Lexical v. Morphosyntactic Cues to Dependencies

Paula Czarnowska

UNIVERSITY OF CAMBRIDGE

July 15, 2020

Overview

- overview of a methodology for analysing behavior of neural dependency parsers: controlled language alterations

- a few experiments on Polish dependency parsing (with Universal Dependencies)
  still work in progress

Motivation

- the relations between words can be signaled in a variety of ways; primarily through word order and morphological markings, but lexical information can also serve as a cue

Motivation

- the relations between words can be signaled in a variety of ways; primarily through word order and morphological markings, but lexical information can also serve as a cue
- in DELPH-IN the lexical information is only exploited during parse ranking

Motivation

- the relations between words can be signaled in a variety of ways; primarily through word order and morphological markings, but lexical information can also serve as a cue
- in DELPH-IN the lexical information is only exploited during parse ranking
- the neural parsers for UD get word-embeddings as inputs, which encode a mixture of lexical and morphosyntactic information

Motivation

- the relations between words can be signaled in a variety of ways; primarily through word order and morphological markings, but lexical information can also serve as a cue
- in DELPH-IN the lexical information is only exploited during parse ranking
- the neural parsers for UD get word-embeddings as inputs, which encode a mixture of lexical and morphosyntactic information
- To what extent the models exploit those different cues? To what extent they are capable of exploiting them?

Why answering this question is important?

- it could reveal the typological biases present in the models

Why answering this question is important?

- it could reveal the typological biases present in the models
- which cues are used has consequences for the model's robustness and its ability to generalize

Controlled Language Alterations

The idea: Methodologically altering the original language data by stripping it of specific information/cue.

Controlled Language Alterations

The idea: Methodologically altering the original language data by stripping it of specific information/cue.

Could be used both at training and testing: to get insight into the models capacity to exploit different cues.

Controlled Language Alterations

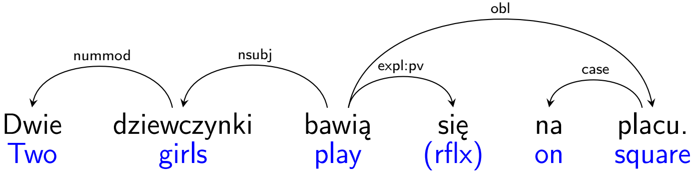
The idea: Methodologically altering the original language data by stripping it of specific information/cue.

Could be used both at training and testing: to get insight into the models capacity to exploit different cues.

Or just at testing: to get insight into what cues the models rely on.

