HUN REN SZTAKI
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS
Comenius University
Bratislava

INDIRECT PROBING EXPERIMENTS

ENDRE HAMERLIK

28.02.2024 – HLT Seminar

ABOUT ME

- PhD student at Comenius University in Bratislava
- Profile: xAI, NLP, Adversarial Neural Networks
- Projects with HUN-REN SZTAKI - HLT Group
  - Probing
  - Summarization
- Collabortion with Kempelen's Institute of Intelligent Technologies
  - News dataset creation
  - Political stance classification
- Working at a hungarian startup
  - Applied NLP in the governmental sector

ABOUT THE PROJECT

- Inspiration: Ács (2023)'s Perturbed probing experiments
- Not yet articulated findings of it:
  - Randomly weighted MLMs
  - Left-context dependence
- Open questions:
  - Explanation for the asymmetric context dependence
  - How valid probing is?

HOW TO TREAT THE INTERNAL REPRESENTATIONS IN VISION MODELS?

- Finding „exciting” examples for specific hidden activations
  - Motivation by Quiroga et. al (2005)
- Dataset examples that maximally activate specific hidden neurons
- Optimizing an input which would excite the selected neurons even more!

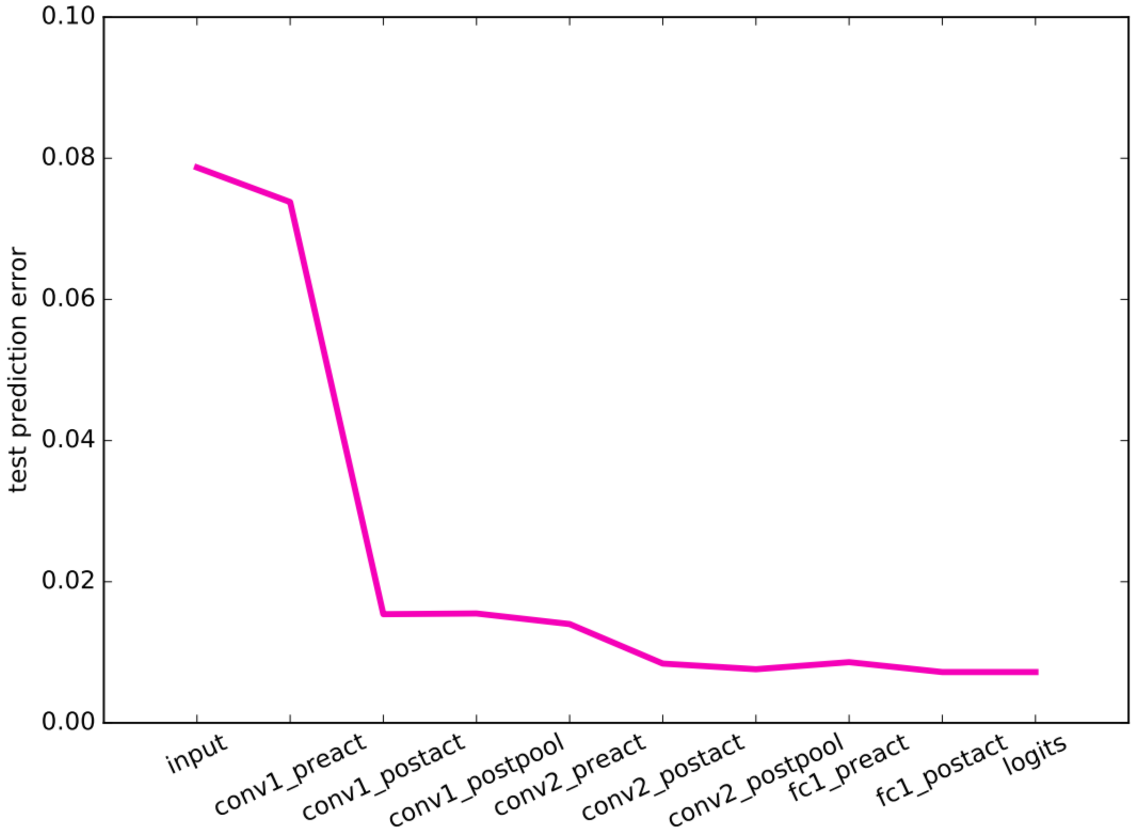
Class visualization
of the class black widow
(BSCV)

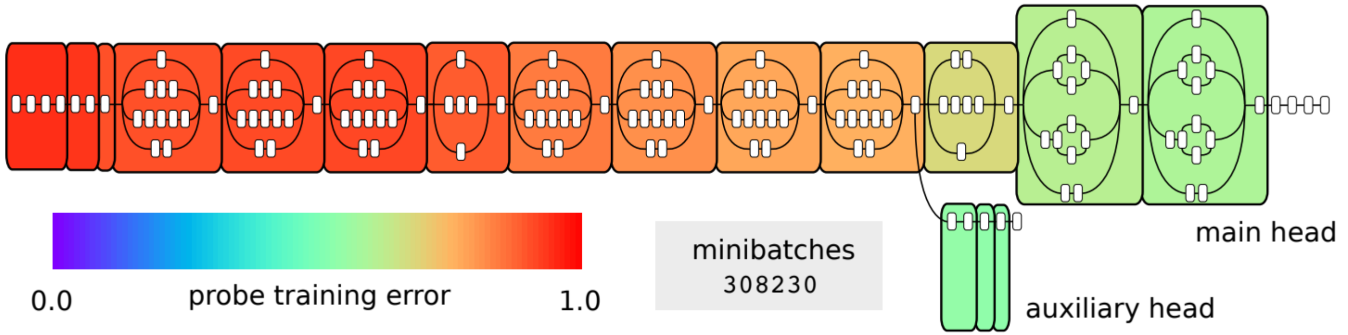
Dataset examples
Of the black widow

Source: Hamerlik, Endre, Deb, Mayukh and Takáč, Martin: „Bi-Source Class Visualization: An Adversarial Neural Networkbased Approach for Unbiased Class Visualization”, DISA Conference (2023).

VALIDATING THE INTERNAL REPRESENTATIONS

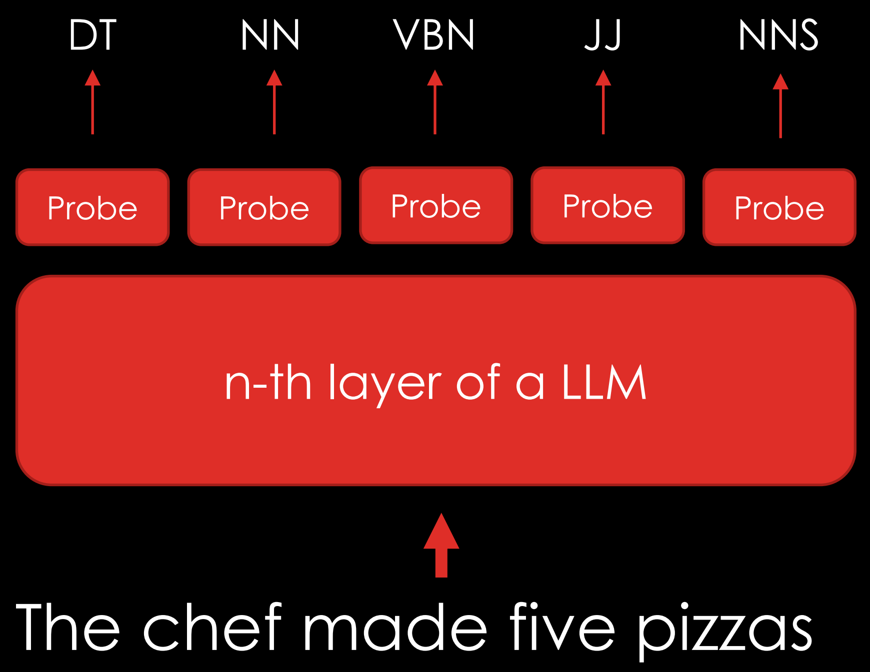
- Internal representation of an example class is the more valid, the more a classifier can predict it's class.
- So we are probing the hidden activations in different layers for a specific set of input samples





Source: Alain, Guillaume, and Yoshua Bengio. "Understanding intermediate layers using linear classifier probes." arXiv preprint arXiv:1610.01644 (2016).

