DEEP LEARNING TECHNIQUES FOR HANDLING “DE-DA” TURKISH CLITICS

Alperen Değirmenci – Hasan Öztürk

Advisors: Suzan Üsküdarlı – Onur Güngör

Department of Computer Engineering, Boğaziçi University

Introduction and Motivation

- In Turkish, “de/da” misspelling error is quite common, even among native speakers. Because its usage depends on the context rather than the morphology of the words.
- Arıkan et al. [1] obtained state-of-the-art results in 2019.
- Google published a study named BERT for NLP pre-training in 2018. It has achieved state-of-the-art performance in 11 natural language understanding tasks. [2]
- Our goal is to employ BERT to improve Arıkan et al. [1] ’s results.

BERT

- BERT is a new language representation model that stands for **Bidirectional Encoder Representations from Transformers.**
- BERT employs “masked language model” (MLM) pre-training objective, which masks a random word in a sentence and tries to predict the word based on its context.

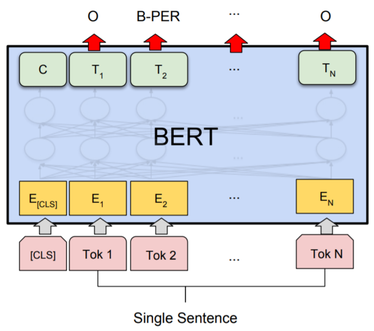

Figure 1 - BERT Sequence Tagging Representation

- Arıkan et al. represents the input with **Word2Vec**, **GloVe** and **FastText** embeddings.
- We appended BERT embeddings on top of them in order to better capture the context.
- We obtained BERT embeddings in two ways:

  BERT Multilingual: A single language-model pre-trained from a corpora in 104 languages -- including Turkish.
  BERT with Custom Data: Embeddings obtained from a training with a large Turkish corpus.

Model

Input Representation: Word Embeddings

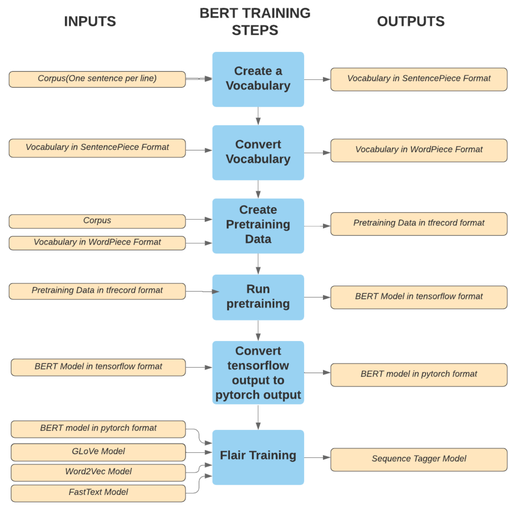

Figure 2 - BERT pre-training steps

A vocabulary is created to capture the common strings and substrings such as "**-mek**/**-mak**" or "**-ler/lar**“ using **SentencePiece** [3].

Bence	O
daha	O
iyide	B-ERR
olabilirdi	O
.	O

Figure 3 - CoNLL Data Format

Model: Bidirectional LSTM with Two Layers

- A sequence tagging model is trained on a Bidirectional LSTM with two layers using Flair NLP framework.
- Every word in the input is labeled with one of two tags: **B-ERR** & **O** to indicate if the word has an error. These tags are used to calculate the cost function.

Figure 4 - Flair training steps

Results

Embeddings					P	R	F1	Acc
G	F	W	MB	CB	%	%	%	%
+	+	+			91.56	82.28	86.67	76.48
+	+	+	+		91.11	78.68	84.44	73.07
+	+	+		+	88.25	83.54	85.83	75.18

Figure 5 - Results in the automatically created test set.

Embeddings					Accuracy
G	F	W	MB	CB	%
+	+	+			64
+	+	+	+		76
+	+	+		+	78

Figure 6 - Results in 100 manually curated challenging sentences.
***G:** GloVe **F:** FastText **W:** Word2Vec **MB:** Multilingual BERT Embeddings **CB:** Custom BERT Embeddings*

SUCCESSES	FAILS
Input: Sende başını alıp gitme ne olur. **Output:** Sende başını alıp gitme ne olur.	**Input:** Gömleğin önünüde iliklersen iyi olur. **Output:** Gömleğin önünüde iliklersen iyi olur.
Input: Gerçekleri tarih yazar tarihide Galatasaray. **Output:** Gerçekleri tarih yazar tarihide Galatasaray.	**Input:** Yediğinde içtiğinde senin olsun bize gördüklerini anlat. **Output:** Yediğinde içtiğinde senin olsun bize gördüklerini anlat.
Input: Olsun demekte zor artık. **Output:** Olsun demekte zor artık.	**Input:** Kimselerede bakmadım senden daha güzel. **Output:** Kimselerede bakmadım senden daha güzel.

Figure 6 - Some results from sentences.

Access the Web Interface

Future Work

- Completing BERT pre-training on a powerful GPU grid, e.g. TRUBA
- Extending the model for other clitics such as “**mi/mı**” and “**ki**”
- Integrating the model into a spellchecker
- Improving the model’s web interface

References

[1] Ugurcan Arikan Onur Gungor and Suzan Uskudarli. Detecting clitics related orthographic errors in turkish. Recent Advances in Natural Language Processing 2019, Sept. 2019.
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[3] Taku Kudo and John Richardson.SentencePiece: A simple and language independent sub-word tokenizer and detokenizer for neural text processing. InProceedings of the 2018 Con-ference on Empirical Methods in Natural Language Processing: System Demonstrations, pp.66–71, Brussels, Belgium, November 2018.