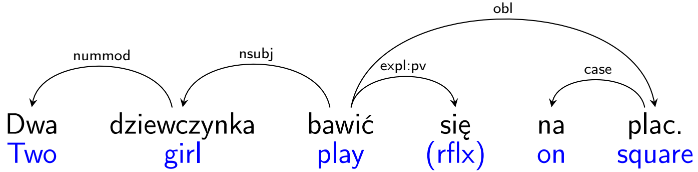
Some Examples

Lemmatisation



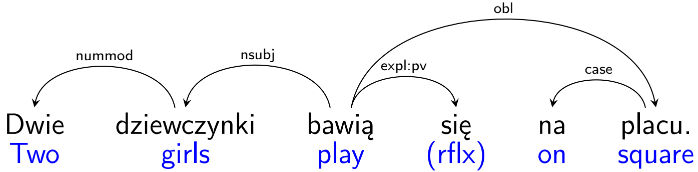
*Results in ungrammatical sentences

Lemmatisation



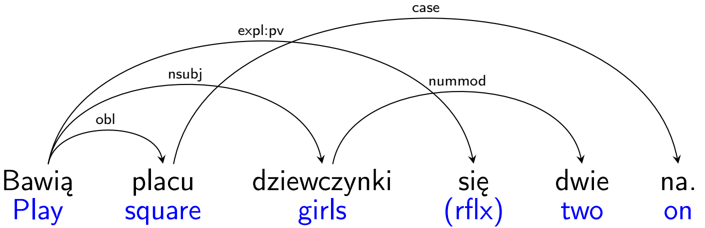
*Results in ungrammatical sentences

Word Order Permutation



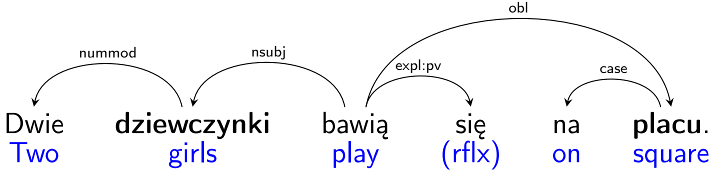
*Results in ungrammatical sentences

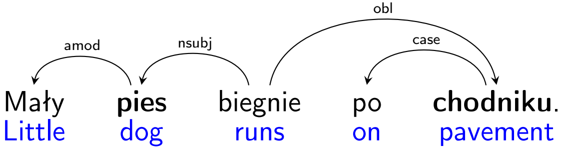
Word Order Permutation



*Results in ungrammatical sentences

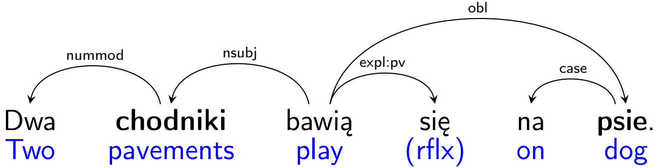
Mixed Noun Lexemes

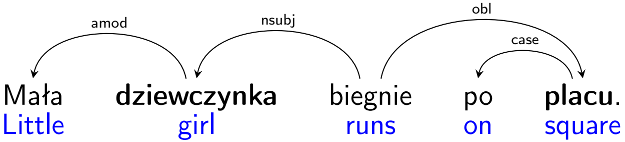




*Results in (mostly) grammatical sentences

Mixed Noun Lexemes





*Results in (mostly) grammatical sentences

Other experiments (not discussed here)

- Removing case marking
- Mixing in very rare/nonce words
- Varying the word order of the core elements in the clause; SVO, SOV etc.
- And more...

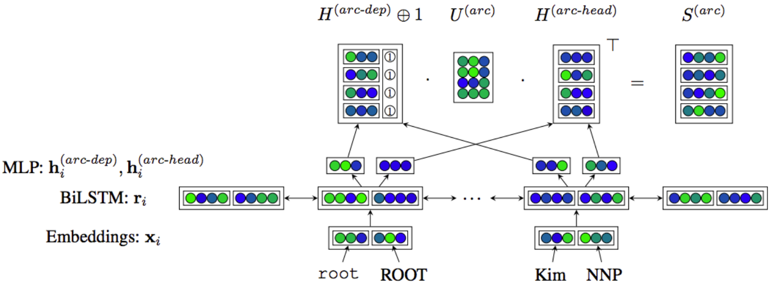
Closely Related Work

- Ravfogel et al. (2019) create synthetic versions of English, by changing the typological parameters, and experiment with RNNs on predicting agreement features for verbs.
- Gulordava et al. (2018) substitute content words by random words with matching POS and morphology, and experiment on predicting long-distance number agreement.
- Kasai and Frank (2019) evaluate parsers in the absence of lexical information, by zeroing out word embeddings (they become OOVs).
- Zheng et al. (2020) craft adversarial examples for parsers by replacing few words in an input sentence, while maintaining both syntactic **and semantic** coherence.

Experiments

Experimental Details

Model: (Dozat and Manning, 2017)



Experimental Details

Different inputs:

- fastText (Bojanowski et al., 2017) – an embedding for a word is constructed by summing the embeddings of its n-grams
- CNN over characters (Kim et al., 2016)
- fastText + CNN
- BERT (Devlin et al., 2019)

Experimental Details

Different inputs:

- fastText (Bojanowski et al., 2017) – an embedding for a word is constructed by summing the embeddings of its n-grams
- CNN over characters (Kim et al., 2016)
- fastText + CNN
- BERT (Devlin et al., 2019)

Data:

- The Polish PDB-UD treebank (Wróblewska, 2018)
- 22,152 sentences (dev set 2215 sents with average length of 16 tokens)

Experimental Details

Different inputs:

- fastText (Bojanowski et al., 2017) – an embedding for a word is constructed by summing the embeddings of its n-grams
- CNN over characters (Kim et al., 2016)
- fastText + CNN
- BERT (Devlin et al., 2019)

Data:

