НИУ ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Автоматический анализ грамматических ошибок в предложении на английском языке

Выполнила студентка 19ФПЛ1 Шханукова Милана
Научный руководитель: Шадрина Елена Викторовна

2020-2021

Описание

Задача error detection в предложениях на английском языке относительно их представлений в виде токенов и синтаксического дерева.

Цель:

Проанализировать способы детектирования ошибки в предложении и их влияние на общее качество детекции.

Задачи:

1. Выбрать несколько типов представления предложения как последовательности.
2. Построить архитектуру нейронной сети.
3. Сделать независимое сравнение результатов согласно выбранным представлениям предложений.

Описание

Объект - обнаружение грамматических ошибок в предложении
Предмет - использование нейронных сетей при анализе предложения

Основные методы:
1. Синтаксический анализ.
2. Статистический анализ.
3. Компьютерное моделирование.

Гипотеза

Для обнаружения ошибки в предложении достаточно лишь его синтаксической структуры или грамматических показателей каждого элемента из заданной последовательности.

words	He	**do**	his	task	always
tokens	PRP	**VBP**	PRP$	NN	NN
tree	nsubj	**ROOT**	poss	dobj	npadvmod
mistakes	none	**MIST**	none	none	none

токенизирование выполнено с помощью библиотеки Spacy

PRP - personal pronoun,
PRP$ - possesive pronoun
VBP - Verb non-3rd person singular present forms
NN - common nouns

nsubj - nominal subject
poss - modifier
dobj - direct object
npadvmod - noun phrase as adverbial modifier

Лингвистическая направленность гипотезы

«Языковая компетенция» *- как полное знание о родном языке, которое позволяет «идеальному говорящему-слушающему» судить о правильности и осмысленности высказываний. (Н. Хомский)*

Это знание описывается грамматикой, включающей в себя **набор правил**, регулирующих порождение всех возможных в данном языке структур предложений путем **преобразования исходной конструкции**, а также описание **грамматических отношений** в самих предложениях и между ними.

1. **Rei и Yannanakoadakis (2016) embeddings, biLSTM**

2. **«Context is Key: Grammatical Error Detection with Contextual Word Representations» Samuel Bell, Yannakoudakis, Marek Rei**, 2020

