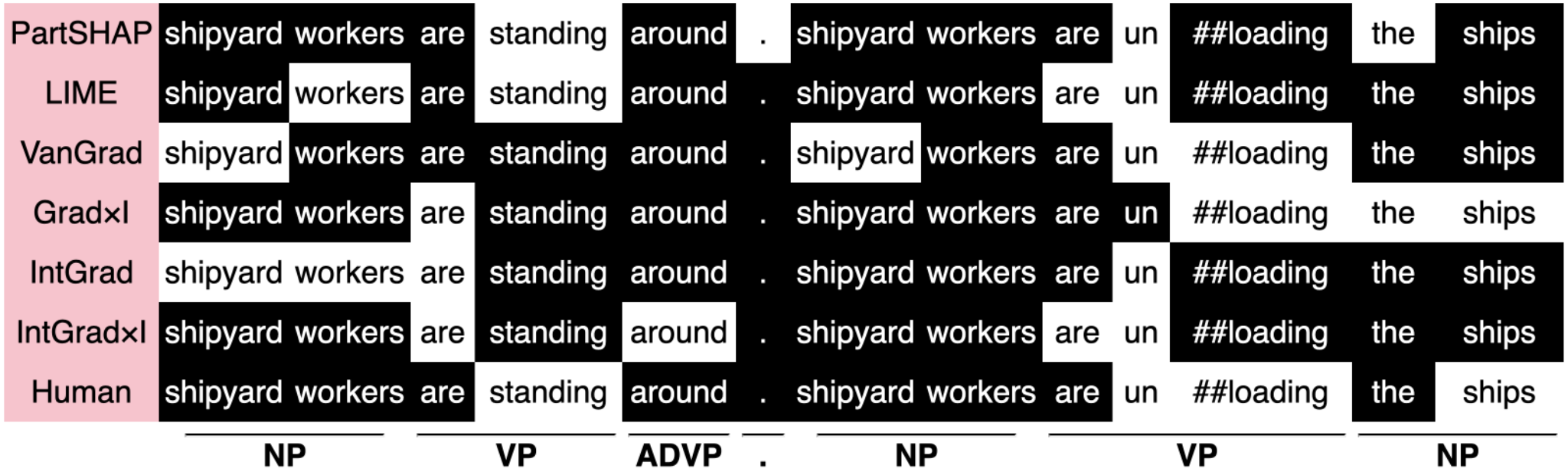
Are **tokens too strict**
for evaluating feature attribution methods?

Are tokens too strict?

- Post-hoc attribution methods assign importance scores to tokens
- Post-hoc attribution methods tend to disagree on token importance
  - Top-4 important tokens method_A ≠ Top-4 important tokens method_B

- Are we evaluating too strictly?
- Should we evaluate at a lower granularity level?

Example, $k = 4$


PartSHAP shipyard workers are standing around . shipyard workers are un ##loading the ships
LIME shipyard workers are standing around . shipyard workers are un ##loading the ships
VanGrad shipyard workers are standing around . shipyard workers are un ##loading the ships
Grad×I shipyard workers are standing around . shipyard workers are un ##loading the ships
IntGrad shipyard workers are standing around . shipyard workers are un ##loading the ships
IntGrad×I shipyard workers are standing around . shipyard workers are un ##loading the ships
Human shipyard workers are standing around . shipyard workers are un ##loading the ships
NP VP ADVP . NP VP NP

In this paper…

- We study potential sources of disagreement across attribution methods from a **linguistic perspective**.

- We find that
  - different methods systematically select different word classes
  - methods that agree most with other methods and with humans display similar linguistic preferences
  - token-level differences between methods are smoothed out if we compare them on the syntactic span level

In this paper…

- We also systematically investigate the interaction between top-*k* and spans and propose an improved configuration for selecting important tokens.

