Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5563–5568
Marseille, 11–16 May 2020

Reproducing a Morphosyntactic Tagger with a Meta-BiLSTM Model over Context Sensitive Token Encodings

Yung Han Khoe
CLS, Radboud University
Nijmegen, The Netherlands
yhkhoe@protonmail.com

Abstract

Reproducibility is generally regarded as being a requirement for any form of experimental science. Even so, reproduction of research results is only recently beginning to be practiced and acknowledged. In the context of the REPROLANG 2020 shared task, we contribute to this trend by reproducing the work reported on by Bohnet et al. (2018) on morphosyntactic tagging. Their meta-BiLSTM model achieved state-of-the-art results across a wide range of languages. This was done by integrating sentence-level and single-word context through synchronized training by a meta-model. Our reproduction only partially confirms the main results of the paper in terms of outperforming earlier models. The results of our reproductions improve on earlier models on the morphological tagging task, but not on the part-of-speech tagging task. Furthermore, even where we improve on earlier models, we fail to match the F1-scores reported for the meta-BiLSTM model. Because we chose not to contact the original authors for our reproduction study, the uncertainty about the degree of parallelism that was achieved between the original study and our reproduction limits the value of our findings as an assessment of the reliability of the original results. At the same time, however, it underscores the relevance of our reproduction effort in regard to the reproducibility and interpretability of those findings. The discrepancies between our findings and the original results demonstrate that there is room for improvement in many aspects of reporting regarding the reproducibility of the experiments. In addition, we suggest that different reporting choices could improve the interpretability of the results.

Keywords: reproduction, morphosyntactic tagging, meta-bilstm

1. Introduction

Determining the state-of-the-art in language related machine learning is not a trivial task. This issue has been recognized for some time (Armstrong et al., 2009; Lops et al., 2011), and has recently been receiving more attention (Dacrema et al., 2019; Lin, 2019). It can be regarded as a symptom of the underlying problem of a research and publication culture that focuses on beating previous benchmarks while disregarding the need to contribute to scientific knowledge and understanding (Sculley et al., 2018).

Judging whether real progress has been made in any field of machine learning can be difficult because of two main problems. Firsty, the reported results are not always reproducible without investing an unreasonable amount of time and effort. Secondly, the comparisons that are made do not always make sense. This is often because the baseline that is chosen is generally too weak, or because not enough effort is made to optimize the baseline.

An inverse relationship between the popularity of a scientific method or model, and the reliability of the reporting on its results has been argued for (Ioannidis, 2005) and was empirically demonstrated in some fields of science (Pfeiffer and Hoffmann, 2009). The meta-BiLSTM model presented in the work we reproduce can certainly be considered as belonging to a class of models that is at the height of its popularity in the field of machine learning. It is therefore appropriate to evaluate the adequacy of the baselines that were chosen to compare the model with. In this case, we estimate the problem of choosing a weak baseline to be limited. The results of the paper were compared to outcomes of a recent shared task in which multiple teams optimized their own models. These teams obviously have sufficient incentive to choose a properly competitive method and to fine-tune it optimally.

In the present work, we therefore focus our attention on determining the degree to which results reported by Bohnet et al. (2018) on part-of-speech tagging and morphological tagging are reproducible. As requested by the organizers of the REPROLANG 2020 shared task, we obtain the data and tools based on the information provided in the paper. We use these resources to replicate the main experiments of the paper as closely as possible. Theoretically, such a close reproduction of the reported results should be feasible. In practice, however, this is often not trivial, because many unreported details of the training and testing procedure can affect performance (Said and Bellogín, 2014).

In addition to their main experiments, Bohnet et al. (2018) performed a grid search on one data set, visualized in Figure 3 of their paper, which was done to investigate the sensitivity of the meta-BiLSTM model to changes in the network size of the word and character model. Instead of testing a large number of hyperparameter settings on one data set, we test two different network configurations on all data sets included in our reproduction.

