Explicitly modeling case improves neural dependency parsing

Clara Vania and Adam Lopez
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
c.vania@ed.ac.uk, alopez@inf.ed.ac.uk

Abstract

Neural dependency parsing models that compose word representations from characters can presumably exploit morphosyntax when making attachment decisions. How much do they know about morphology? We investigate how well they handle morphological case, which is important for parsing. Our experiments on Czech, German and Russian suggest that adding explicit morphological case—either oracle or predicted—improves neural dependency parsing, indicating that the learned representations in these models do not fully encode the morphological knowledge that they need, and can still benefit from targeted forms of explicit linguistic modeling.

1 Introduction

Parsing morphologically rich languages (MRLs) is difficult due to the complex relationship of syntax to morphology. But the success of neural networks offer an appealing solution to this problem by computing word representation from characters. Character-level models (Ling et al., 2015; Kim et al., 2016) learn relationship between similar word forms and have shown to be effective for parsing MRLs (Ballesteros et al., 2015; Dozat et al., 2017; Shi et al., 2017; Björkelund et al., 2017). Does that mean that we can do away with explicit modeling of morphology altogether? Consider two challenges in parsing MRLs raised by Tsarfaty et al. (2010, 2013):

- *Can we represent words abstractly so as to reflect shared morphological aspects between them?*
- *Which types of morphological information should we include in the parsing model?*

It is tempting to hypothesize that character-level models effectively solve the first problem. For the second, Tsarfaty et al. (2010) and Seeker and Kuhn (2013) reported that morphological case is beneficial across morphologically rich languages with extensive case systems, where *case syncretism* is pervasive and often hurts parsing performance. But these studies focus on vintage parsers; do neural parsers with character-level representations also solve this second problem?

We attempt to answer this question by asking whether an explicit model of morphological case helps dependency parsing, and our results show that it does. Furthermore, a pipeline model in which we feed predicted case to the parser outperforms multi-task learning in which case prediction is an auxiliary task. These results suggest that neural dependency parsers do not adequately infer this crucial linguistic feature directly from the input text.

2 Dependency Parsing Model

We use a neural graph-based dependency parser similar to that of Kiperwasser and Goldberg (2016) and Zhang et al. (2017) for all our experiments. We treat our parser as a black box and experiment only with the input representations of the parser. Let $w = w_1, \ldots, w_{|w|}$ be an input sentence of length $|w|$ and let w_0 denote an artificial ROOT token. For each input token w_i, we compute the context-independent representation, $\mathbf{e}(w_i)$ with a bidirectional LSTM (bi-LSTM) over characters. We concatenate the result with its part-of-speech (POS) representation, $\mathbf{t}_i$: $\mathbf{x}_i = [\mathbf{e}(w_i); \mathbf{t}_i]$. We then feed $\mathbf{x}_i$ to a word-level bi-LSTM encoder to learn a contextual word representation $\mathbf{w}_i$. The model uses these representations to compute the probability $p(h_i, \ell_i \mid w, i)$ of head $h_i \in \{0, ..., |w|\}/i$ and label ℓ_i of word w_i.

3 Experiments

We experiment with three fusional languages with extensive case systems: Czech, German, and Rus-

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 356–358
Brussels, Belgium, November 1, 2018.

Language	Input	Dev	Test
Czech (68.5K)	word	89.9	89.3
	char	91.2	90.6
	char (multi-task)	91.6	91.0
	char + predicted case	**92.2**	**91.8**
	char + gold case	92.3	91.9
	char + full analysis	92.5	92.0
German (14.1K)	word	86.7	84.5
	char	87.5	84.5
	char (multi-task)	**87.9**	84.4
	char + predicted case	87.8	**86.4**
	char + gold case	90.2	86.9
	char + full analysis	89.7	86.5
Russian (48.8K)	word	89.5	90.1
	char	91.6	92.4
	char (multi-task)	92.2	92.6
	char + predicted case	**92.5**	**93.3**
	char + gold case	92.8	93.5
	char + full analysis	92.6	93.3

Table 1: Label Attachment Score (LAS) results. For each language, we show the number of training sentences.

sian; and we consider four forms of input ($\mathbf{e}(w_i)$, §2): **word** (embedding), **char**acters, characters with **gold** case, and characters with **predicted** case. For the latter two, we append the case label to the character sequence, e.g. ⟨`b, a, t, Acc`⟩ represents *bat* with accusative case. Using the same method, we also supply the gold **full analysis**, to tease out the importance of case specifically. Finally, we experiment with **multi-task** learning (MTL; Søgaard and Goldberg, 2016; Coavoux and Crabbé, 2017), using the bi-LSTM states of the lower layer of the bi-LSTM encoder to predict case feature. Table 1 summarizes the results.

