Supplementary Material "Explaining Character-Aware Neural Networks for Word-Level Prediction: Do They Discover Linguistic Rules?"

Fréderic Godin, Kris Demuynck, Joni Dambre, Wesley De Neve and Thomas Demeester

IDLab, Ghent University - imec, Ghent, Belgium
firstname.lastname@ugent.be

1 Introduction

This document contains supplementary material for the paper *"Explaining character-aware neural networks for word-level prediction: Do they discover linguistic rules?"*. It consist of some additional information on dataset selection, used morphological classes, per class individual training results and the full statistical analysis associated with Section 5.4 in the paper.

2 Notes on Dataset Selection

In the paper (Silfverberg and Hulden, 2017) that introduced the morphological segmentations for a subset of the Universal Dependencies dataset 1.4, sentences from the training dataset were selected for constructing the test set. Although the paper mentions that all test set sentences for Finnish, Spanish and Swedish were selected from the Universal Dependencies test sets, this is not the case for Finnish. The first 515 lines of *fi-ud-train.conllu* were used for selecting 300 test set words. Given that a new sentence starts at line 521, we removed the first 520 lines from the Finnish training set. This is only 0.3% of the full training set, and consequently, this will have a negligible impact on our conclusions. Note that, for Spanish and Swedish, the segmented words were indeed selected from their respective test sets. The above observations were also confirmed by the first author of the original paper.

3 Overview Morphological Classes Used

In Table 1, Table 2 and Table 3, all class types and corresponding feature class values for Finnish, Spanish and Swedish are listed. During training, each class type has a specific multinomial regression layer which predicts a single value for that class type. However, all class types are jointly trained.

4 Individual Results Morphological Tagging

In Table 4, Table 5 and Table 6, the individual results for each morphological feature class for Finnish, Spanish and Swedish can be found.

5 Full statistical analysis for "Interactions of learned patterns"

From the full UD test set, we selected all words that end with the character *a* and evaluated the morphological feature type gender for all of them. We selected three groups:

- Words that have the label gender=feminine and are classified as gender=feminine, called wf_pf. This group contains 219 words.
- Words that do not have the label gender=feminine are classified as gender=feminine, wnf_pf. This group contains 44 words.
- Words that do not have the label gender=feminine are classified as either gender=NA or gender=masc, i.e. not-feminine, called wnf_pnf. This group contains 199 words.

For each group, we calculated the contributions of all possible character sets of different length within each word and selected the highest contribution score and the lowest contribution score for each word. In other words, we look for the sets of characters that generate the strongest positive and negative contributions for predicting the class gender=feminine. These two contribution scores are the determining factors for certain classification decisions.

5.1 Maximum contribution scores

Based on a Kruskal-Wallis test, a statistically significant difference was found between the three groups, $H(2) = 50,600$, $p < 0.001$. Pairwise comparisons with adjusted p-values showed no significant difference in positive contributions scores between the groups wnf_pf and wnf_pnf ($p = 1.000$). Hence, non-feminine words have similar positive contribution scores, independent of the classification result. Furthermore, significant differences were found between the positive contribution scores of the groups wf_pf and wnf_pf ($p < 0.001$) and the groups wf_pf and wnf_pnf ($p < 0.001$), indicating a difference between the positive contributions of feminine words and non-feminine words.

5.2 Minimum contribution scores

Based on a Kruskal-Wallis test, an overall statistically significant difference was found between the three groups, $H(2) = 36.710$, $p < 0.001$. Pairwise comparisons with adjusted p-values showed that there was no significant difference between the groups wf_pf and wnf_pf ($p = 0.585$), showing that the negative contribution scores of words classified as feminine are similar despite that the fact that the words from wnf_pf are not feminine. A strong significant difference was found between the groups wf_pf and wnf_pnf ($p < 0.001$) and a borderline significant difference between the groups wnf_pnf and wnf_pf ($p < 0.070$). Consequently, there is a clear difference between the negative contributions of non-feminine words that are classified as not-feminine and words that are classified as feminine. Moreover, words that are wrongly classified as feminine have similar negative contribution scores as words classified correctly as feminine.

References

Miikka Silfverberg and Mans Hulden. 2017. Automatic morpheme segmentation and labeling in universal dependencies resources. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*.

Table 1: Overview of classes used for Finnish.

Class type	Values
Number	_NA_ Sing Plur
PartForm	_NA_ Past Pres Agt Neg
Case	_NA_ Ela Ine Ins Par Ill Com Nom All Acc Ade Gen Ess Abl Tra Abe
Person	_NA_ 1 2 3
Derivation	_NA_ Ja Minen Sti Vs Tar Llinen Inen U Ttaa Ttain Lainen Ton
Person[psor]	_NA_ 1 2 3
VerbForm	_NA_ Inf Part Fin
Mood	_NA_ Imp Cnd Pot Ind
Tense	_NA_ Past Pres
Clitic	_NA_ Pa,S Han Ko Pa Han,Pa Han,Ko Ko,S S Kin Kaan Ka
Degree	_NA_ Pos Cmp Sup
Voice	_NA_ Pass Act

Table 2: Overview of classes used for Spanish.

Class type	Values
Person	_NA_ 1 2 3
Mood	_NA_ Imp Ind Sub Cnd
Tense	_NA_ Fut Imp Pres Past
Gender	_NA_ Fem Masc
VerbForm	_NA_ Inf Ger Part Fin
Number	_NA_ Sing Plur

Table 3: Overview of classes used for Swedish.

Class type	Values
Gender	_NA_ Neut Masc Fem Com
Degree	_NA_ Sup Cmp Pos
Number	_NA_ Sing Plur
Case	_NA_ Gen Nom Acc
Poss	_NA_ Yes
Voice	_NA_ Act Pass
Tense	_NA_ Pres Past
Definite	_NA_ Ind Def
VerbForm	_NA_ Sup Part Inf Fin Stem

Table 4: Per class accuracy on the Finnish test set.

	Number	Partform	Case	Person	Derivation	Person[psor]
Maj. Vote	64.42%	94.33%	28.49%	89.17%	98.43%	96.05%
CNN	89.40%	96.97%	87.00%	95.81%	99.07%	98.49%
BiLSTM	89.67%	97.86%	87.89%	95.77%	99.11%	99.29%

	Verbform	Mood	Tense	Clitic	Degree	Voice
Maj. Vote	77.54%	87.77%	89.09%	98.49%	84.16%	78.49%
CNN	93.05%	95.90%	96.17%	99.51%	92.70%	93.59%
BiLSTM	93.19%	96.13%	95.99%	99.51%	92.97%	94.12%

Table 5: Per class accuracy on the Spanish test set.

	Person	Mood	Tense	Gender	Verbform	Number
Maj. Vote	85.26%	87.62%	85.99%	54.40%	75.49%	45.56%
CNN	91.84%	93.51%	91.11%	84.62%	88.08%	84.41%
BiLSTM	91.95%	93.41%	90.90%	84.31%	89.02%	86.40%

Table 6: Per class accuracy on the Swedish test set.

	Gender	Degree	Number	Case	Poss	Voice	Tense	Definite	Verbform
Maj. Vote	46.64%	84.57%	42.21%	62.73%	99.67%	83.99%	87.57%	41.54%	79.19%
CNN	86.18%	93.78%	79.45%	87.79%	99.94%	94.60%	94.29%	83.83%	90.98%
BiLSTM	83.97%	94.26%	78.72%	86.04%	99.97%	93.75%	93.84%	83.86%	90.64%