GTU Department of Computer Engineering
CSE 484 - Spring 2023
Homework 3 Report

Emirkan Burak Yılmaz
1901042659

1 Contents

2 Introduction

In this homework, the objective was to develop a classifier that determines whether the Turkish suffixes "de" and "ki" should be separated or not in a given sentence. The challenge arises from the fact that these suffixes can be written either separately or combined with the preceding word, leading to ambiguity. For detailed experiments and results, please consult the Jupyter Notebook file named **'tr_suffix_checker.ipynb'**.

3 Dataset Preparation

The training dataset was generated by extracting sentences from a Turkish Wikipedia dump, specifically focusing on instances where the suffixes "de" or "ki" were present. The labeling process involved categorizing sentences as either true (correctly spelled) or false (incorrectly spelled) based on the prescribed correct form.

1. Sentences containing the specified suffixes were extracted from the dataset.
2. To adhere to the homework instructions, all instances of the suffixes "de" and "ki" were modified to ensure they were not written separately. The modification involved combining the suffix with the preceding word.
3. The dataset was then partitioned into two categories: modified samples labeled as false, indicating a spelling error, and unmodified samples labeled as true, indicating no spelling error.

During the dataset preparation phase, a significant challenge emerged when handling sentences containing multiple suffixes. To address this issue, two primary approaches were considered. The first involved removing all suffixes from a sentence and then individually appending each suffix to create new sentences. However, this method presented a drawback as it resulted in some sentences losing their intended meaning. For example, the sentence "O tepedeki evler de bizim" transformed into "O tepe evler de bizim" after the removal of the first suffix. The second approach considered was the elimination of all sentences containing multiple suffixes. However, this solution came at the cost of losing approximately 9000 sample sentences. To strike a balance, the chosen strategy involved modifying only the first suffix in such sentences while leaving the others untouched. This compromise aimed to retain the integrity of sentence meanings while addressing the challenge posed by multiple suffixes in a more nuanced manner.

Suffix	Number of Samples
de	39048
ki	15661

Class	Number of Samples
True	48518
False	6191

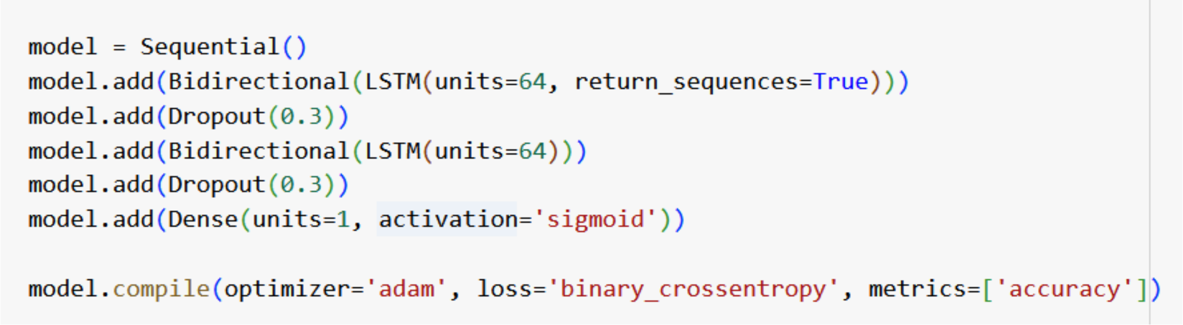
4 Word Embeddings with Word2Vec

In utilizing Word2Vec embeddings, my approach is rooted in the pursuit of capturing semantic information and contextual nuances of the Turkish language. Unlike character-based embeddings that focus solely on the spelling of words, Word2Vec allows us to represent words as vectors in a continuous semantic space, considering their meaning and context within sentences. By incorporating semantical information, the model gains a more nuanced understanding of the language, enabling it to generalize effectively to unseen examples with similar semantic contexts. This approach enriches the representation of words, fostering a more robust and context-aware model for addressing the intricacies of Turkish grammar, particularly in discerning the ambiguous "de" and "ki" suffixes.

Initially, I trained a Word2Vec model on a Turkish Wikipedia dump with 100 dimensions and a window size of 5. Subsequently, upon discovering an already trained model with 400 dimensions and a window size of 15, I transitioned to utilizing that pre-trained model for enhanced representation.