Effect of case We found that the oracle condition of adding gold case improves the parsing performance for all languages, and indeed explains all of the gains of a full morphological analysis. In German, *case syncretism* is pervasive—a single surface form can represent multiple cases—and we see improvement of up to 2.4 LAS points on test set. This results suggest that the character-level models still struggle to disambiguate case when they learn only from the input text.

Language	%case	Dev		Test	
		PL	MT	PL	MT
Czech	66.5	95.4	**96.7**	95.2	**96.6**
German	36.2	**92.6**	92.0	90.8	**91.4**
Russian	55.8	95.8	**96.5**	95.9	**96.5**

Table 2: Case accuracy for case-annotated tokens, for pipeline (**PL**) vs. multitask (**MT**) setup. **%case** shows percentage of training tokens annotated with case.

We then look at the performance when we replace gold case with predicted case. We train a morphological tagger to predict case information. The tagger has the same structure as the parser's encoder, with an additional feedforward neural network with one hidden layer followed by a softmax layer. We found that predicted case improves accuracy, although the effect is different across languages. These results are interesting, since in vintage parsers, predicted case usually harmed accuracy (Tsarfaty et al., 2010). However, we note that our taggers use gold POS, which might help.

Pipeline model vs. Multi-task learning In general, MTL models achieve similar or slightly better performance than the character-only models, suggesting that supplying case in this way is beneficial. However, we found that using predicted case in a pipeline model gives more improvements than MTL. We also observe an interesting pattern in which MTL achieves better tagging accuracy than the pipeline model but lower performance in parsing (Table 2). This is surprising since it suggests that the MTL model must learn to effectively encode case in the model's representation, but must not effectively use it for parsing.

4 Conclusion

Vintage dependency parsers rely on hand-crafted feature engineering to encode morphology. The recent success of character-level models for many NLP tasks motivates us to ask whether their learned representations are powerful enough to completely replace this feature engineering. By empirically testing this using a single feature known to be important—morphological case—we have shown that they are not. Experiments with multi-task learning suggest that although MTL gives better performance, it is still underperformed by a traditional pipeline model.

References

Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved transition-based parsing by modeling characters instead of words with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 349–359, Lisbon, Portugal. Association for Computational Linguistics.

Anders Björkelund, Agnieszka Falenska, Xiang Yu, and Jonas Kuhn. 2017. IMS at the CoNLL 2017 ud shared task: Crfs and perceptrons meet neural networks. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 40–51, Vancouver, Canada. Association for Computational Linguistics.

Maximin Coavoux and Benoit Crabbé. 2017. Multilingual lexicalized constituency parsing with word-level auxiliary tasks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 331–336. Association for Computational Linguistics.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, Vancouver, Canada. Association for Computational Linguistics.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander Rush. 2016. Character-aware neural language models. In *Proceedings of the 2016 Conference on Artificial Intelligence (AAAI)*.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530, Lisbon, Portugal. Association for Computational Linguistics.

Wolfgang Seeker and Jonas Kuhn. 2013. Morphological and syntactic case in statistical dependency parsing. *Computational Linguistics*, 39(1):23–55.

Tianze Shi, Felix G. Wu, Xilun Chen, and Yao Cheng. 2017. Combining global models for parsing universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 31–39, Vancouver, Canada. Association for Computational Linguistics.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235. Association for Computational Linguistics.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Marie Candito, Jennifer Foster, Yannick Versley, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages (spmrl): What, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, SPMRL '10, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Reut Tsarfaty, Djamé Seddah, Sandra Kübler, and Joakim Nivre. 2013. Parsing morphologically rich languages: Introduction to the special issue. *Comput. Linguist.*, 39(1):15–22.

Xingxing Zhang, Jianpeng Cheng, and Mirella Lapata. 2017. Dependency parsing as head selection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 665–676, Valencia, Spain. Association for Computational Linguistics.