It has already been noted by Bohnet et al. (2018) that some data sets on which they report do not contain meaningful part-of-speech tags, and that for some data sets the morphological tagging task is trivial. We exclude these data sets from our study. In this way we aim to improve the interpretability of the results in the context of determining the state-of-the-art.

The non-triviality of closely reproducing results from machine learning experiments is reflected in our results. We fail to match the reported results on either of the two tagging tasks using either of the two model configurations. On the morphological tagging tasks, however, our results still confirm that the meta-BiLSTM model improves on earlier models.

2. Method

For our reproduction, we used only publicly available resources. We did not contact the authors of the original paper. The publicly accessible information on GitHub, however, does include communication between the authors and third parties. This source of information was taken into account in our reproduction effort. A Docker image of our reproduction experiments is released at:
`https://gitlab.com/yhkhoe/rep_meta_tagger`, where the experiments are divided over 8 tags[1]:
v1.2-feats-rep1-gpu
(commit ef7e780e3c4ccba4553abd9484fd4a78f4dff9af),
v1.2-feats-rep2-gpu
(commit ef7e780e3c4ccba4553abd9484fd4a78f4dff9af),
v1.2-feats-rep1-cpu
(commit 2253edbfe469930ed8bb559539eaaf4ea669f2eb),
v1.2-feats-rep2-cpu
(commit 4eacaa2a170de7dede0810e0aaa54fdf14cce9be),
v1.2-xtag-rep1-gpu
(commit 53666f199e5ea42ebc8b004ef37b09fad5e3a58d),
v1.2-xtag-rep2-gpu
(commit e9bafc014ccfea29db2d201ffcd146ea2ae7059b),
v1.2-xtag-rep1-cpu
(commit 114073fb1d5283deb64918f3fb418ef11da97c0e),
and v1.2-xtag-rep2-cpu
(commit efeb98be0be75598e24a65f5a4cff2ad5dfff4d3).
In line with Dacrema et al. (2019) we only consider experimental results reproducible if the available source code requires only minimal modifications to work correctly. Therefore, we do not replicate the ablation study that is reported on in Tables 5 to 8. Replicating these experiments is not possible by setting parameters of the model. It would instead require significant changes to the source code. This would make the reproduction results much less reliable. We hope that the authors will still choose to make their model configurable in a way that makes these experiments easy to replicate. This seems like an improvement that would require a modest effort on their part, compared to the effort it takes for others to implement these modifications.

2.1. Data

We used training/development data from the CoNLL 2017 Shared Task (Zeman et al., 2017a)[2]. As was done in the original paper, we used the versions of the data that already included morphology predicted by UDPipe (Straka and Straková, 2017). This data was already pre-split into training and development sets. The only exception was the Galician TreeGal dataset of which the UDPipe enriched version was not pre-split. We do not know how this was handled by Bohnet et al. (2018), and therefore exclude this data set from our reproduction. This omission only affects results on the part-of-speech tagging task, as the Galician TreeGal data was not used for the morphological features task.
Precomputed word embeddings were used that were provided with the shared task (Ginter et al., 2017)[3]. To evaluate the model, we used test data that was made public after the shared task ended (Nivre et al., 2017) [4].

Some of the languages in the data do not have meaningful XPOS tags. These thirteen language sets are Danish, Spanish, Basque, French, French Sequoia, Hungarian, Croatian, Indonesian, Japanese, Dutch Lassy Small, Norwegian Bokmaal, Norwegian Nynorsk, and Russian SynTagRus, which was noted by Bohnet et al. (2018) in the caption of Table 2 of their document. They included the accuracy scores for these languages in that table, even though they are excluded from the macro-average. We chose to exclude these data sets from our reproduction.

