Explaining A Deep Learning Based Model for Spelling Correction of "de/da" Clitics in Turkish through Understanding Their Dependence to Various Regions of A Sentence

Alperen Değirmenci – Hasan Öztürk

Advisors: Suzan Üsküdarlı – Onur Güngör

Department of Computer Engineering, Boğaziçi University

Introduction and Motivation

- A deep learning-based spellchecker is built to detect the spelling errors of Turkish “de/da” clitics with a SOTA score of **88.26%** F1-Score on this task.
- The spellchecker is treated as a black-box which is used to spellcheck perturbed sentences to provide an explanation in terms of the most significant features related to the prediction.
- A Game with a Purpose (GWAP) type web application called **"-de/-da Takıntısı"** was developed to gather information from Turkish speakers regarding how they predict the correct spelling of perturbed sentences containing "de/da". These results are being analyzed and compared with the findings of the model explanation.

Perturbation Based Explanation Approach

- There has been efforts to build explanation techniques for black-box NLP models [1]. In this study, a model-agnostic explanation approach is proposed to interpret sequence labeling models.
- Sentences are perturbed with a certain strategy and their effect on the model is observed.

Tenis topu evde kaldı. ⟶ Tenis evde kaldı.

Perturbation example 1: Remove the two previous word.

Arabayı ben de sürdüm. ⟶ Arabayı ben de __düm.

Perturbation example 2: Remove a common prefix in the following word.

Algorithm 1 Perturbation based explanation algorithm

Input: S ▷ Set of sentences which contain “de/da”
$Perturbations$ ▷ Set of perturbation functions
$spellchecker$ ▷ “-de/-da” Spellchecker function
Output: $Perturbation\_Impact$ ▷ set of $\overline{\Delta P}$ for each perturbation
for $pr \in Perturbations$ **do**
  for $s \in S$ **do**
    $w_d \leftarrow get\_deda\_word(s)$
    $\langle w_d, p_0 \rangle \leftarrow spellchecker(s, w_d)$
    $s_{pr} \leftarrow pr(s)$
    $\langle w_d, p_{pr} \rangle \leftarrow spellchecker(s_{pr}, w_d)$
    $\Delta P_{pr} = p_{pr} - p_0$
    $\Delta P_{pr_c} \leftarrow \Delta P_{pr_c} + \Delta P_{pr}$
  end for
  $\overline{\Delta P_{pr}} \leftarrow \frac{\Delta P_{pr_c}}{|S|}$
  $Perturbation\_Impact[i].append(\langle pr, \overline{\Delta_{Ppr}} \rangle)$
end for
return $Perturbation\_Impact$

Perturbation based explanation algorithm

w_d *is a word that contains a "de/da" and* ***spellchecker*** *is the model trained to check the spelling of "de/da" words and it returns the labeling probability of a given word in a sentence.*

- The average Δ_{P_i}'s for each perturbation strategy indicates the effect of each perturbation to the prediction of the model. If the Δ_{P_i} for a particular strategy is relatively high, then this means that that perturbation is important for determining the "de/da" usage in a sentence.

Results of the Perturbation

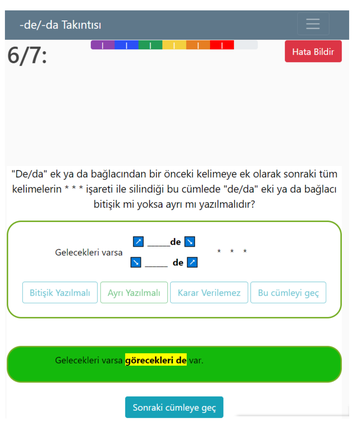
Perturbation Description	Δ_P
Remove the first letter of the previous word	0.0021
Remove the last letter of the previous word	0.0033
Remove the previous word	**0.33**
Remove the two previous word	**0.32**
Remove the three previous word	**0.32**
Remove the following word	0.0002
Remove the two following word	0.0026
Remove the three following word	0.0009
Remove a common prefix in the previous word	0.0016
Remove a common suffix in the previous word	0.03
Remove a common prefix in the following word	0.013
Remove a common suffix in the following word	0.012

Results of different perturbation techniques

These results show that the most important elements in a sentence determining the decisions of our model are the words which comes before the “de/da” clitic.