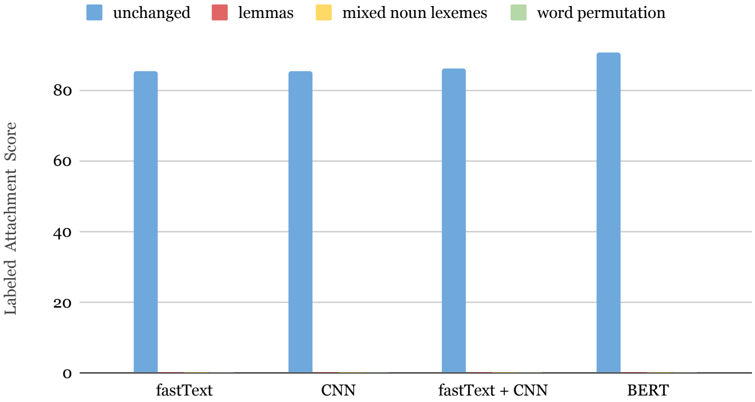
- The Polish PDB-UD treebank (Wróblewska, 2018)
- 22,152 sentences (dev set 2215 sents with average length of 16 tokens)

Metric reported in all graphs: Labeled Attachment Score

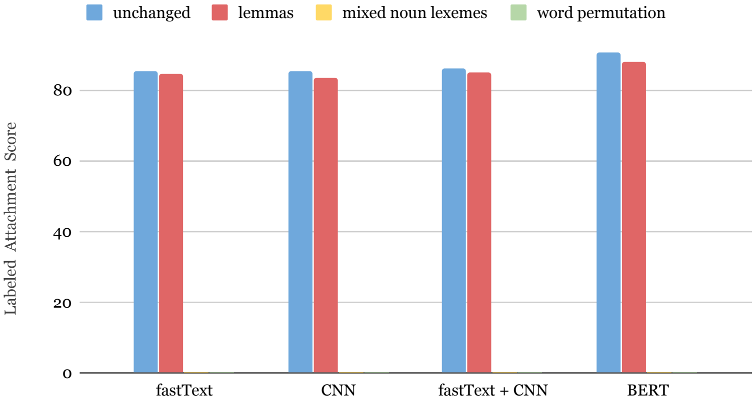
1. An insight into the models' capability to exploit different cues