Setup

- e-SNLI dataset (Camburu *et al.*, 2018)
- We finetune DistilBERT on a multi-class NLI task
  - Output labels: [contradiction, entailment, neutral]
  - Output labels indicate the relation between a premise and hypothesis
  - Balanced classes
  - ~0.9 accuracy

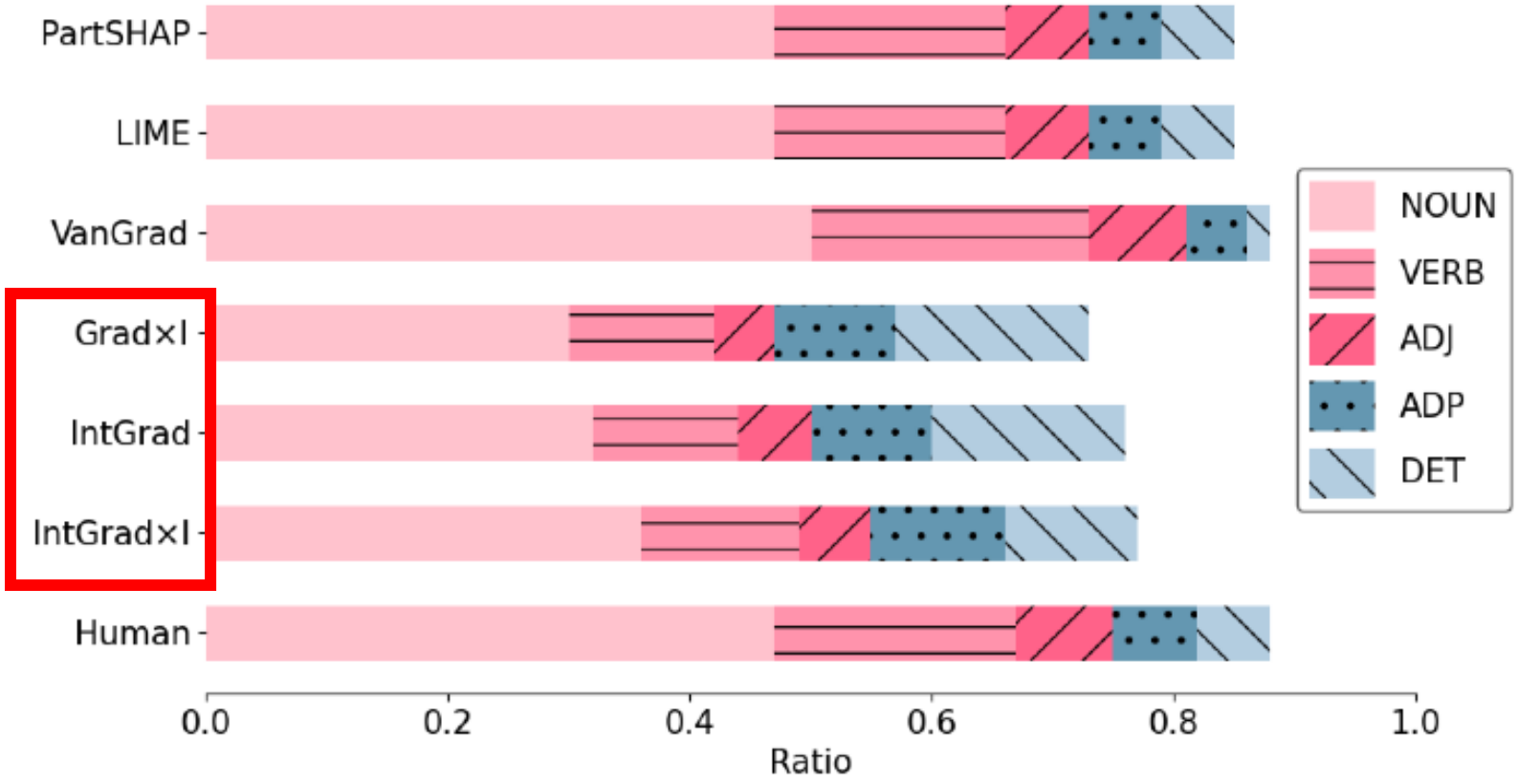
Word class preferences

- Among the tokens in a sentence,
  which word classes are methods most likely to target?