INTERNAL REPRESENTATIONS OF MASKED LANGUAGE MODELS



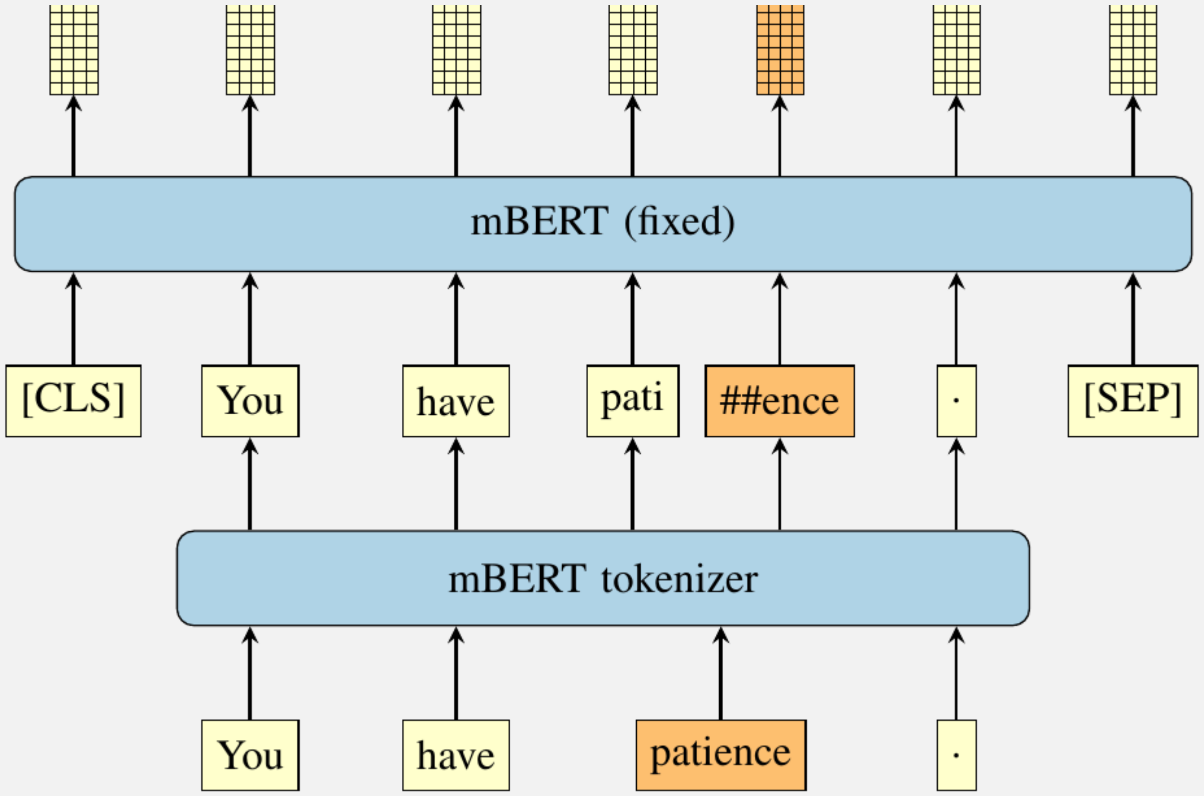
- The input space of the LLMs is rather sparse → Optimization of Maximally Exciting Inputs (MEIs) is extremely hard. Even if possible, it would need too much regularization
- No smooth transition between two discrete words
- Usually, the goal is to predict lower-level features from the representations learnt unsipervisedly.

PREVIOUSLY ON

…

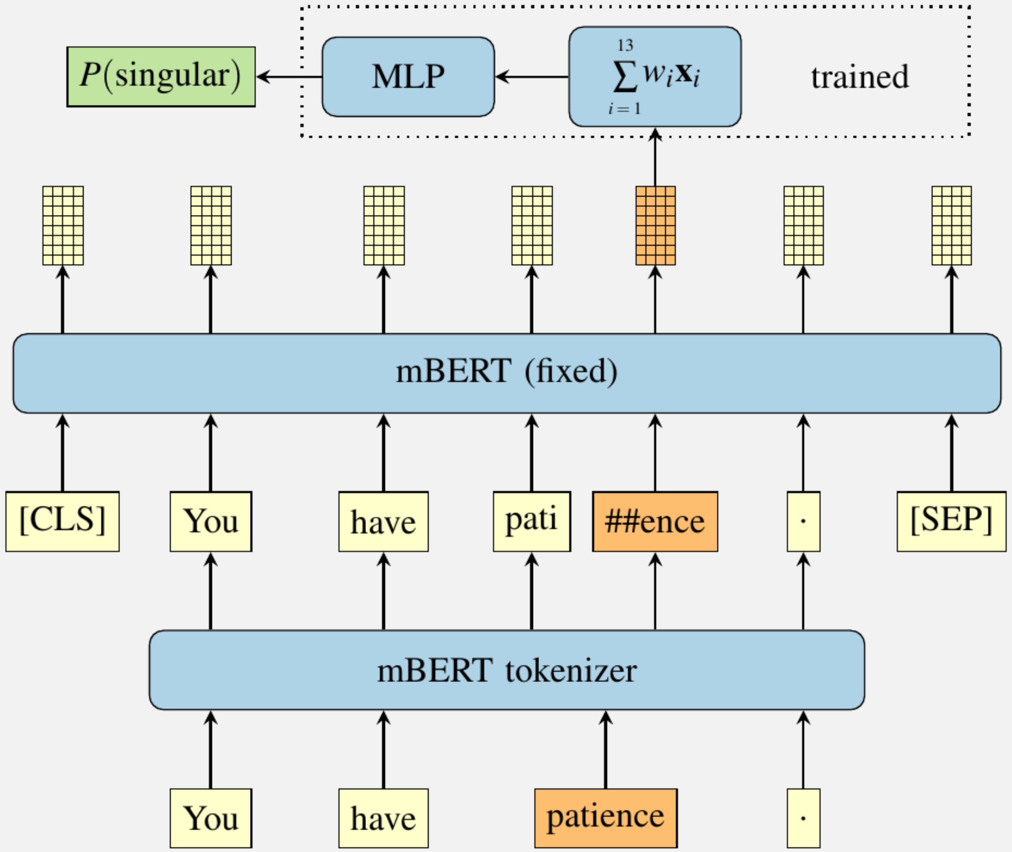
EVALUATING THE INTERNAL REPRESENTATIONS - PROBING

- The embeddings in question are n-times 768d vectors
- Post-hoc analysis of these embeddings is hard
- Training a diagnostic classifier to evaluate the embeddings (Köhn, 2015)
- The accuracy of the diagnostic classifier will be indicative of the degree to which a chosen feature is encoded in the embedding