It is mentioned by Bohnet et al. (2018) that the morphological tagging task is trivial for some languages. Nonetheless, they report accuracy scores for these languages and they do not mention excluding them from the macro-average they calculated. They also do not provide a list of these languages. We suggest that excluding the scores on a trivial tagging tasks can improve the interpretability of the results, and we therefore do not include the following data sets in our reporting: Indonesian, Brazilian Portuguese, Chinese, Vietnamese, Japanese, Korean, English LinES, and Swedish LinES.

Including results for these data sets would only distort the overview of what constitutes the state-of-the-art for these tasks, because they can artificially inflate the average scores. Overall, this means that we include 41 out of 55 data sets in our reproduction of the part-of-speech tagging results and 46 out of 54 for the morphological features task.

We also do not reproduce the tagging results on the Penn Treebank. This experiment is not documented to the same degree as the other experiments. It is, for example, not reported in the paper which word embeddings, if any, were used for the experiment. In addition, this data set, although widely used, is not publicly available. We would therefore not have been able to provide a url to this data as is required for submissions to the REPROLANG 2020 shared task.

2.2. Model

The model reported on by Bohnet et al. (2018) uses two separate, sentence-level recurrent networks to learn context sensitive initial character and word-based representations. A meta-level BiLSTM model combines these into a unified representation, which is then used for part-of-speech and morphological tagging. Note that while the character, word and meta-models are trained synchronously, their network architectures, hyperparameters and loss functions are configured individually. Although the authors did not provide a link to the code in their article, we managed to locate code on GitHub[5] that is based on their paper.

[1] We include the tags and commit hashses here to fulfill the requirements of the REPROLANG 2020 shared task

[2] Training/development data was downloaded from:
`http://hdl.handle.net/11234/1-1983`

[3] Word embeddings were downloaded from:
`http://hdl.handle.net/11234/1-1989`

[4] Test data was downloaded from:
`http://hdl.handle.net/11234/1-2184`

[5] The code was downloaded from:
`https://github.com/google/meta_tagger`

lang.	CoNLL	DQM	Meta	Rep1	Rep2
cs_cac	95.16	95.16	**96.91**	*96.58*	96.00
cs	95.86	95.86	97.28	***97.34***	96.88
fi	97.37	97.37	**97.81**	*97.57*	*97.57*
sl	94.74	94.74	**95.54**	94.74	*94.91*
la_ittb	94.79	94.79	**95.56**	95.40	*95.44*
grc	84.47	84.47	**86.51**	84.56	*85.22*
bg	96.71	96.71	**97.05**	96.68	*96.81*
ca	98.58	98.58	**98.72**	*98.58*	98.55
grc_pro	97.51	97.51	**97.72**	97.29	*97.35*
pt	83.04	83.04	**84.39**	*83.86*	83.41
cu	96.20	96.20	**96.49**	95.76	*96.18*
it	97.93	97.93	**98.08**	97.88	*97.94*
fa	97.12	97.12	**97.32**	*97.05*	*97.05*
ru	96.73	96.73	**96.95**	96.64	*96.90*
sv	96.40	96.40	**96.64**	96.31	*96.40*
ko	93.02	93.02	**93.45**	*93.37*	93.27
sk	85.00	85.00	**85.88**	84.83	*85.01*
nl	90.61	90.61	**91.10**	90.79	*90.85*
fi_ftb	95.31	95.31	**95.56**	95.06	*95.32*
de	97.29	97.29	**97.39**	97.14	*97.26*
tr	93.11	93.11	93.43	93.29	***93.53***
hi	97.01	97.01	**97.13**	*97.03*	96.89
es_anc	98.73	98.73	**98.78**	*98.71*	98.69
ro	96.98	96.98	**97.08**	*97.07*	96.96
la_pro	96.93	96.93	**97.00**	96.70	*96.72*
pl	91.97	91.97	**92.12**	91.36	*91.52*
ar	87.66	87.66	**87.82**	*87.72*	87.62
gl	97.50	97.50	**97.53**	97.27	*97.39*
sv_lines	94.84	94.84	**94.90**	94.44	*94.55*
cs_cltt	89.98	89.98	**90.09**	89.05	*89.70*
lv	80.05	80.05	**80.20**	79.16	*79.53*
zh	**88.40**	85.07	85.10	*85.17*	85.02
en_lines	**95.41**	**95.41**	95.39	94.94	*95.02*
ur	**92.30**	**92.30**	92.21	92.29	*92.15*
he	83.24	**82.45**	82.16	*82.20*	82.01
vi	75.42	**73.56**	73.12	*73.26*	73.29
en	94.82	**94.82**	94.66	*94.60*	94.50
en_part	95.08	95.08	94.81	*94.99*	**95.33**
pt_br	**98.22**	**98.22**	98.11	*98.10*	*98.09*
et	**95.05**	**95.05**	94.72	*94.67*	94.89
el	**97.76**	**97.76**	97.53	97.51	*97.53*
macr-av	91.01	90.91	**91.20**	90.88	*90.94*