the models are trained and evaluated on the altered data

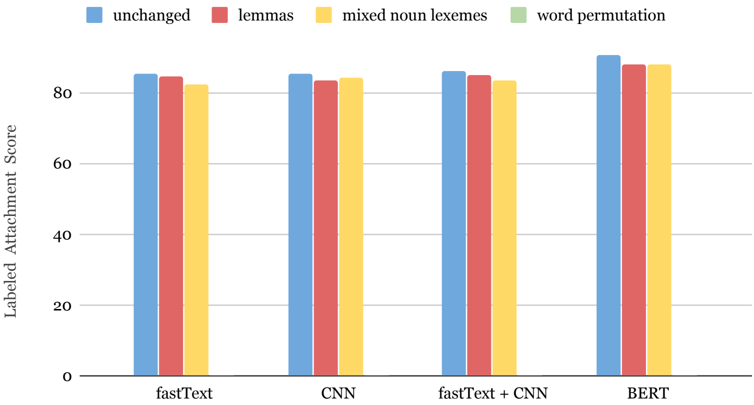
Polish dependency parsing results



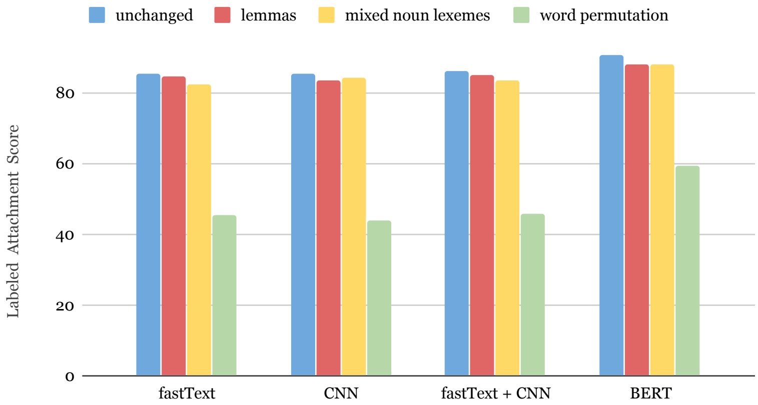
Polish dependency parsing results



Polish dependency parsing results



Polish dependency parsing results



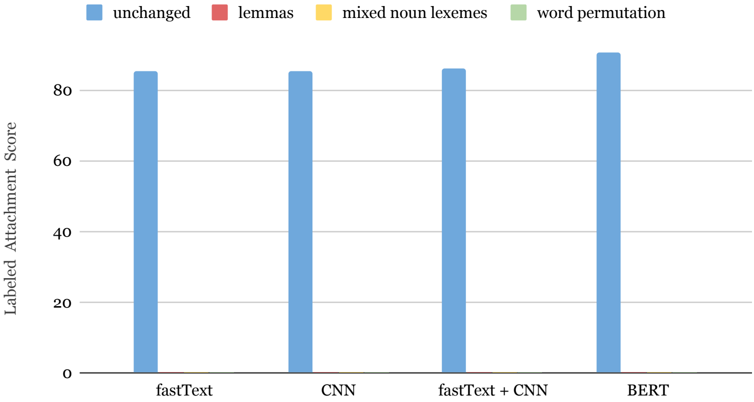
Insights

- the models can exploit different cues to make predictions
- the models do not have to make use of morphological cues to get good performance

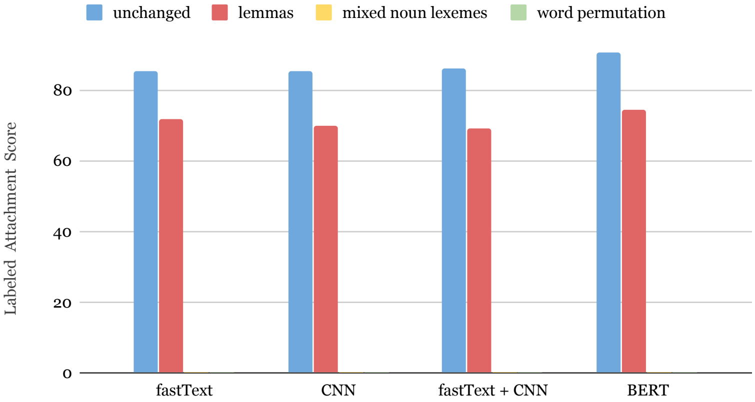
2. An insight into what cues are exploited by models trained on unaltered data.