METHODS

PROBE TRAINING PROCEDURE

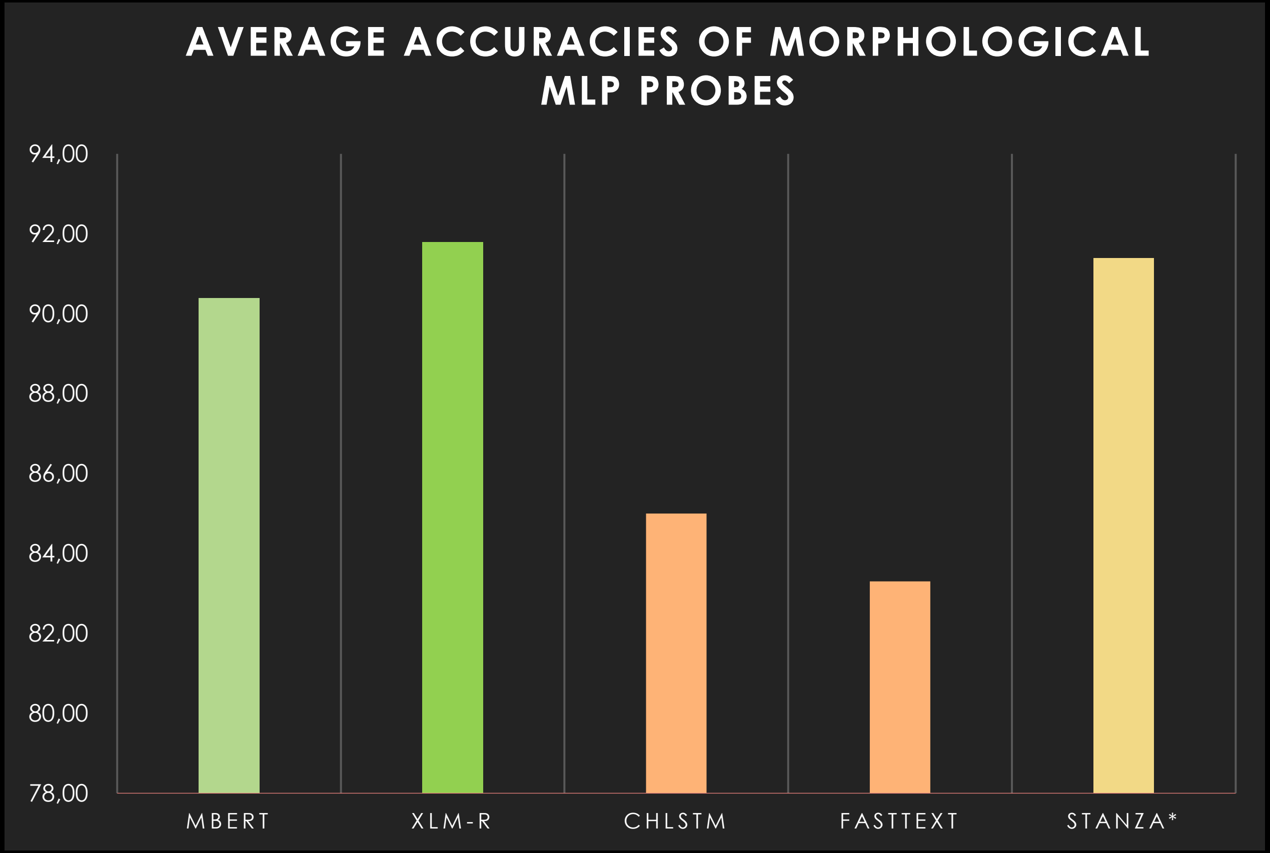


- We train a diagnostic classifier for each task separately
- MLP with a single hidden layer with 50 neurons
- Trained using Adam optimizer (Kingma and Ba 2015)
- With α = 0.001, β1 = 0.9, β2 = 0.999
- Early stopping based on development loss and accuracy.
- Implemented a 20% dropout between both the input and hidden layer of the MLP and between the hidden and the output layers.
- batch size 128
- Accuracies averaged over 10 runs

SELECTED RESULTS OF THE PREVIOUS WORK

RECAP 1: AVERAGE PROBING ACCURACIES

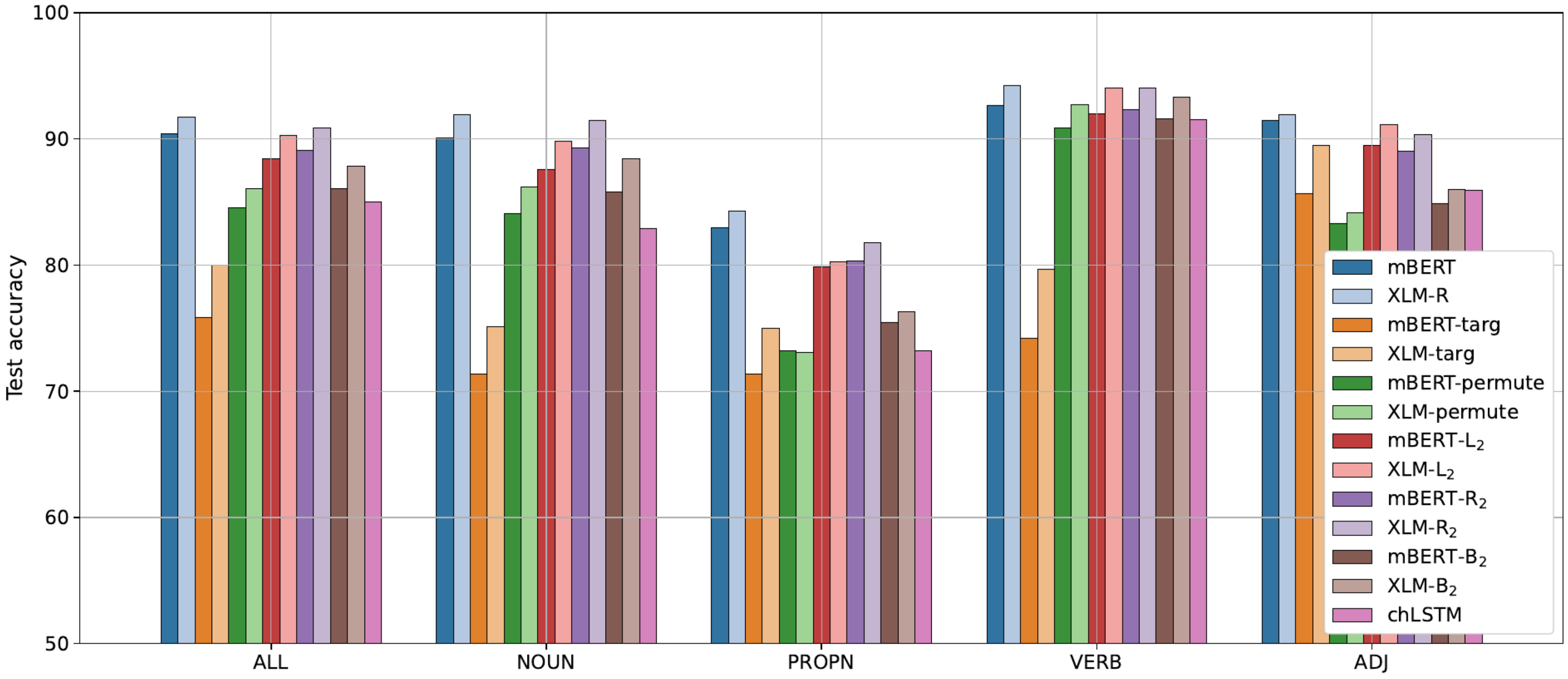
- Both evaluated models (mBERT and XLM-R) perform on the level of the „skyline" morphological tagger (STANZA)
- Both mBERT and XLM-R outperform the two baselines (chLSTM and fastText)
- XLM-R outperforms mBERT in most of the cases; by 2% on average