Table 1: Results for XPOS tags. Column 1 shows the language acronym. Column 2 lists the winning results from the CoNLL 2017 shared task. The column named DQM shows the results of Dozat et al. (2017). The Meta column contains the results reported by Bohnet et al. (2018). The last 2 columns are our reproduction scores, using 3 BiLSTM layers for characters and words, a BiLSTM size of 400 for character, word and meta-models (Rep1) or 2 BiLSTM layers for characters and words, a BiLSTM size of 300 for character, word and meta-models (Rep2). The highest score for each language is in bold. The reproduction (Rep1 or Rep2) closest to the original result (Meta) is in italics.

2.2.1. Model configuration

We tried to replicate the model configuration as closely as possible, using the selection of hyperparameter values listed in Table 1 of the original paper. These values are reported to have resulted from optimization using a limited set of languages. These languages, however, are not listed.

It was slightly confusing that the dropout rates as reported in the paper are configured as their complements, that is embedding keeping probabilities, in the model's configuration file. This naming inconsistency also does not seem to serve any purpose.

In addition, information given on GitHub is somewhat contradictory. The README states that "the settings for the number of LSTM layers, cells, etc. are smaller than the sizes used in the paper". Furthermore, a question concerning the hyperparameters was posted on GitHub by a third party. In response to this issue, the paper's first author answers that the configuration file has been updated with values that should result in similar accuracy scores as reported in the paper. These values, however, are still lower than those reported in the paper. We therefore performed two different reproductions. First we used the configuration reported by Bohnet et al. (2018) in Table 1, which has three BiLSTM layers for characters and words, and uses a BiLSTM size of 400 for the character, word and meta-models. We then also tried the parameter settings from the source code repository, which has two BiLSTM layers for characters and words, and uses a BiLSTM size of 300 for the character, word and meta-models. The two reproductions were configured in the same way in all other respects.

The GitHub README describes the code as being based on the study we aim to reproduce. However, it implements a different training scheme. Instead of training in a single loop over the word, character and meta-models, it loops over the three models consecutively. Unfortunately the training schedule could not be adjusted through the model configuration settings. Because we chose not to contact the authors and to make only minimal code adjustments, we used the code as it was made available.

We configured a maximum value of 1000 early stopping steps to improve the model. In this case we used the value from the configuration file on GitHub, as details about how early stopping was configured were not mentioned in the paper.

Because we had limited access to processing resources, we performed our experiments on different systems with or without a graphics processing unit (GPU). To run the model without the use of a GPU, we used batch sizes of 40000 for words and 80000 for characters that are defined in the configuration file on GitHub. As this caused errors while using a GPU, we used the smaller default batch sizes of 5000 for words and 10000 for characters in that case. These are the values that are hard coded as the defaults in the source code. Because of our processing resource limitations, we were not able to optimize any hyperparameters, such as the learning rate, that are linked to the batch sizes by performing a parameter search. Experiments that were run without a GPU used TensorFlow version 1.14.0. For experiments that were run on a GPU, we used TensorFlow version 1.12.0. This earlier version of TensorFlow was used

because it is the latest release that works with CUDA version 9.0, which is the version of the GPU software we had to work with. All experiments used Python version 2.7.12.