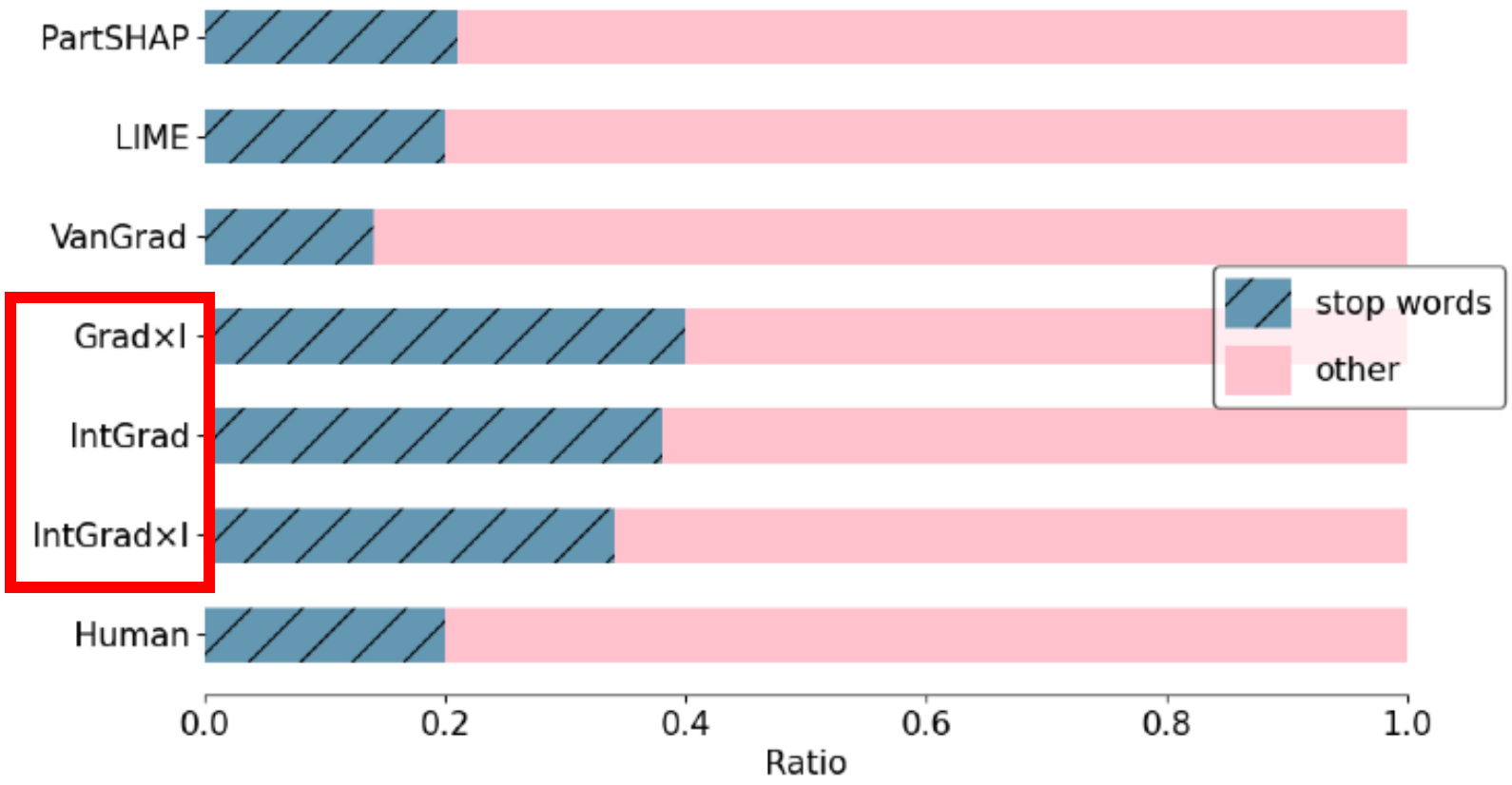
Word class preferences

- Among the tokens in a sentence,
  which word classes are methods most likely to target?
  - Part-of-speech (POS)
  - Stop words
  - Punctuation