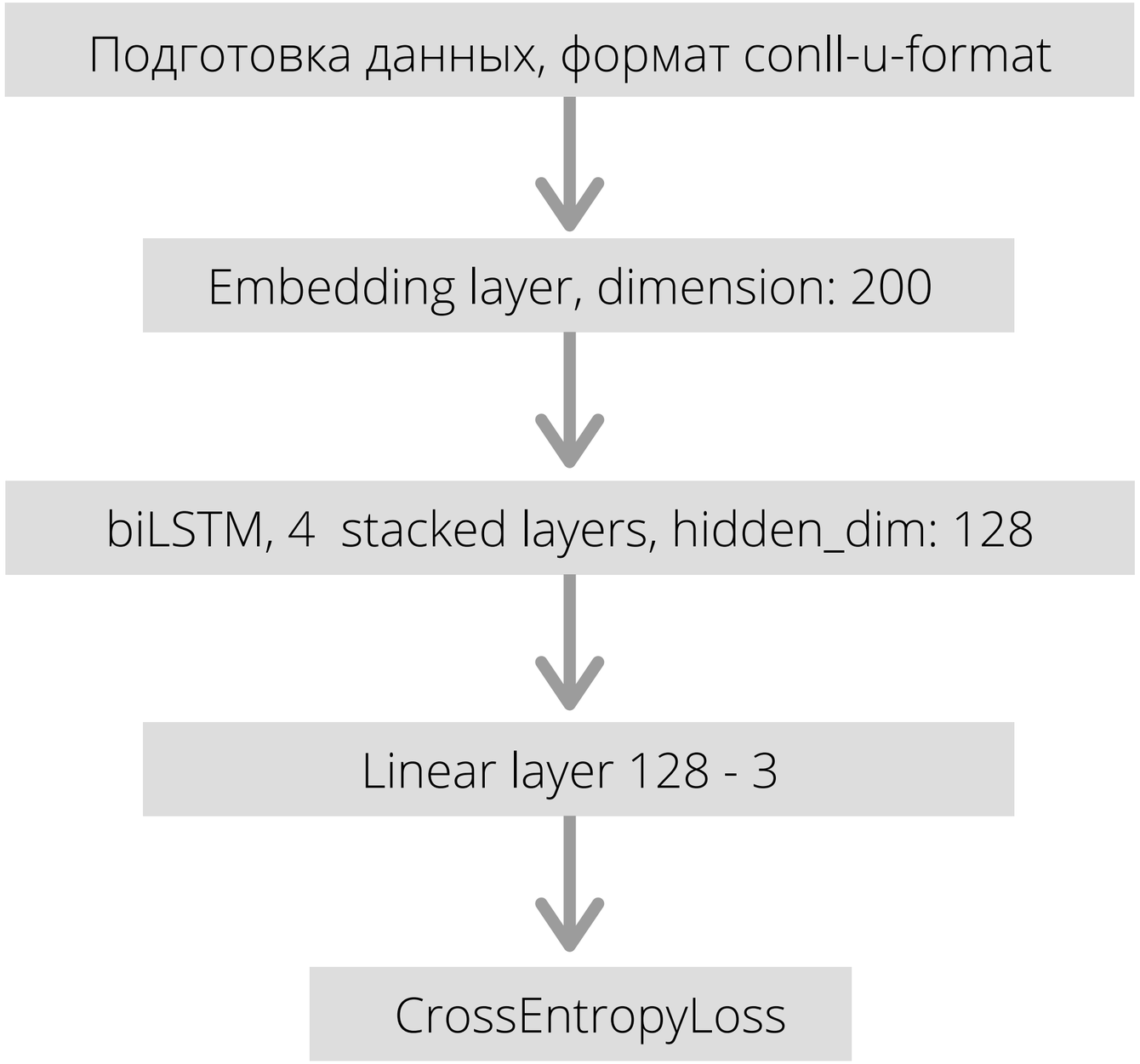
3. **«Lex-Pos Feature-Based Grammar Error Detection System for the English Language» Nancy Agarwal, Mudasir Ahmad Wani * and Patrick Bours**

Предыдущие работы

В курсовой работе рассмотрены исследования, которые отражают специфику поставленной проблемы.