2.3. Evaluation

As was done by Bohnet et al. (2018), we produced F1-scores using the official evaluation script from the CoNLL 2017 shared task (Zeman et al., 2017a). We did, however, need to make some slight adjustments to the script in order for it to properly process the model's output. We verified that our changes did not affect F1-scores for the fine-grained language specific part-of-speech (XPOS) tags and morphological features by recalculating scores on the system outputs (Zeman et al., 2017b) produced by Dozat et al. (2017) for the 2017 CoNLL shared task. Our modified script produced the exact same F1-scores for Czech-CAC (where Bohnet et al. (2018) outperformed earlier models by the widest margin), Arabic (where all models produced similar results), and Greek (where Bohnet et al. (2018) was most outperformed by earlier models). We note that Bohnet et al. (2018) generally, but with some exceptions, label the F1-scores they report as accuracy scores. We suggest that more consistent naming of these types of scores could improve the reproducibility and interpretability of experimental results. It is specifically confusing in this case, because the evaluation script of the CoNLL 2017 shared task produces both F1-scores and accuracy scores[6].

3. Results

3.1. Part-of-speech tagging results

Our results in Table 1 show the part-of-speech tagging scores of the meta-BiLSTM model to be reproducible to a limited degree.

Our comparison of two configurations of the model shows the parameter values from the paper resulting in scores that are closest to the reported F1-scores for 16 out of 41 data sets, with 2 ties. This indicates that the configuration suggested in the source code repository outperforms the configuration in the paper in terms of matching the reported scores. The macro-averages confirm this.

Overall we fail to achieve the same F1-scores as those reported by Bohnet et al. (2018). For the 32 language data sets where they reported higher scores than earlier models, our closest reproduction (Rep2) shows lower scores with only one exception. In 16 of these cases, however, our results are still better than or equal to those reported by the CoNLL 2017 shared task winner (Dozat et al., 2017). For languages where Bohnet et al. (2018) did not improve on earlier models, our closest reproduction outperforms the earlier models on one data set. Overall our most successful reproduction outperforms the earlier models on only 16 out of 41 data sets, with 1 tie. The macro-averages of both of our reproductions are also lower than those for the CoNLL 2017 shared task winners, while only our second reproduction has a slightly higher macro-average than Dozat et al. (2017) across the data sets that we include.

3.2. Morphological tagging results

The results in Table 2 show that our F1-scores are generally lower than those reported by Bohnet et al. (2018). However, our scores are in most cases still higher than those reported for earlier models.

Our comparison of two model configurations shows the parameter values from the paper resulting in F1-scores that are closest to the reported scores for 25 out of 46 data sets, with 1 tie. In contrast with the part-of-speech tagging results, this indicates that the configuration suggested in the paper outperforms the configuration in the source code in terms of matching the F1-scores we aimed to reproduce. Based on the macro-averages, however, our second reproduction performs slightly better in achieving scores that equal the originally reported results. Nonetheless, both reproductions show macro-averages that outperform the earlier models. The higher performance of our second reproduction (Rep2) seems to be largely attributable to a small number of data sets that show highly variable scores across models, such as Greek and Hungarian. We therefore discuss how our first reproduction of the morphological features task (Rep1) compares to the original model results and to results from earlier models.

Our first reproduction fails to achieve the same or better results as the original meta-BiLSTM study with the exception of four data sets. For the 36 data sets where they outperformed earlier models, we still outperform those earlier models in 29 cases, with 1 tie (Spanish). We also outperform earlier models on one data set where the original model did not. Overall, we achieve higher scores than earlier models on 30 out of 46 data sets. This means that our results support the finding that the meta-BiLSTM model improves on earlier results on the morphological features task. The macro-average over the data sets included in our reproduction also confirms this.