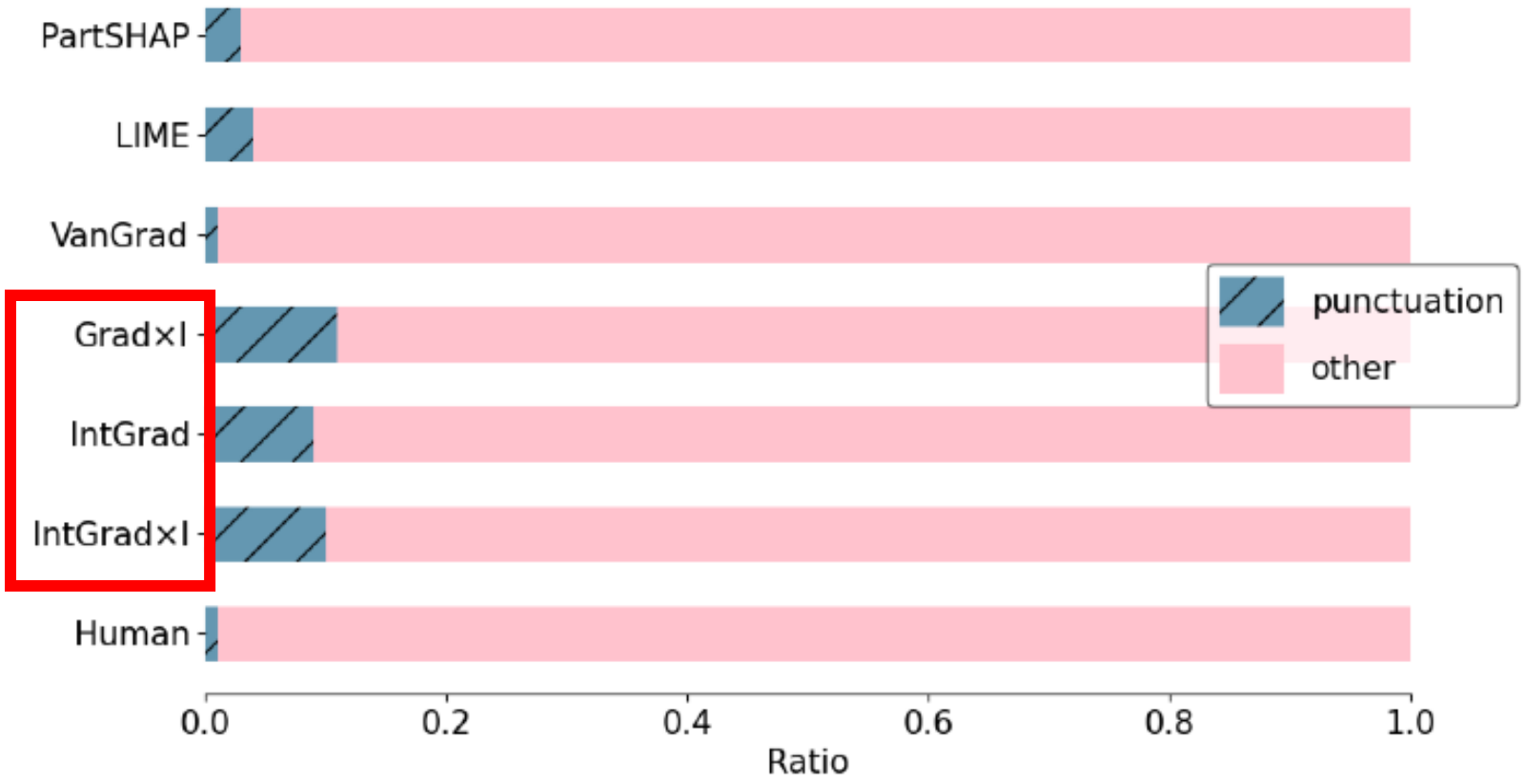
Word class preferences: *POS*



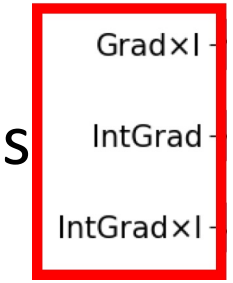
Word class preferences: *stop words*



Word class preferences: *punctuation*

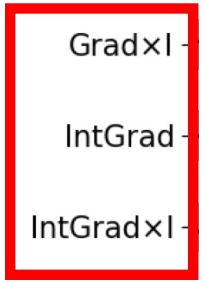


Word class preferences

- Hypothesis: preferences 

≠ preferences other methods

--> Largely confirmed through pairwise Chi-Square tests

- We find that 

are those with lowest agreement with other methods and with human preference (Kamp *et al.*, 2023).