Модель

В данной структуре отражены лишь главные составляющие алгоритма



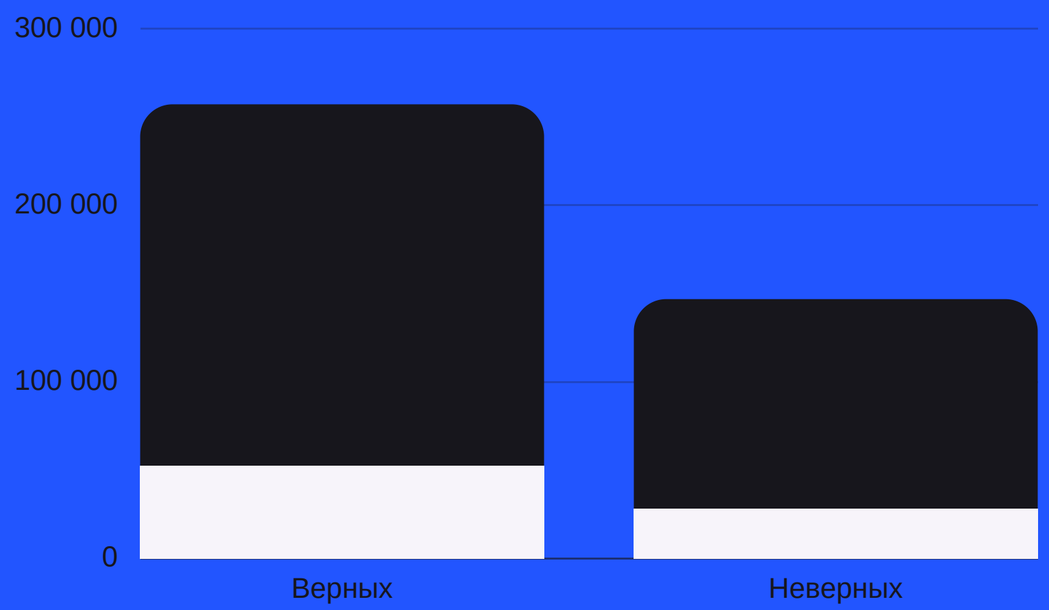
Данные

Общее число предложений
484297, короче 20 слов 403317

Корпуса:

1. Lang - 8
2. FCE
3. Write and Improve

0.36% неверных предложений во всем датасете
0.43% неверных предложений использовано при обучении



Метрика и функция потерь

Loss

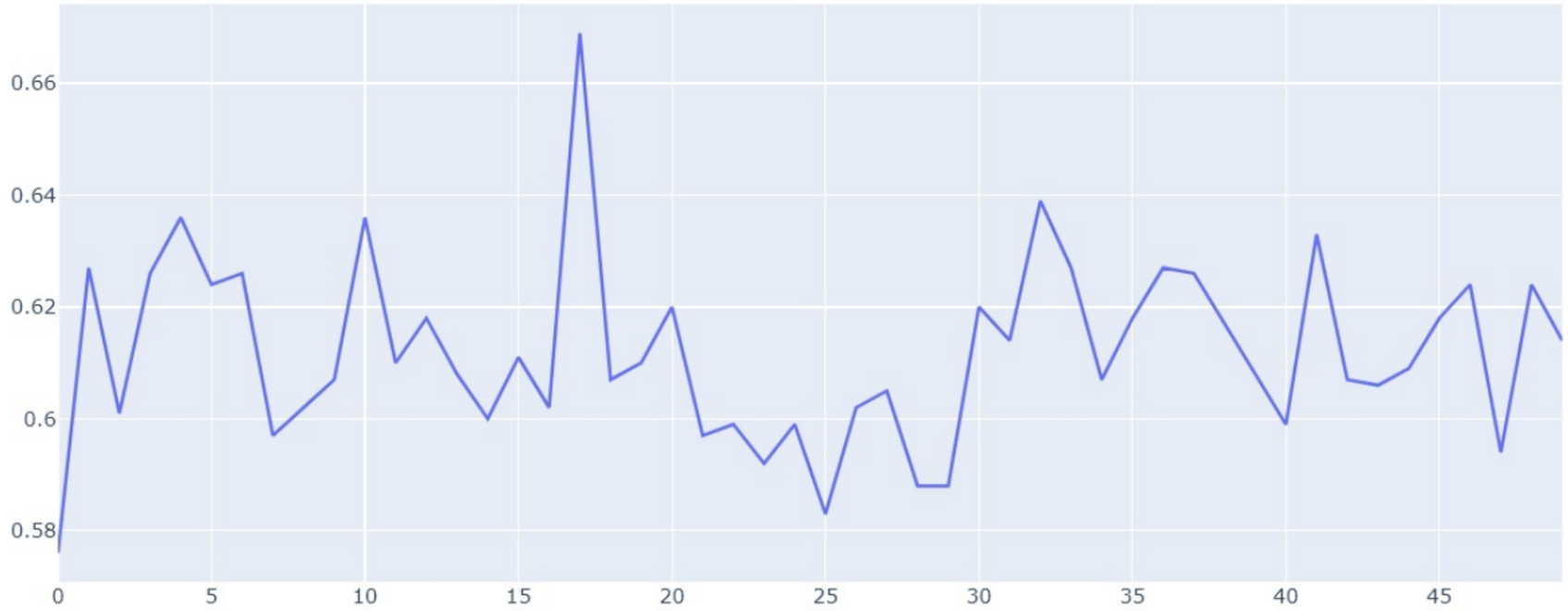
CrossEntropyLoss - количественная оценка разницы между двумя распределениями вероятностей.