4. Conclusion

It seems likely that the parallelism between the original study and our reproduction is limited in different ways, for example in terms of the number of early stopping steps, the ratio between the batch sizes and learning rates, and the training schedule. This in turn limits the value of our findings as an assessment of the reliability of the original results. At the same time, however, it makes our reproduction effort all the more relevant in regard to the reproducibility and interpretability of those findings, based solely on reporting and materials that are publicly available.

The results we report can confirm the main finding of the work of Bohnet et al. (2018) in regard to only one of the tagging tasks. Our reproduction shows that improvement on earlier results in tagging of morphological features, but not in part-of-speech tagging.

The variability between the original and reproduced scores, together with the fact that we outperform all other models on a small number of data sets, support the suggestion by Bohnet et al. (2018) that performance can be increased by

[6]Dozat et al. (2017) also label their scores on part-of-speech tagging and morphological features as accuracy scores. However, they are correctly labeled as F1-scores in the results of the CoNLL 2017 shared task. We verified that these scores are indeed F1-scores by generating scores on the system outputs of the shared task using the official evaluation script.

lang	CoNLL	DQM	Meta	Rep1	Rep2
cs_cac	90.72	94.66	**96.41**	*95.30*	95.19
ru_syn	94.55	96.70	**97.53**	*97.13*	96.88
cs	93.14	96.32	**97.14**	*96.72*	96.44
la_ittb	94.28	96.45	**97.12**	*96.68*	96.66
sl	90.08	95.26	**96.03**	*95.30*	95.20
ca	97.23	97.85	**98.13**	*98.00*	97.88
fi_ftb	93.43	95.96	**96.42**	*96.13*	95.99
no_bok	95.56	96.95	**97.26**	96.91	*96.95*
grc_pro	90.24	91.35	**92.22**	*91.39*	91.23
fr_seq	96.10	96.62	**97.62**	*96.92*	96.90
la_pro	89.22	91.52	**92.35**	*91.80*	91.72
es_anc	97.72	98.15	**98.32**	*98.22*	98.14
da	94.83	96.62	**96.94**	*96.68*	96.60
fi	92.43	94.29	**94.83**	94.47	*94.60*
sv	95.15	96.52	**96.84**	96.60	*96.75*
pt	94.62	95.89	**96.27**	*95.90*	95.79
grc	88.00	90.39	**91.13**	*90.67*	90.39
no_nyn	95.25	96.79	**97.08**	96.66	*96.72*
de	83.11	89.78	**90.70**	*90.00*	89.48
ru	87.27	91.99	**92.69**	92.02	*92.53*
hi	91.03	90.72	91.78	*93.53*	**93.76**
cu	88.90	88.93	**89.82**	88.59	*88.86*
fa	96.34	97.23	**97.45**	97.20	*97.21*
tr	87.03	89.39	**90.21**	89.75	*89.87*
en_part	92.69	93.93	**94.40**	94.20	*94.29*
sk	81.23	87.54	**88.48**	87.17	*87.75*
eu	89.57	92.48	**93.04**	*92.45*	92.36
es	96.34	96.42	**96.68**	96.42	*96.67*
ar	87.15	85.45	**88.29**	*87.69*	87.61
it	97.37	97.72	**97.86**	*97.85*	97.77
nl_lassy	97.55	98.04	98.15	*98.17*	**98.24**
nl	90.04	92.06	**92.47**	92.24	*92.37*
pl	86.53	91.71	**92.14**	*91.85*	91.44
ur	81.03	83.16	**84.02**	*83.77*	83.34
bg	96.47	97.71	**97.82**	*97.64*	97.62
hr	85.82	90.64	**91.50**	*90.81*	90.66
he	**85.06**	79.34	79.76	79.54	*79.66*
et	84.62	88.18	88.25	**88.34**	*88.22*
fr	**96.12**	95.98	95.98	*95.95*	96.05
gl	**99.78**	99.72	99.72	*99.74*	*99.74*
ro	96.24	**97.26**	**97.26**	*97.13*	97.10
cs_cltt	87.88	**90.41**	*90.36*	89.76	*89.94*
lv	84.14	**87.00**	86.92	86.21	*86.70*
el	91.37	94.00	93.92	93.53	***94.15***
hu	72.61	82.67	82.44	80.69	***82.95***
en	94.49	**95.93**	95.71	*95.54*	95.50
macr-av	91.09	93.12	**93.64**	93.25	*93.30*