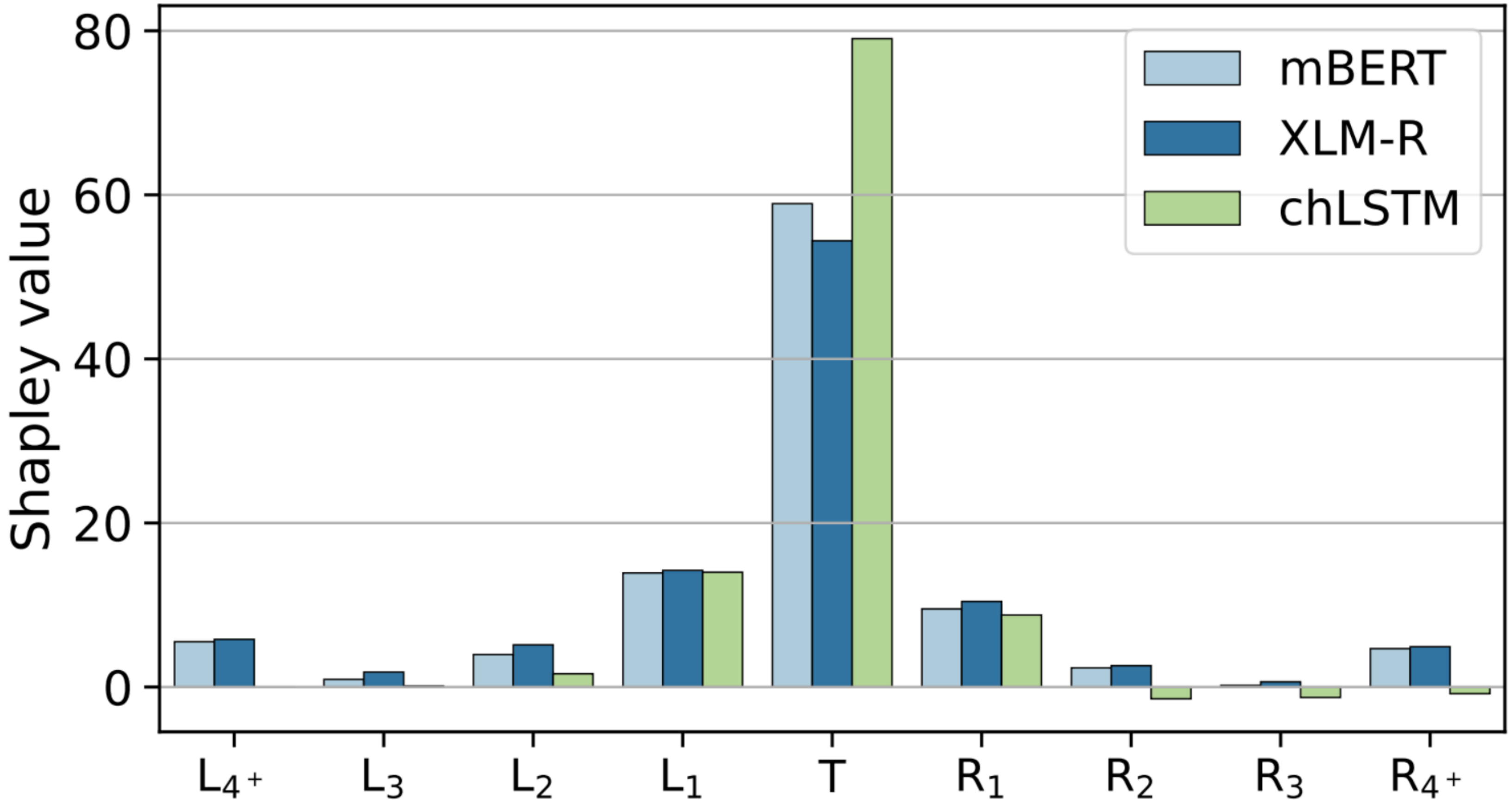
PERTURBATIONS IN THE INPUT SPACE – PROBING CONTROLS

Method	Explanation	Example
Original		Then he ripped open Hermione 's letter and **read** it out loud .
TARG	mask target word	Then he ripped open Hermione 's letter and **[M]** it out loud .
L_2	mask previous 2 words	Then he ripped open Hermione 's [M] [M] **read** it out loud .
R_2	mask next 2 words	Then he ripped open Hermione 's letter and **read** [M] [M] loud .
B_2	mask 2 on each side	Then he ripped open Hermione 's [M] [M] **read** [M] [M] loud .
PERMUTE	shuffle word order	and open **read** Then letter . it out he ripped 's Hermione loud