Human Based Computation Game

- We developed a GWAP to validate the findings of the perturbation-based explanation approach.
- In this game some words are masked in a sentence and the user is asked to predict the correct spelling of the “de/da” clitic.

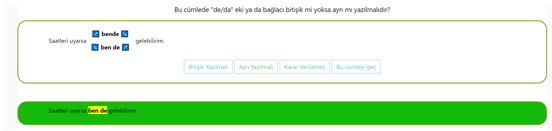


A screenshot from the GWAP

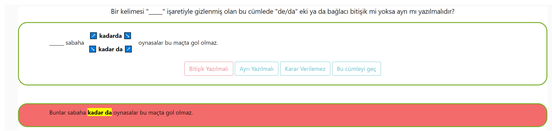
dedatakintisi.derlem.com

Game Modes

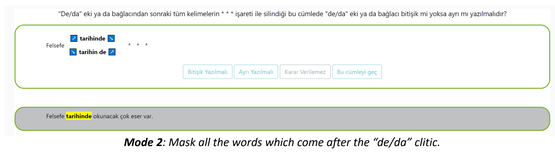
There are 7 modes in the game, each of which measures the significance of a particular region of a sentence.



Mode 0*: Do not mask anything.*



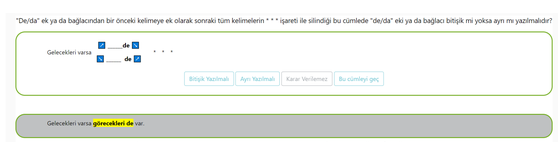
Mode 1*: Mask a random word.*



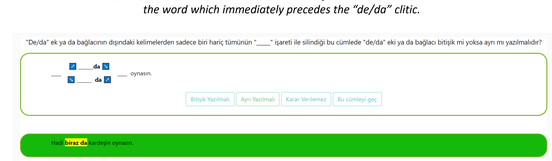
Mode 2*: Mask all the words which come after the “de/da” clitic.*

Mode 3*: Mask all the words which come before the “de/da” clitic excluding the one which immediately precedes the “de/da” clitic.*

Mode 4*: Mask all the words which come before the “de/da” clitic.*



Mode 5*: Mask all the words which come after the “de/da” clitic and the word which immediately precedes the “de/da” clitic.*

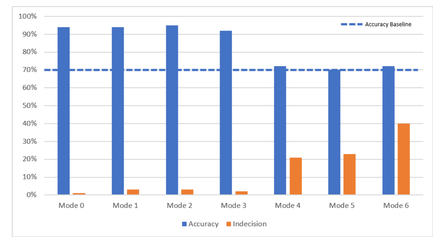


Mode 6*: Mask all the words except a random word.*

The “de/da” Spellchecker

Preliminary Results of the GWAP Experiment

- 447 people participated to the experiment and 13536 questions were solved in total.



- According to results of Mode 0, our participants had 94% success in determining of correct spelling of “de/da” clitic.
- Accuracy of the Mode 2 and 4 are **95%** and **72%** respectively. This shows that words which come **before** the “de/da” clitic are more important than the words which come **after.**
- Accuracy of the Mode 2 and 5 are **95%** and **70%** respectively. This shows that the word which immediately precedes the “de/da” clitic has a crucial role in the correct spelling.

Discussion and Future Work

- The findings of the explanation approach can be used to improve the spellchecker.
- This model has the potential to be integrated into a general-purpose spellchecker.

Acknowledgements

We would like to thank Suzan Üsküdarlı and Onur Güngör for their exceptional support throughout this year. We are also grateful to everyone who contributed to the experiment.

References

[1]: Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).