Table 2: F1-scores for morphological features. Column 1 shows the language acronym. Column 2 lists the winning results from the CoNLL 2017 shared task. The DQM column shows the results of the reimplementation of Dozat et al. (2017) by Bohnet et al. (2018). The Meta column contains the results reported by Bohnet et al. (2018) for their own model. The results in the Rep1 and Rep2 columns are our replicated scores, using 3 BiLSTM layers for characters and words, a BiLSTM size of 400 for character, word and meta-models (Rep1) or 2 BiLSTM layers for characters and words, a BiLSTM size of 300 for character, word and meta-models (Rep2). The highest score for each language is in bold. The reproduction (Rep1 or Rep2) closest to the original result (Meta) is in italics.

means of a grid search on a per model basis. Of course, this is likely also the case for earlier models.

Our reproduction effort reveals room for improvement in two main areas. Firstly, more detailed information on how the model was configured would facilitate reproduction of the results. More consistent naming conventions between the paper and information provided in the source code repository could also contribute to this. Exact detail on how the published code differed from the paper on which it was based would also have been helpful. Secondly, we suggest excluding results that according to the authors are meaningless or constitute a trivial task. Leaving such results out would facilitate interpretation of the results in terms of determining the state-of-the-art in the field. It would also lead to a quantitative decrease in the effort required to reproduce those results.

We conclude that the reproducibility of these type of machine learning results can be improved. This could be done by instituting more extensive requirements on the publication of source code and configuration details, similarly to how the REPROLANG 2020 shared task requires a working source code image of the reported results.

5. Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments and suggestions. We furthermore thank everyone at the Centre for Language and Speech Technology at Radboud University who provided feedback on this paper. Finally, we thank the organizers of the REPROLANG 2020 shared task for their efforts in promoting reproducibility in our field of science.

6. Bibliographical References

Armstrong, T. G., Moffat, A., Webber, W., and Zobel, J. (2009). Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 601–610. ACM.

Bohnet, B., McDonald, R., Simões, G., Andor, D., Pitler, E., and Maynez, J. (2018). Morphosyntactic tagging with a meta-bilstm model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652.

Dacrema, M. F., Cremonesi, P., and Jannach, D. (2019). Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 101–109. ACM.

Dozat, T., Qi, P., and Manning, C. D. (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.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8):e124.

Lin, J. (2019). The neural hype and comparisons against weak baselines. In *ACM SIGIR Forum*, volume 52, pages 40–51. ACM.

Lops, P., De Gemmis, M., and Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer.

Pfeiffer, T. and Hoffmann, R. (2009). Large-scale assessment of the effect of popularity on the reliability of research. *PLoS One*, 4(6):e5996.

Said, A. and Bellogín, A. (2014). Rival: a toolkit to foster reproducibility in recommender system evaluation. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 371–372. ACM.

Sculley, D., Snoek, J., Wiltschko, A., and Rahimi, A. (2018). Winner's curse? on pace, progress, and empirical rigor. In *Proceedings of the 6th International Conference on Learning Representa- tions, Workshop Track (ICLR 2018)*.