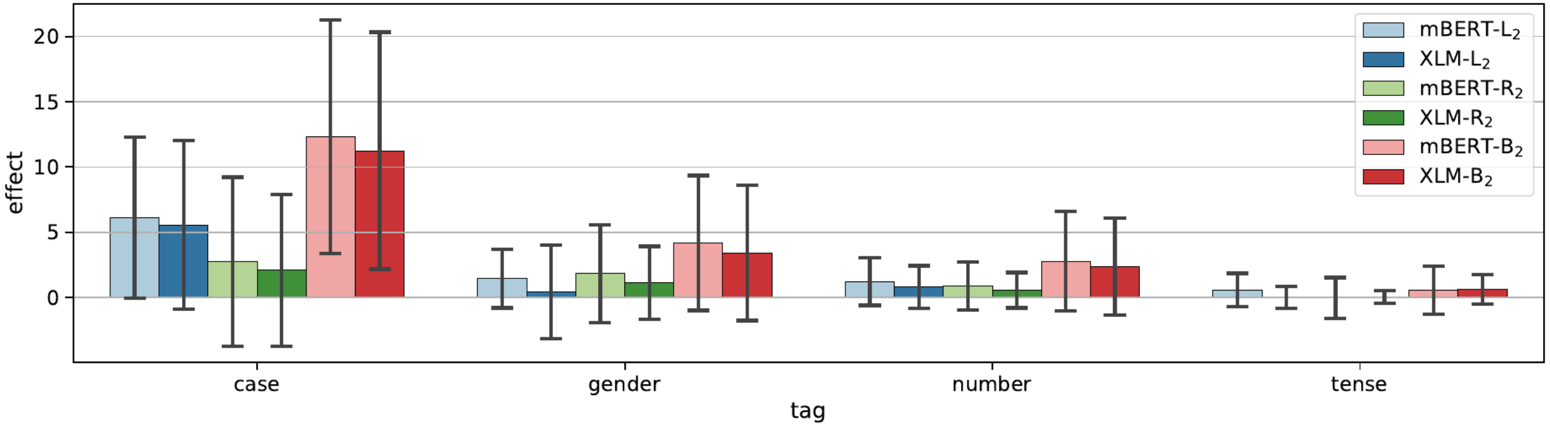
RECAP 2: PERTURBATION – THE BIG PICTURE

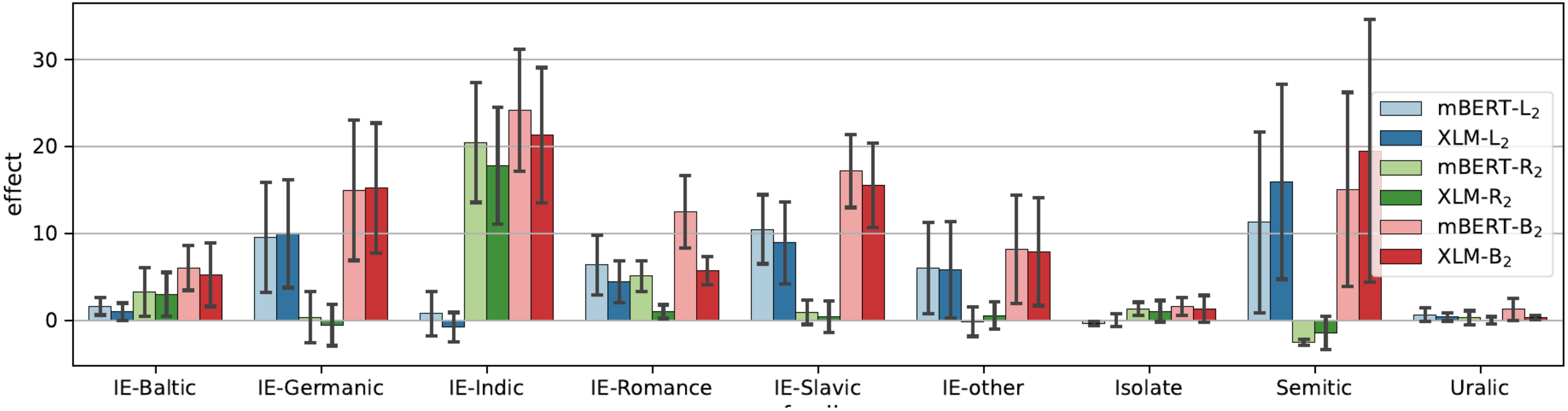


RECAP 3: SHAPLEY VALUES



$$E(m,t,p) = 1 - \frac{\mathrm{Acc}(m,t,p)}{\mathrm{Acc}(m,t)}$$





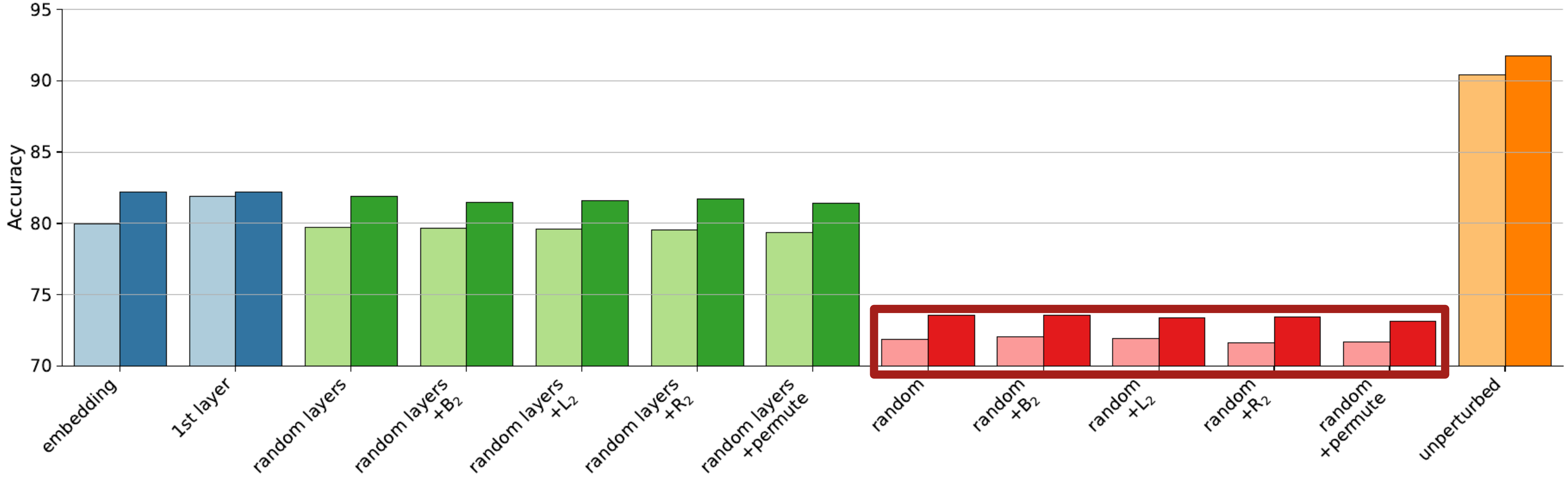
RANDOM BERTS

- MLMs with such a big parameter size can easily learn downstream NLP tasks (Kovaleva et al., 2019).

Questions:

1. What does random BERTs rely on?
2. How does randomizing the token embeddings affect the probing accuracy?
3. How does randomizing all the layers affect the probing accuracy?
4. How does perturbations affect the probing of random BERTs?

RESULTS 3: RANDOM MODELS



RANDOM MODELS SUMMARY

- The accuracies of random models' (with a pretrained embedding layer) morphological probes match the accuracies of their embedding layers' probe
  - i.e., even the Transformer-based **random MLMs rely mostly on the word-identities represented by their embeddings**
  - Randomly initialized language models are **capable of tasks requiring information about word identities only**
- Probing perturbed and unperturbed representations of random MLMs does not make a big difference. Thus, word identities are the most significant factor; the order of words is almost irrelevant

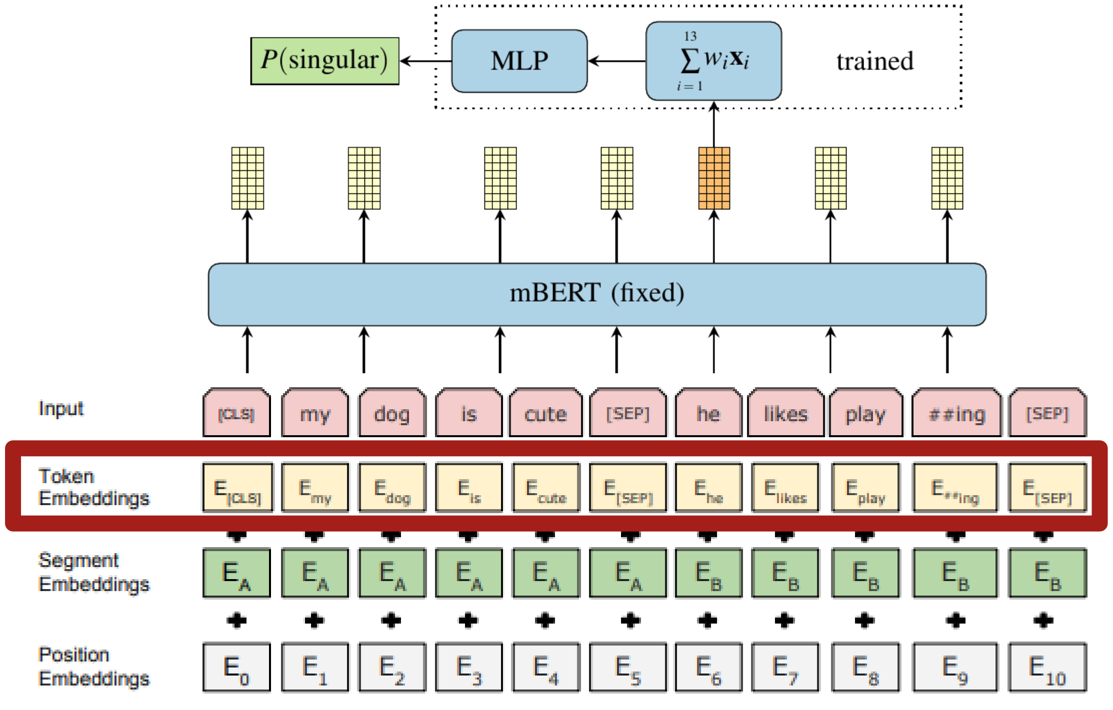


Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

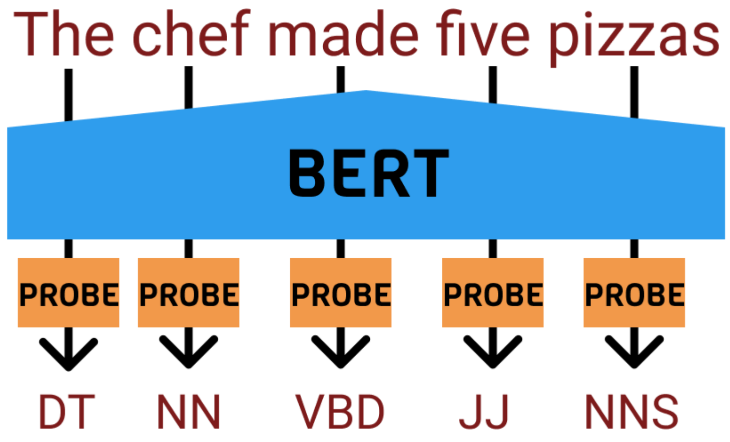
MIXED-PROBING EXPERIMENTS

ONGOING WORK

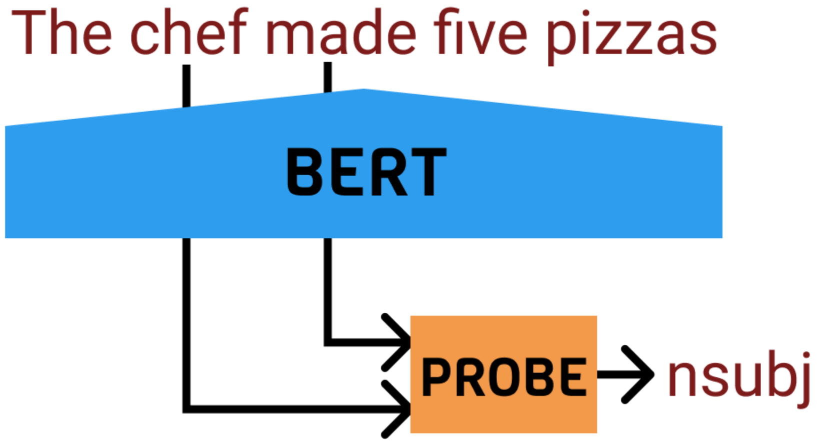
- Evaluating the „most important" context via Dependency trees (Universal Dependencies)
- Validating the usage of the representations found by probing (addressing Belinkov's (2022) critique)

EDGE-PROBING

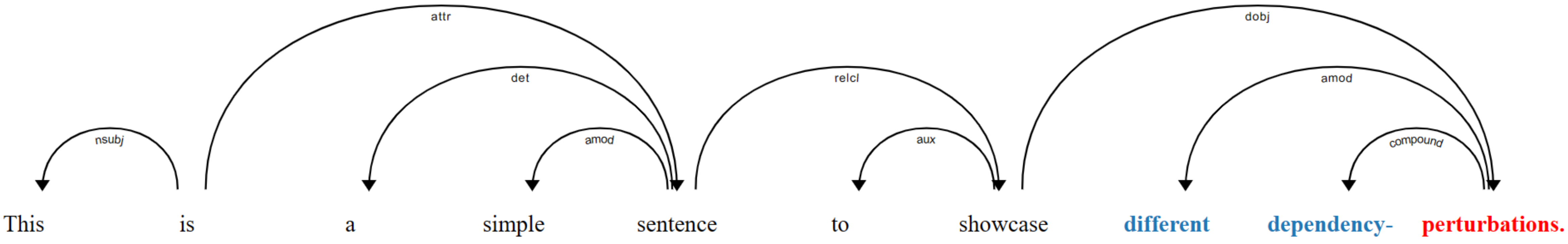
Part-of-speech!