Kamp, J., Beinborn, L., & Fokkens, A. (2023). Dynamic Top-k Estimation Consolidates Disagreement between Feature Attribution Methods

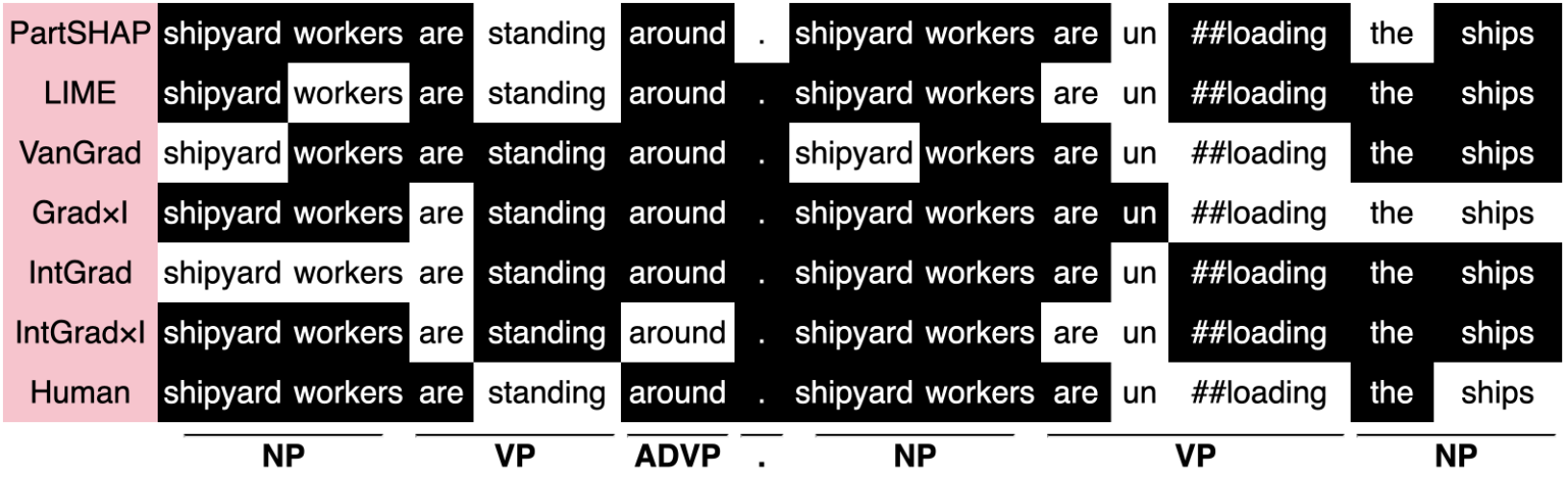
Are tokens too strict?

- Different methods: different word class preferences
- Do we expect too much in terms of token-level agreement?
- Lexical bias / (un)frequency of words may play a role

Are tokens too strict?

- Different methods: different word class preferences
- Do we expect too much in terms of token-level agreement?
- Lexical bias / (un)frequency of words may play a role

- Do methods perhaps agree more at the syntactic span level?



Span definition

- How to obtain syntactic span units?
  - Flair Chunker (Akbik *et al*., 2018)
  - Discrete, non-overlapping phrases

- We consider that a span is *targeted* if it contains ≥1 *targeted* token
  - On average, 4 tokens --> 3.5 spans

Head vs. modifier preference

- Which patterns in the data explain disagreement?
- We zoom in on Vanilla Gradient (nouns ☺) vs. Gradient × Input (nouns ☹)

Head vs. modifier preference

- Which patterns in the data explain disagreement?
- We zoom in on Vanilla Gradient (nouns ☺) vs. Gradient × Input (nouns ☹)
- Spans can be complex, e.g. [DET, ADV, ADJ, NOUN]