7. Language Resource References

Ginter, F., Hajič, J., Luotolahti, J., Straka, M., and Zeman, D. (2017). CoNLL 2017 shared task - automatically annotated raw texts and word embeddings. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Nivre, J., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Bauer, J., Bengoetxea, K., Bhat, R. A., Bick, E., Bosco, C., Bouma, G., Bowman, S., Burchardt, A., Candito, M., Caron, G., Cebiroğlu Eryiğit, G., Celano, G. G. A., Cetin, S., Chalub, F., Choi, J., Cho, Y., Cinková, S., Çöltekin, Ç., Connor, M., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dobrovoljc, K., Dozat, T., Droganova, K., Eli, M., Elkahky, A., Erjavec, T., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Habash, N., Hajič, J., Hajič jr., J., Hà Mỹ, L., Harris, K., Haug, D., Hladká, B., Hlaváčová, J., Hohle, P., Ion, R., Irimia, E., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kanayama, H., Kanerva, J., Kayadelen, T., Kettnerová, V., Kirchner, J., Kotsyba, N., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lando, T., Lê H`ông, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Mendonça, G., Missilä, A., Mititelu, V., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, S., Moskalevskyi, B., Muischnek, K., Mustafina, N., Müürisep, K., Nainwani, P., Nedoluzhko, A., Nguy˜ên Thị, L., Nguy˜ên Thị Minh, H., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Osenova, P., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Pitler, E., Plank, B., Popel, M., Pretkalniņa, L., Prokopidis, P., Puolakainen, T., Pyysalo, S., Rademaker, A., Real, L., Reddy, S., Rehm, G., Rinaldi, L., Rituma, L., Rosa, R., Rovati, D., Saleh, S., Sanguinetti, M., Saulīte, B., Sawanakunanon, Y., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shakurova, L., Shen, M., Shimada, A., Shohibussirri, M., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Stella, A., Strnadová, J., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Tanaka, T., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., van Noord, G., Varga, V., Vincze, V., Washington, J. N., Yu, Z., Žabokrtský, Z., Zeman, D., and Zhu, H. (2017). Universal Dependencies 2.0 - CoNLL 2017 Shared Task Development and Test Data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., et al. (2017a). Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19. Association for Computational Linguistics.

Zeman, D., Potthast, M., Straka, M., Popel, M., Dozat, T., Qi, P., Manning, C., Shi, T., Wu, F. G., Chen, X., Cheng, Y., Björkelund, A., Falenska, A., Yu, X., Kuhn, J., Che, W., Guo, J., Wang, Y., Zheng, B., Zhao, H., Liu, Y., Teng, D., Liu, T., Lim, K., Poibeau, T., Sato, M., Manabe, H., Noji, H., Matsumoto, Y., Kırnap, Ö., Önder, B. F., Yuret, D., Straková, J., Vania, C., Zhang, X., Lopez, A., Heinecke, J., Asadullah, M., Kanerva, J., Luotolahti, J., Ginter, F., Kuan, Y., Sofroniev, P., Schill, E., Hinrichs, E., Nguyen, D. Q., Dras, M., Johnson, M., Qian, X., Liu, Y., Vilares, D., Gómez-Rodríguez, C., Aufrant, L., Wisniewski, G., Yvon, F., Dumitrescu, S. D., Boroş, T., Tufiş, D., Das, A., Zaffar, A., Sarkar, S., Wang, H., Zhao, H., Zhang, Z., Hornby, R., Taylor, C., Park, J., de Lhoneux, M., Shao, Y., Basirat, A., Kiperwasser, E., Stymne, S., Goldberg, Y., Nivre, J., Akkuş, B. K., Azizoglu, H., Cakici, R., Moor, C., Merlo, P., Henderson, J., Wang, H., Ji, T., Wu, Y., Lan, M., de la Clergerie, E., Sagot, B., Seddah, D., More, A., Tsarfaty, R., Kanayama, H., Muraoka, M., Yoshikawa, K., Garcia, M., and Gamallo, P. (2017b). CoNLL 2017 shared task system outputs. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.