Partial dependency info!



DEPENDENCY TREE

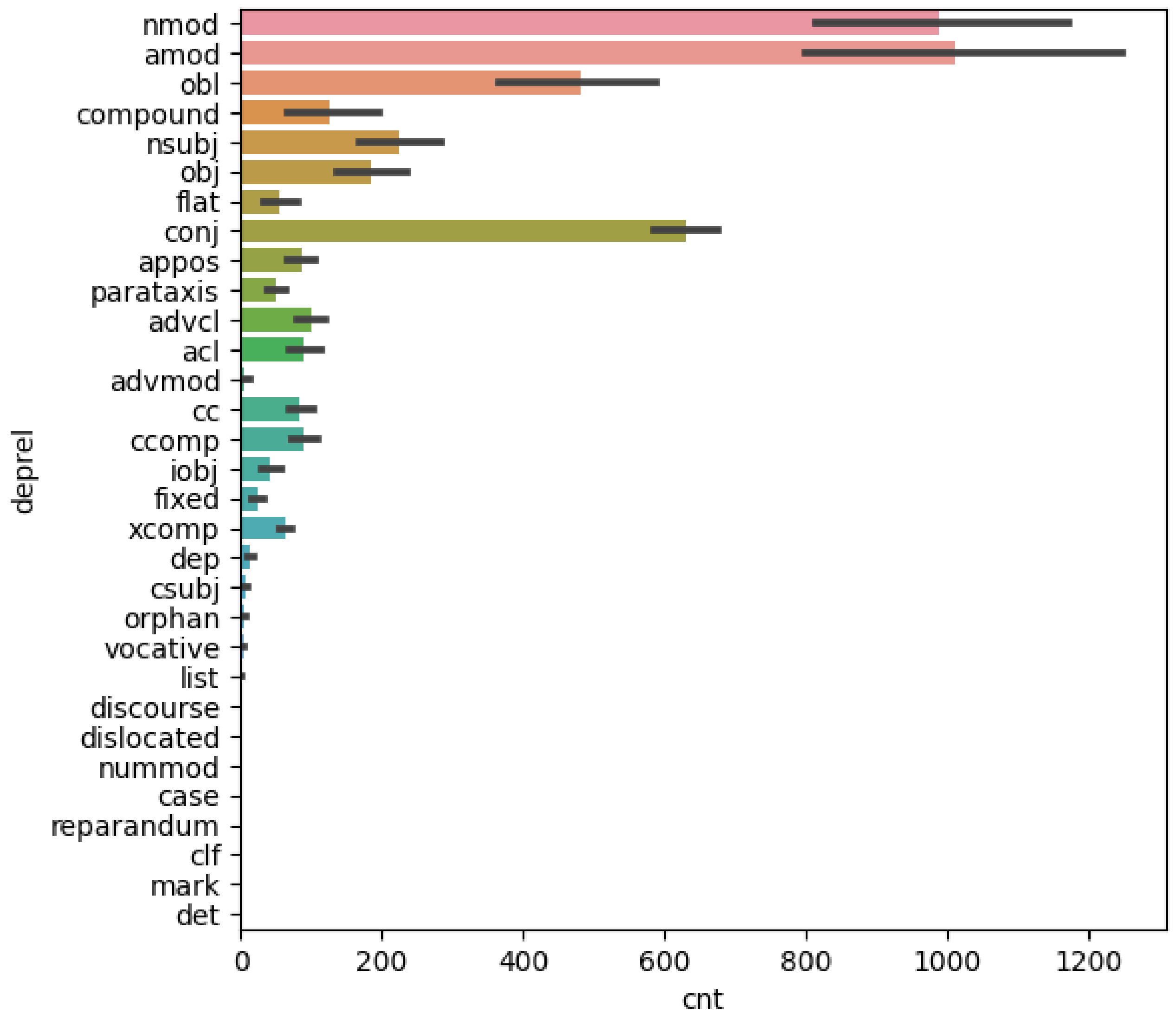


Method	Explanation	Example
Unperturbed		This is a simple sentence to **showcase** different dependency-perturbations.
DEP$_1$	Masks a randomly selected child of the target in the dependency tree.	This is a simple sentence to **showcase** different dependency- [M].
RAND$_1$	Masks a random word, excluding the target.	This is a [M] sentence to **showcase** different dependency-perturbations.
DEP$_R$	Masks a randomly selected pathway from the target to a leaf node in the dependency tree.	This is a simple sentence to **showcase** [M] dependency- [M].
DEP$_R$ + TARG	Masks out both what DEP$_R$ does, and the target word itself in a specific sentence.	This is a simple sentence to [M] [M] dependency- [M].

DEPTREE TAGS

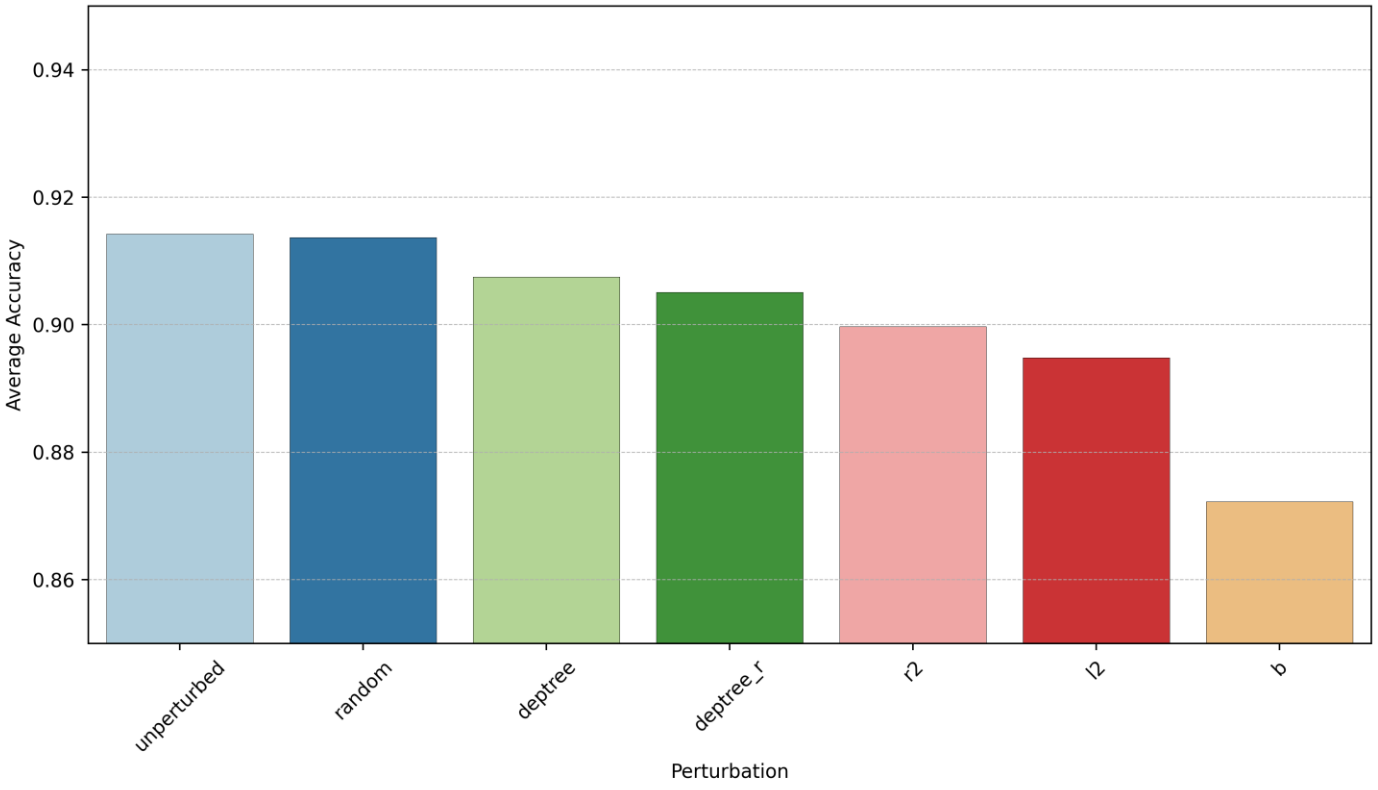
- amod: adjectival modifier
  - **Large house**
- nmod: nominal modifier
  - **Prezident's office**
- conj: conjunct
  - Nice **and** big **house**
- obl: oblique (prev. nmod)
  - They will **arrive** on **Sunday**