F1

F1 - гармоническое среднее между precision и recall. Является оценкой алгоритма и позволяет раскрыть его качество.

Train на токенах

График не сходится, loss не уменьшается равномерно.



обучение только на токенах, 50 эпох, lr = 0.001. batch_size = 32

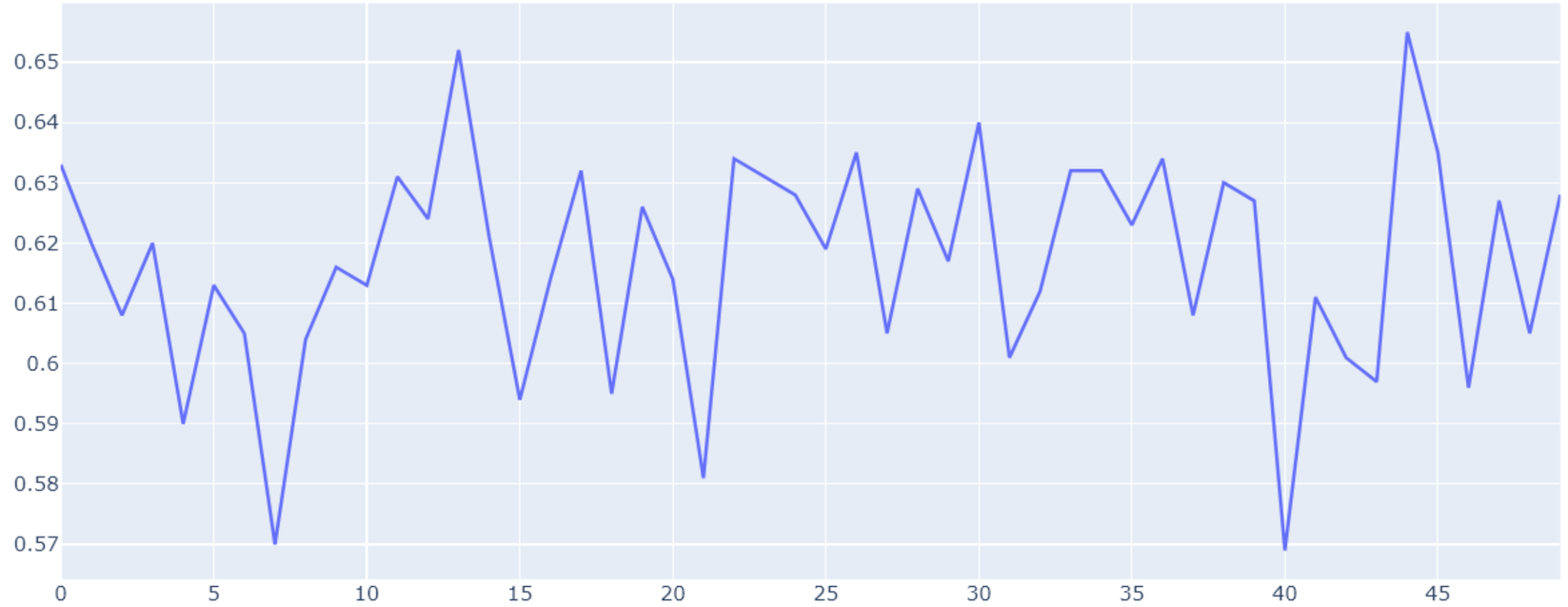
F1 на токенах

Также нет единого увеличения f1-метрики, сходимость нарушена.