Head vs. modifier preference

- Which patterns in the data explain disagreement?
- We zoom in on Vanilla Gradient (nouns ☺) vs. Gradient × Input (nouns ☹)
- Spans can be complex, e.g. [DET, ADV, ADJ, NOUN]
- What if we look at something simple? → *targeted* NP spans: [DET, NOUN]

Head vs. modifier preference

- Which patterns in the data explain disagreement?
- We zoom in on Vanilla Gradient (nouns ☺) vs. Gradient × Input (nouns ☹)
- Spans can be complex, e.g. [DET, ADV, ADJ, NOUN]
- What if we look at something simple? → *targeted* NP spans: [DET, NOUN]
  - Vanilla Gradient (☺) targets NOUN ~100% of the times
  - Gradient × Input (☹) targets DET ~50% of the times
  - Ratio of targeted tokens in these spans is comparable (57-60%): both methods mostly target either DET or NOUN (rarely both).

Head vs. modifier preference

- Which patterns in the data explain disagreement?
- We zoom in on Vanilla Gradient (nouns ☺) vs. Gradient × Input (nouns ☹)
- Spans can be complex, e.g. [DET, ADV, ADJ, NOUN]
- What if we look at something simple? → *targeted* NP spans: [DET, NOUN]
  - Vanilla Gradient (☺) targets NOUN ~100% of the times
  - Gradient × Input (☹) targets DET ~50% of the times
  - Ratio of targeted tokens in these spans is comparable (57-60%): both methods mostly target either DET or NOUN (rarely both).

- Take-away: methods targeting different word classes may translate to systematic, alternating differences within syntactic spans.

Span agreement

Agreement results (ranges from 0.5 to 1.0)

- Token-level: 0.61
- Span-level: 0.69

Span agreement

Agreement results (ranges from 0.5 to 1.0)

- Token-level: 0.61 ……(+0.07)
- Span-level: 0.69 ……(+0.12)

(we considered a random baseline showing that
agreeing on low values of *k*
is similarly difficult for tokens vs. spans)

Dynamic *k* for selecting important tokens

- We used dynamic *k* to automatically obtain the top-*k* selections of important tokens (opposed to a fixed *k*, e.g. top-4).

Dynamic *k* for selecting important tokens

- We used dynamic *k* to automatically obtain the top-*k* selections of important tokens (opposed to a fixed *k*, e.g. top-4).

- Downside: dynamic *k* may select high values of *k,* inflating the agreement score
  --> we explore different dynamic *k* settings

Dynamic *k* for selecting important tokens

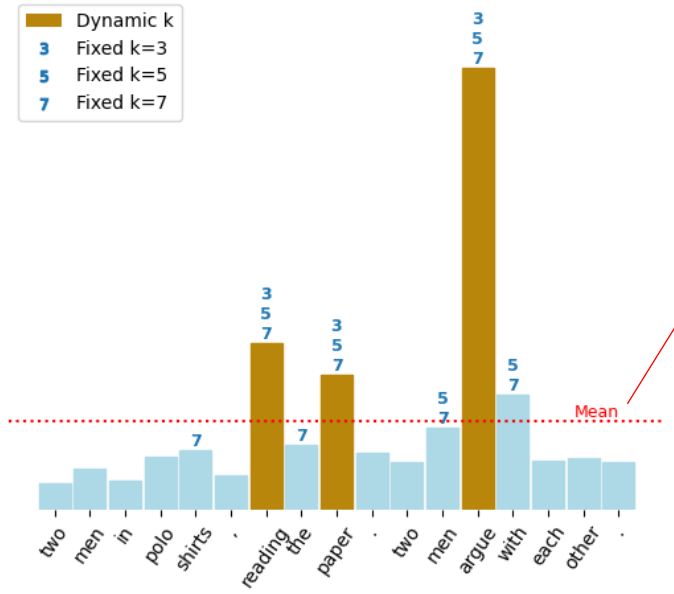
- We used dynamic *k* to automatically obtain the top-*k* selections of important tokens (opposed to a fixed *k*, e.g. top-4).