- morphological hypotheses based on the Agreement rules

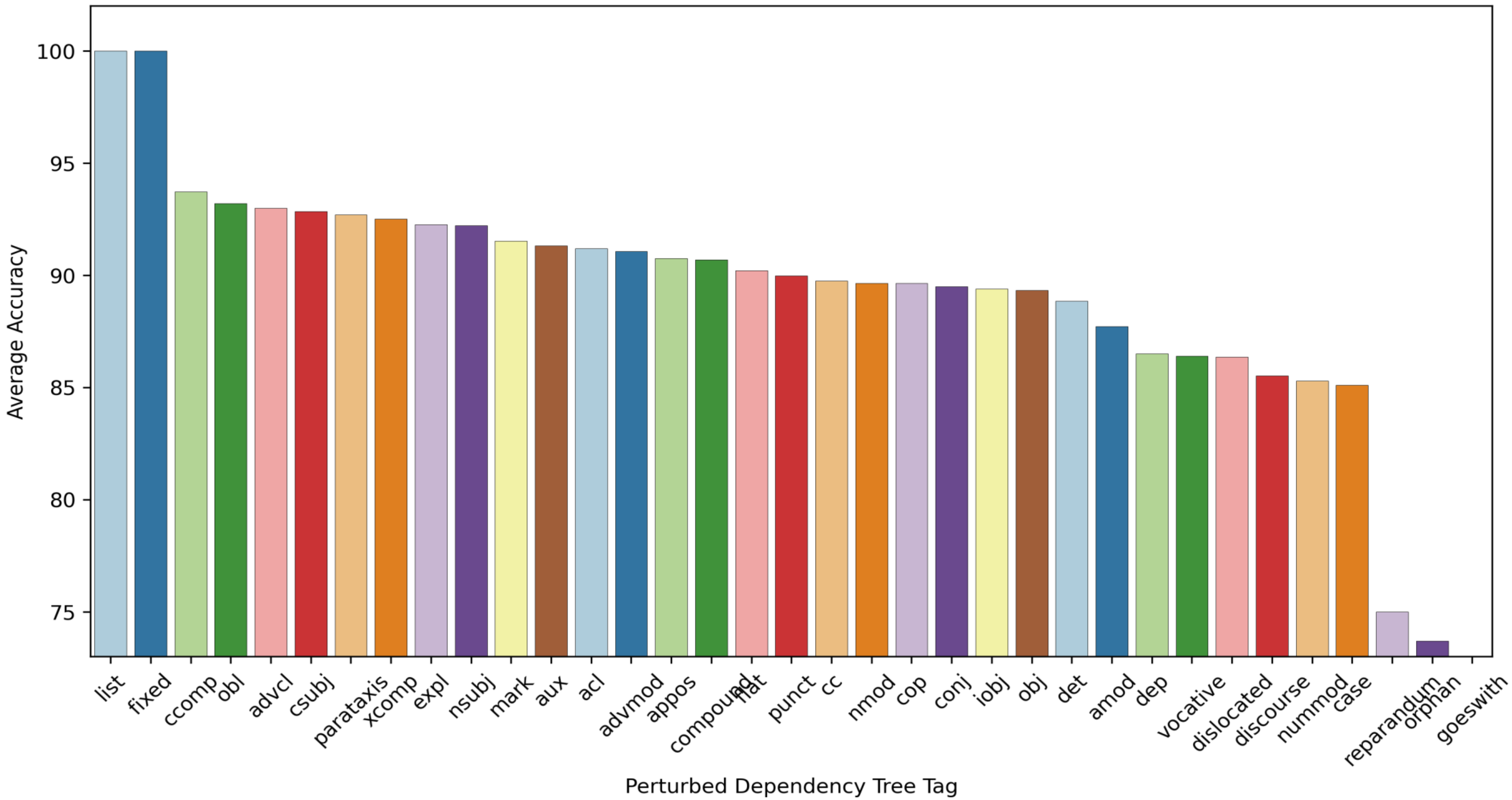


THE BIG PICTURE

- Random perturbations have the smallest effect
- Deptree and Deptree-r perturbations aren't affecting the Accuracy as much as L2, R2 or B2

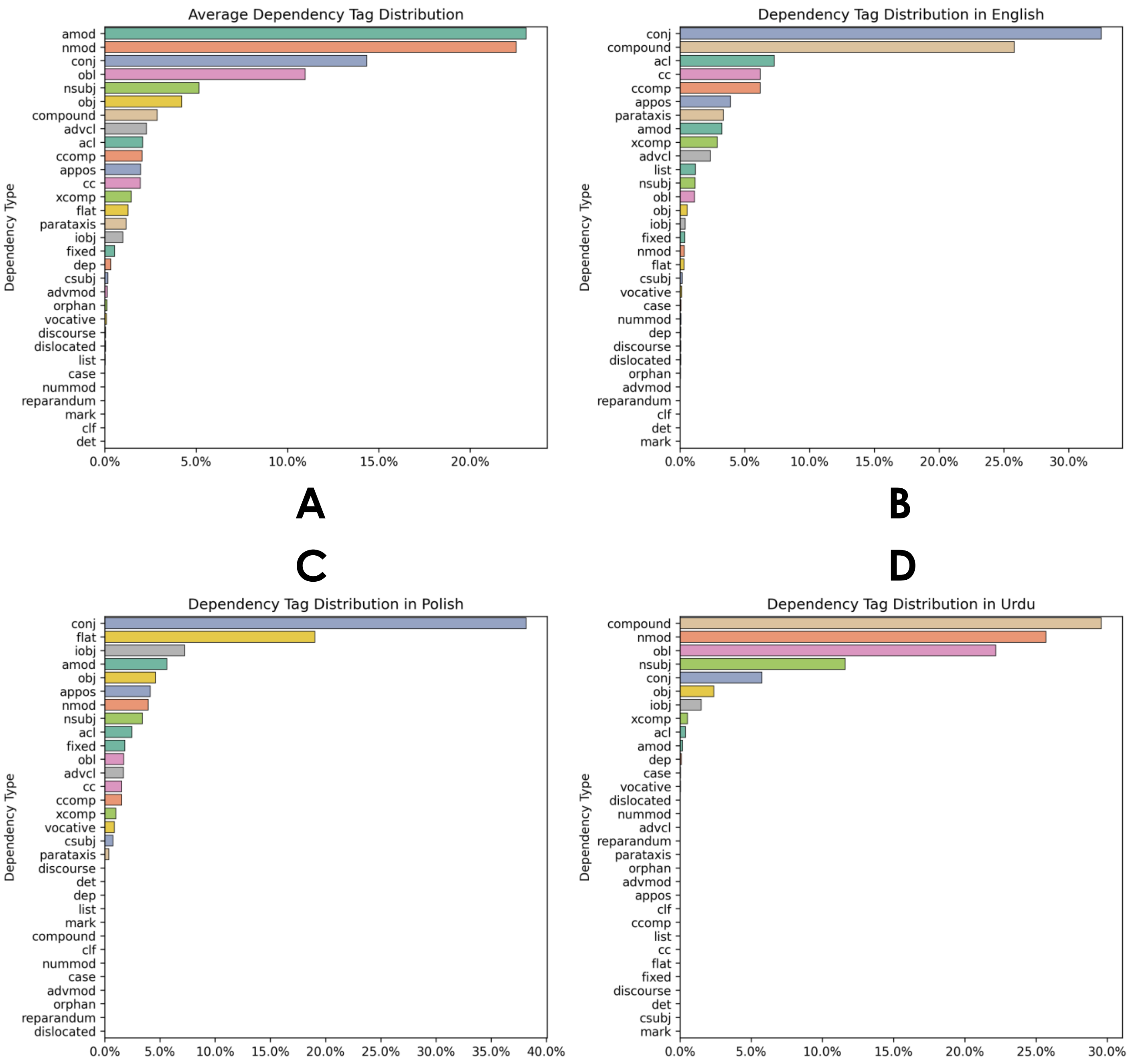


AVERAGE RESULTS BY DT TAG



DEPTREE TAG DISTRIBUTIONS

A: Average
B: English
C: Polish
D: Urdu



WHICH DEPTREE RELATIONS DO AFFECT AGREEMENTS?