the models are only evaluated on the altered data

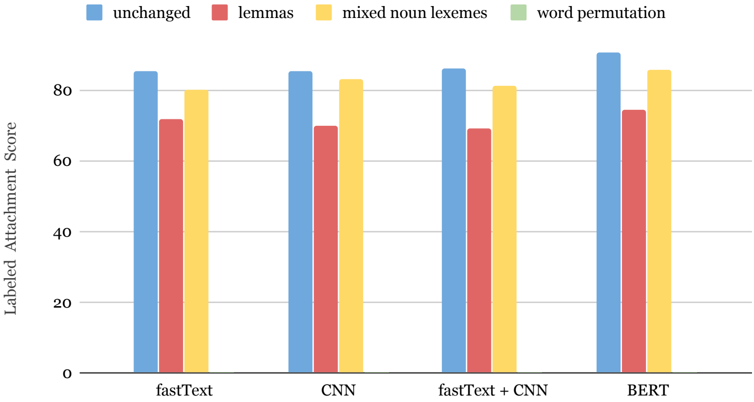
Polish dependency parsing results



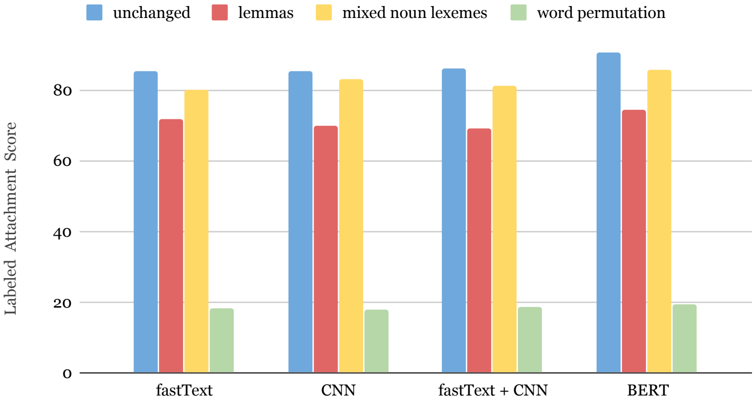
Polish dependency parsing results



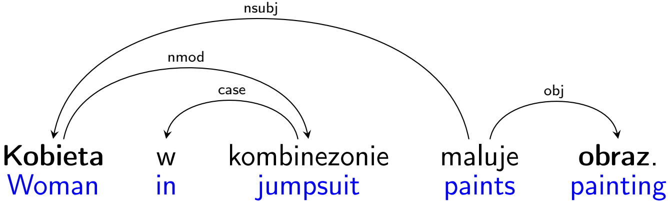
Polish dependency parsing results



Polish dependency parsing results

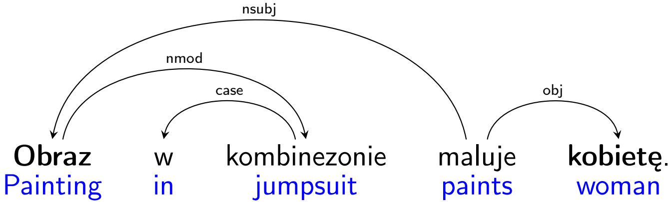


Zooming Into Core Verbal Arguments: Rotation of Core Verbal Arguments



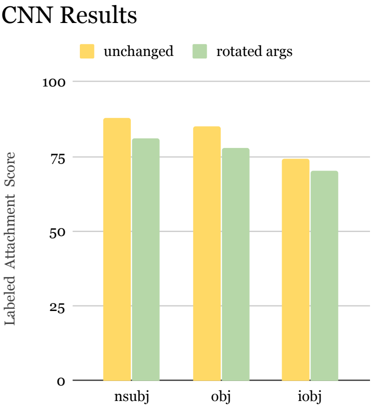
*Results in grammatical sentences

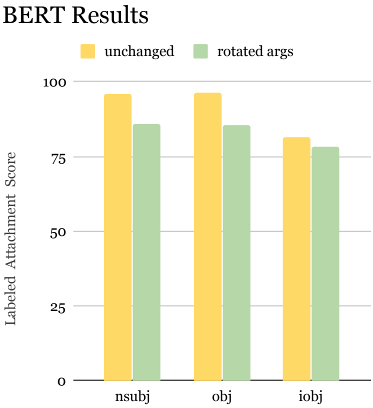
Zooming Into Core Verbal Arguments: Rotation of Core Verbal Arguments



*Results in grammatical sentences

Zooming Into Core Verbal Arguments





*Evaluation on sentences with transitive verbs.

Concluding Remarks and Questions

- the models use a mixture of different cues
- they do rely on morphology and strongly rely on word order
- but at times the lexical signal overpowers the morphosyntactic cues

Concluding Remarks and Questions

- the models use a mixture of different cues
- they do rely on morphology and strongly rely on word order
- but at times the lexical signal overpowers the morphosyntactic cues

Is this *semantic overfitting* a big issue?

Concluding Remarks and Questions

- the models use a mixture of different cues
- they do rely on morphology and strongly rely on word order
- but at times the lexical signal overpowers the morphosyntactic cues

Is this *semantic overfitting* a big issue?

Could this methodology help to further reveal whether the models have a 'preference' for any particular morphosyntactic signal, e.g. rigid word order over flexible word order, adpositions over case markings?

Compared to other interpretability approaches (Belinkov et al., 2020)

Unlike probing (Hewitt and Manning, 2019; Hewitt and Liang, 2019) the goal is not to reveal whether the model's representations capture a specific feature, but to understand how the different parts of the model's 'knowledge' are used to make predictions.

The approach is related to constructing challenge sets (McCoy et al., 2019; Paperno et al., 2016), but quite different – we have multiple altered versions of the original data that can be ungrammatical.

References (1)

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5. Association for Computational Linguistics, 2020.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

References (2)

Timothy Dozat and Christopher D. Manning. Deep biaffine attention for neural dependency parsing. *ICLR*, 2017.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205. Association for Computational Linguistics, 2018.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*, 2019.

References (3)

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

Jungo Kasai and Robert Frank. Jabberwocky parsing: Dependency parsing with lexical noise. *Proceedings of the Society for Computation in Linguistics*, 2(1):113–123, 2019.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. Character-aware neural language models. In *Thirtieth AAAI conference on artificial intelligence*, 2016.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

References (4)

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. Studying the inductive biases of RNNs with synthetic variations of natural languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3532–3542, 2019.

References (5)

Alina Wróblewska. Extended and enhanced polish dependency bank in universal dependencies format. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 173–182. Association for Computational Linguistics, 2018.

Xiaoqing Zheng, Jiehang Zeng, Yi Zhou, Cho-Jui Hsieh, Minhao Cheng, and Xuan-Jing Huang. Evaluating and enhancing the robustness of neural network-based dependency parsing models with adversarial examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6600–6610, 2020.