обучение только на токенах, 50 эпох, lr = 0.001. batch_size = 32

Train на синтаксическом дереве

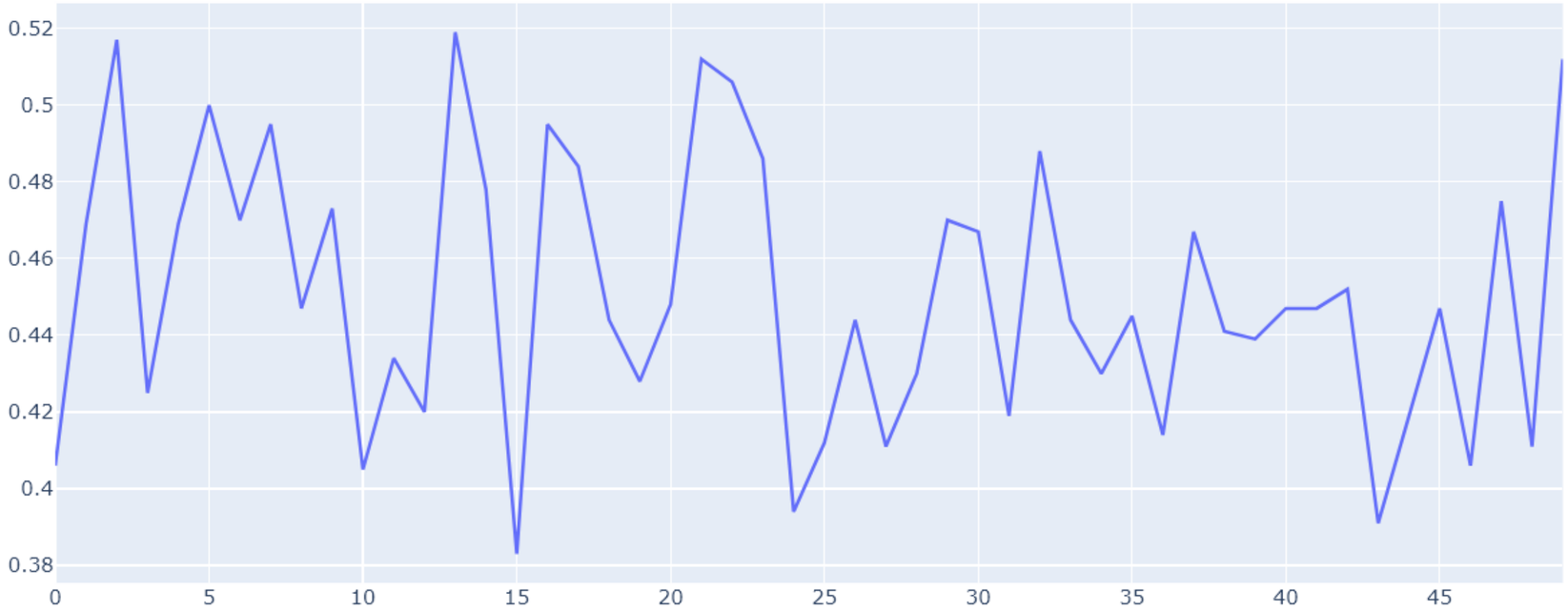
График не сходится, на каждой эпохе loss сильно различается с предыдущим.



обучение только на токенах, 50 эпох, lr = 0.001. batch_size = 32

F1 на синтаксическом дереве

F1 метрика только на 4 эпохах превышает вероятность 0.5, что говорит о неадекватности модели.



обучение только на токенах, 50 эпох, lr = 0.001. batch_size = 32

Результаты

Модель не смогла сойтись, метрики не превысили 0.528 для f1

	Обучение на токенах	Обучение на синтаксическом дереве
Train loss (min)	0.537	0.569
F1 (max)	0.528	0.519

Заключение

1. **Гипотеза опровергнута**
2. **Модель показала невозможность сходимости на абстракции предложений**
3. **Без представлений слов невозможно сделать предсказания**

Причины несходимости:

1. Повторение токенов на разных местах в предложении
2. Маленький словарь токенов
3. Нехватка информации с токенов.
4. Ошибки в токенизации Spacy