- Downside: dynamic *k* may select high values of *k,* inflating the agreement score
  --> we explore different dynamic *k* settings

- Requirements to keep the resulting values of *k* “low” :

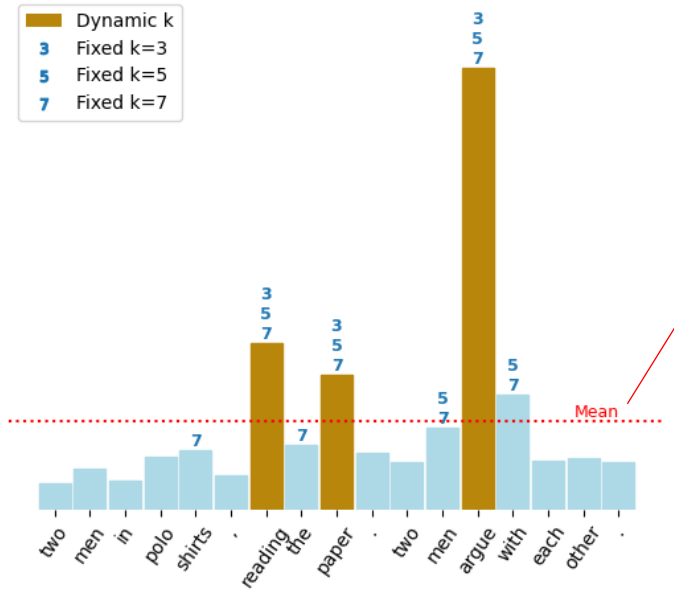
1) close to the human preference of 4±3;
2) ability to outperform a pseudo-random baseline (see paper for details)

Adjusting Dynamic *k*



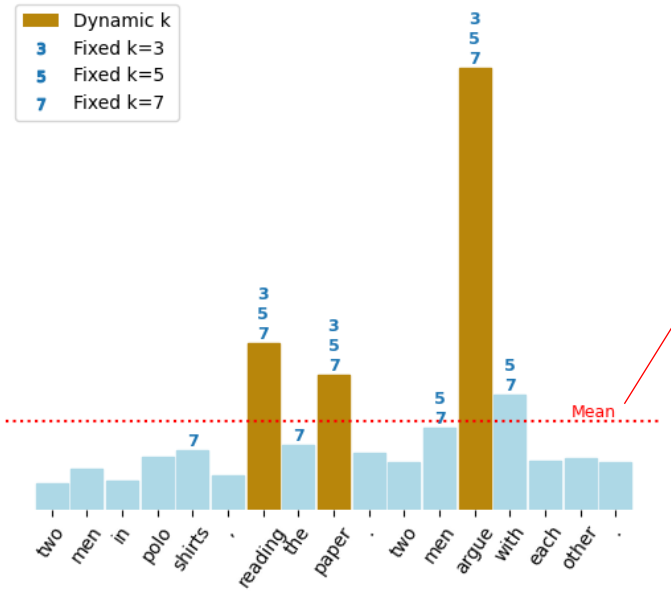
μ (current threshold)

Adjusting Dynamic *k*



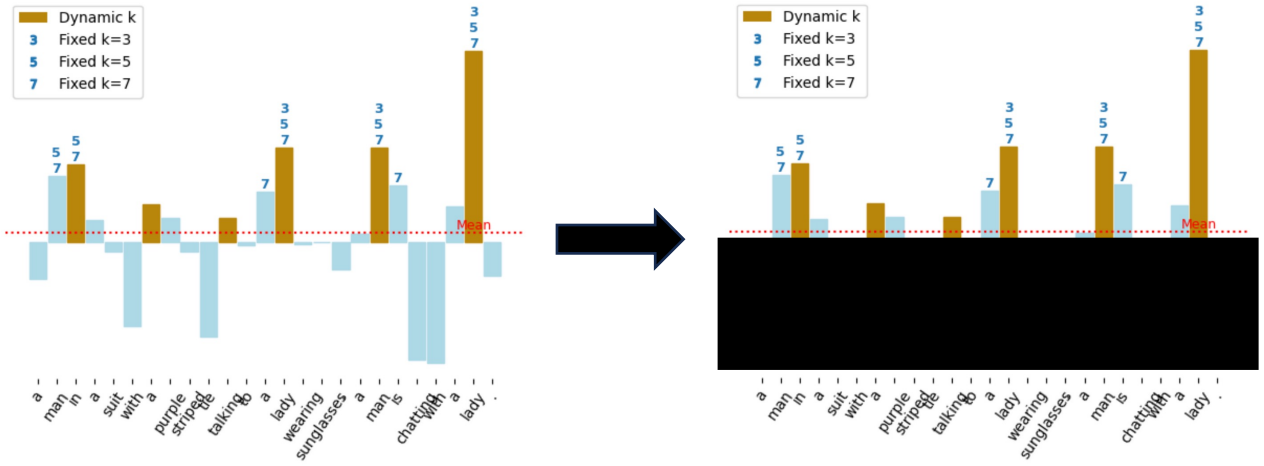
μ (current threshold)
μ + σ
μ - σ
μ + 2σ
μ - 2σ
median