5 Neural Network Architecture

The neural network architecture implemented in this project addresses the intricate nature of the Turkish language, where the spelling of the suffixes "de" and "ki" is contingent on the contextual meaning of the sentence. For instance, in the sentence “Öğrenciler de geldi” (Students also came), the suffix "de" is separated, adding the meaning of "also." Conversely, in “Öğrencilerde gelişme var.” (There is progress in students), it is not separated and signifies locative information. To capture semantic context accurately, a Bidirectional Long Short-Term Memory (LSTM) architecture was chosen. LSTMs excel in capturing long-term dependencies, crucial for understanding nuanced relations in Turkish sentences. The bidirectional aspect enhances context comprehension by processing sequences in both directions, effectively capturing the intricate nuances of free word order in Turkish. Furthermore, dropout layers were incorporated to introduce regularization, enhancing the robustness of the model by preventing overfitting during training.

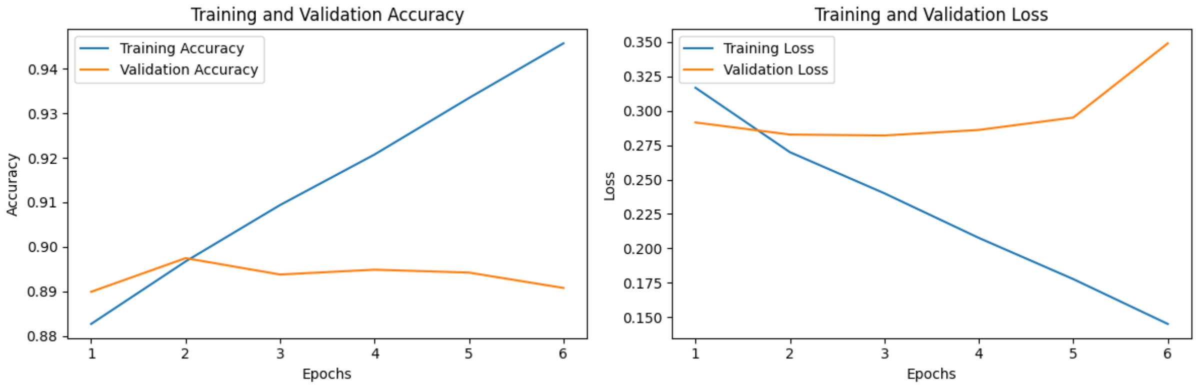


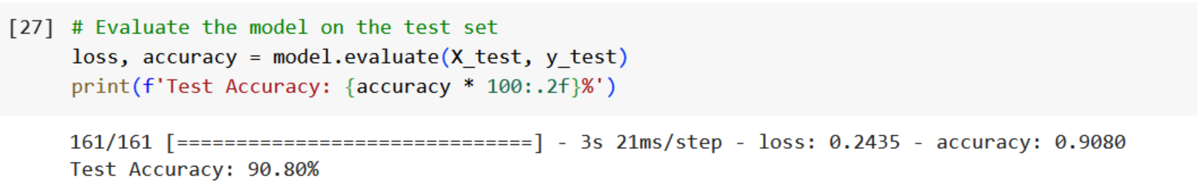
```
model = Sequential()
model.add(Bidirectional(LSTM(units=64, return_sequences=True)))
model.add(Dropout(0.3))
model.add(Bidirectional(LSTM(units=64)))
model.add(Dropout(0.3))
model.add(Dense(units=1, activation='sigmoid'))

model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
```

6 Training

The model was trained on a dataset consisting of sentences as input and binary labels indicating the correctness of suffix separation. Training was performed with the Adam optimizer and binary crossentropy loss. Early stopping was employed to prevent overfitting.





```
[27] # Evaluate the model on the test set
     loss, accuracy = model.evaluate(X_test, y_test)
     print(f'Test Accuracy: {accuracy * 100:.2f}%')

     161/161 [==============================] - 3s 21ms/step - loss: 0.2435 - accuracy: 0.9080
     Test Accuracy: 90.80%
```

7 Evaluation

Examining the evaluation outcomes, it becomes evident that the model effectively learns the ambiguity associated with the 'de' suffix. On the other hand, the generalization performance for the 'ki' suffix is not as robust as ‘de’ suffix. This discrepancy can be attributed to the dataset having a higher number of samples for the 'de' suffix in comparison to the 'ki' suffix. The provided screenshots illustrate the model's predictions, generated by inputting sentences as word embeddings and obtaining binary results indicating true or false.

Please be aware that a prediction result of 'True' indicates that there are no spelling errors related to either the 'de' or 'ki' suffixes. Conversely, 'False' signifies the presence of a spelling error.