- Subject – VERB Agreement <u>in *number and person*</u>
  - E.g., *I am, You are*

- Modifier – NOUN Agreemens* in <u>*gender, number and case*</u>
  - E.g.:

Singular Nominative

- Masculine: **velký pes** (big dog)
- Feminine: **velká kočka** (big cat)
- Neuter: **velké auto** (big car)

Plural Nominative

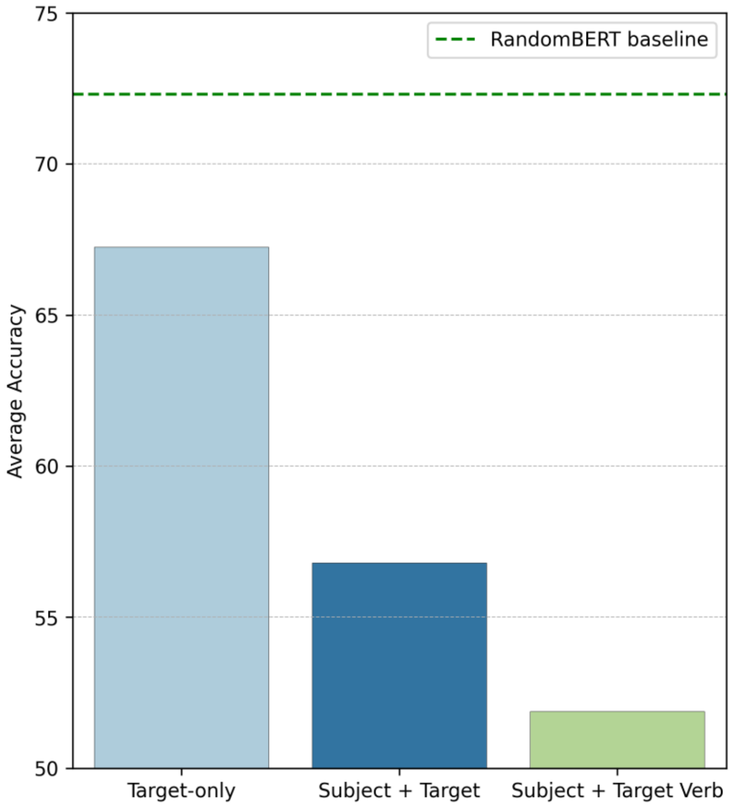
- Masculine: **velcí psi** (big dogs)
- Feminine: **velké kočky** (big cats)
- Neuter: **velká auta** (big cars)

*including possessive structures

AGREEMENT RULE HYPOTHESES TESTING

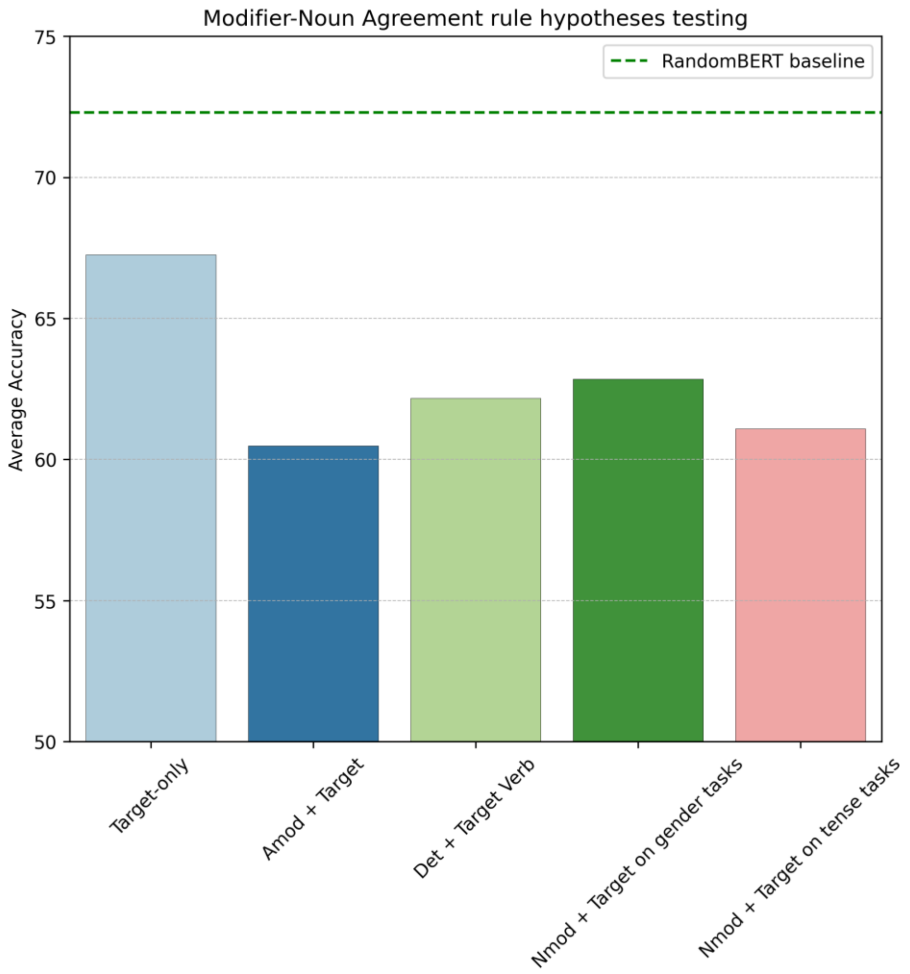
- By definition, agreement rules apply to both the target and its dependent children in the deptree; therefore, both the target and its child node must exhibit the morphological cue relevant to the specific agreement rule. Consequently, agreement rule-based hypotheses were tested by comparing the results of **DEP*R* + TARG** and **TARG** perturbations.

SUBJECT-VERB AGREEMENT



- **[Subject + Target masking]**
  10.46% lower accuracy compared to masking the target only.
- **[Subject + Target verb masking]**
  Performance drop by15%

MODIFIER-NOUN AGREEMENT



- **[Adjectival modifier + Target masking]**
  ~6% lower accuracy compared to masking the target only.
- **[Determinant + Target verb masking]**
  Performance drop by ~5%
- **[Nominal modifier + Target] – Gender tasks only**
  ~5% lower accuracy
- **[Nominal modifier + Target] – Tense tasks only**
  ~6% lower accuracy

PREMISES

- MLMs do not merely rely on the target word's and its neighborhood's representations, but also **selectively integrate contextual cues**, such as dependencies, to enhance morphosyntactic understanding
- Probing can be a precise tool to **investigate the correlation of the internal representation of a MLM and a specific feature double**.
- Future goals:
  - Automatized evaluation of the known agreement rules by mixed-probing
  - **Jointly training an embedding model with a specific diagnostic classifier**

OPEN QUESTIONS

- More data
- Weaker diagnostic classifiers
- What kind of perturbation can increase the accuracy?
- What is the homeostasis of BERT models regarding the <MASK> tokens

THANK YOU FOR YOUR ATTENTION

REFERENCES

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems.
- Kovaleva, O., Romanov, A., Rogers, A., & Rumshisky, A. (2019). Revealing the Dark Secrets of BERT. arXiv preprint arXiv:1908.08593.
- Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644.
- Hamerlik, E., Deb, M., Takáč, M.: Bi-Source Class Visualization: An Adversarial Neural Network-based Approach for Unbiased Class Visualization. In World Symposium on Digital Intelligence for Systems and Machines (DISA2023). 2023.
- Acs, J., Hamerlik, E., Schwartz, R., Smith, N.A., & Kornai, A. (2023). Morphosyntactic probing of multilingual BERT models. Natural Language Engineering. Published online 2023:1-40. doi:10.1017/S1351324923000190