Adjusting Dynamic *k*



μ	(current threshold)
μ + σ	
μ - σ	
μ + 2σ	
μ - 2σ	
median	

μ	of positively important scores to target class
μ + σ	of positively important scores to target class
μ - σ	of positively important scores to target class
μ + 2σ	of positively important scores to target class
μ - 2σ	of positively important scores to target class
median	of positively important scores to target class



Adjusting Dynamic *k*

Many values of *k*!

		Thresholds					
	Method	μ	$\mu+\sigma$	$\mu+2\sigma$	$\mu-\sigma$	$\mu-2\sigma$	*median*
all	PartSHAP	4.54±1.73	2.16±0.95	1.25±0.65	7.36±1.89	7.37±1.89	6.19±1.62
	LIME	5.34±2.35	2.24±1.05	1.23±0.65	8.31±2.86	8.32±2.87	7.13±2.48
	VanGrad	4.58±1.68	2.41±1.02	1.39±0.61	7.63±2.68	7.64±2.69	6.20±2.08
	Grad×I	6.83±2.59	2.39±1.12	0.68±0.65	8.21±2.82	8.28±2.83	7.08±2.51
	IntGrad	7.30±2.63	2.66±1.23	0.64±0.63	8.41±2.88	8.46±2.90	7.41±2.58
	IntGrad×I	5.68±2.37	2.27±1.08	1.02±0.62	8.04±2.80	8.07±2.82	6.83±2.39
>0	PartSHAP	3.34±1.33	1.86±0.82	1.07±0.55	7.00±2.01	7.28±1.93	5.01±1.61
	LIME	3.56±1.56	1.87±0.87	1.06±0.53	7.95±2.91	8.25±2.89	5.59±2.04
	VanGrad	4.58±1.68	2.41±1.02	1.39±0.61	7.63±2.68	7.64±2.69	6.20±2.08
	Grad×I	3.51±1.56	1.75±0.83	0.73±0.56	6.81±2.82	7.86±2.92	4.69±1.88
	IntGrad	3.47±1.60	1.67±0.81	0.62±0.55	6.60±2.81	7.81±3.05	4.54±1.86
	IntGrad×I	3.83±1.69	1.91±0.90	0.98±0.53	7.57±2.74	7.99±2.81	5.38±1.96

Table 1: Values of *k* for different global importance thresholds. The three methods that yield values of k closest to human preference are visually indicated with a dark background.

→ Winner, based on
1) Closeness to human pref.
2) Baseline comparison

:

μ > 0

Inspecting disagreement further

- Linguistic, ”surface” patterns that we observe in the predictions

Inspecting disagreement further

- Underlying mechanics of different methods leading to different scores

- Linguistic, ”surface” patterns that we observe in the predictions

Inspecting disagreement further

- Underlying mechanics of different methods leading to different scores
- Something in between: locality versus lexicality
- Linguistic, ”surface” patterns that we observe in the predictions

Thanks for listening and feel free to reach out!

Code:
github.com/jbkamp/repo-Span-Pref

Website: jbkamp.github.io
E-mail: j.b.kamp@vu.nl
Twitter: jb_kamp