Predictions:
1.) Sentence: 'Aklım hep **sende** kaldı.'
Prediction: True

2.) Sentence: 'Okadar **lezzetliki,** yemeye kıyamıyorum.'
Prediction: True

3.) Sentence: 'Evdeki tüm armutlar bitmiş.'
Prediction: True

4.) Sentence: '**Cezaevinde** sinema ile olan ilgisi devam etti.'
Prediction: True

5.) Sentence: 'Kalemleri **evde** kalmış.'
Prediction: True

6.) Sentence: '**Annemde** bizimle gelicek.'
Prediction: False

7.) Sentence: 'Kalemlerim **annemde** kaldı.'
Prediction: True

8.) Sentence: 'Yöre halkına göre **gölde** bir canavar yaşamaktadır.'
Prediction: True

9.) Sentence: '**Gunumuzde** kardes kulup anlasmasi aktif degildir. '
Prediction: True

10.) Sentence: 'Yeni **öğrencilerde** geziye geldi.'
Prediction: False

11.) Sentence: 'Yeni **öğrencilerde** gelişme var.'
Prediction: True

12.) Sentence: 'Bu **dönemde** asyali kolelerin sayisi cok fazladir.'
Prediction: True

13.) Sentence: 'Onlar aynı **caddede** buyumus ve birbirini seven iki asiktir.'
Prediction: True

14.) Sentence: 'Evlerin **tarihide** eskidir.'
Prediction: True

15.) Sentence: 'Bircok roma eyaleti bu **bolgede** kuruldu.'
Prediction: True

16.) Sentence: 'Liberal koylu **partiside** bu partiye katildi.'
Prediction: False

17.) Sentence: '**Gunumuzde** latince olarak bilinir.'
Prediction: True

18.) Sentence: '**Gunumuzdede** latince olarak bilinir.'
Prediction: True

19.) Sentence: 'Evini **dedesinde** bırakmış.'
Prediction: True

20.) Sentence: 'Onun **dedeside** gelecek.'
Prediction: False

21.) Sentence: '**Evdede** yemek yokmuş.'
Prediction: False

22.) Sentence: 'Herkes **evde** oturuyor.'
Prediction: True

23.) Sentence: 'En **iyiside** onun doğaçlamasıydı.'
Prediction: False

24.) Sentence: 'Onun aklı kırmızı **elbisede** kaldı.'
Prediction: False

25.) Sentence: '**İlkbaharda** bütün doğa canlanır.'
Prediction: True

26.) Sentence: 'Komşunun **köpeğide** durmadan havlıyor.'
Prediction: True

27.) Sentence: 'Beni yanlış **anlamada** o iş öyle yapılmaz.'
Prediction: False

28.) Sentence: 'Yarın **akşamda** bizde ders çalışalım.'
Prediction: True

```
29.) Sentence: 'Masadaki bardağı uzatır mısın?'
Prediction: True

30.) Sentence: 'Çevremizi temiz tutalımki başkaları rahatsız olmasın.'
Prediction: False

31.) Sentence: 'Penceremdeki çiçek soğuktan dondu.'
Prediction: True

32.) Sentence: 'Duydumki unutmuşsun gözlerimin rengini.'
Prediction: False

33.) Sentence: 'Beni dinlemedinki gerçekleri sana anlatayım.'
Prediction: False

34.) Sentence: 'Kitap okuki kelime dağarcığın gelişsin.'
Prediction: True

35.) Sentence: 'Benki hep sizin için çalıştım.'
Prediction: True

36.) Sentence: 'Benimki yine gelmiş.'
Prediction: True

37.) Sentence: 'Kiminki kazanacak göreceğiz.'
Prediction: True

38.) Sentence: 'Patlıcanları ince ince doğraki güzel pişsin.'
Prediction: True

39.) Sentence: 'Yemeklerini yeki çabuk iyileşesin.'
Prediction: False

40.) Sentence: 'Tutki karnım acıktı, o zaman ne yapıcam?'
Prediction: False
```

8 Conclusion

In conclusion, the primary challenge encountered in this homework revolved around the creation of a suitable dataset. Various approaches were explored to curate the most effective dataset, with a fundamental understanding that disambiguating the suffixes 'de' and 'ki' requires knowledge of the sentence's meaning or context. Leveraging Long Short-Term Memory (LSTM) as the neural network architecture proved beneficial for capturing long-term relations and comprehending sentence context. Additionally, the incorporation of Word2Vec embeddings enhanced the model's ability to generalize, ensuring that contextual nuances remained intact even when replacing words with semantically similar counterparts. The trained model achieved a commendable 90% accuracy, demonstrating robust performance in disentangling the ambiguity associated with the 'de' suffix. However, challenges persist with the 'ki' suffix, primarily stemming from a lower number of examples in the dataset compared to 'de', leading to some confusion in disambiguation.