МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева (Самарский университет)»

Институт информатики, математики и электроники

Факультет информатики Кафедра программных систем

ОТЧЁТ

Курсовая работа по курсу

«Нейронные сети глубокого обучения»

Студент П.Д. Поборознюк

Студент А.А. Саенкова

Преподаватель А.Н. Жданова

Самара 2022

Задание

Дано предложение и выбраны некоторые слова в предложении. В результате работы сеть должна выдать прогноз отношения между этими словами.

Например, дано предложение «Остап Бендер прошел путь от простого сына лейтенанта Шмидта до главы конторы “Рога и Копыта”» и слова «Остап» и «конторы». В результате должно быть обозначено отношение между ними - «возглавляет», и должна быть указана вероятность этого отношения, например 0.97403.

Оглавление

[ВВЕДЕНИЕ 4](#_Toc106881117)

[1 Подбор и анализ датасета 5](#_Toc106881120)

[2 Постановка задачи 5](#_Toc106881121)

[3 Архитектура используемой модели 6](#_Toc106881122)

[4. Описание алгоритма обработки данных 9](#_Toc106881123)

[5 Описание экспериментов 11](#_Toc106881124)

[ВЫВОДЫ 18](#_Toc106881125)

[ПРИЛОЖЕНИЕ А 19](#_Toc106881126)

[ПРИЛОЖЕНИЕ Б 21](#_Toc106881127)

[ПРИЛОЖЕНИЕ В 24](#_Toc106881128)

[ПРИЛОЖЕНИЕ Г 27](#_Toc106881129)

[ПРИЛОЖЕНИЕ Д 28](#_Toc106881130)

# ВВЕДЕНИЕ

# Выделение и классификация отношений в тексте является одной из актуальных задач в области нейронных сетей. Решение подобных задач находят своё применение при структуризации неструктурированной информации. Кроме того, это важно в вопросно-ответных, диалоговых и поисковых системах – всегда, когда нужно анализировать вопрос и понимать, к какому типу он относится, а также, какие ограничения накладываются на ответ.

# В данной работе будет рассматриваться реализация выделения отношений для задаваемого через веб-интерфейс предложения и пары слов с использованием передовой нейросети BERT.

# 1 Подбор и анализ датасета

В работе будет использоваться датасет wiki80, содержащий 80 классов отношений по 700 предложений в каждом, всего предложений 56000.

Датасет описывает следующие отношения: child, located in or next to body of water, occupant, contains administrative territorial entity и другие.

# 2 Постановка задачи

Дано множества слов в предложении. Для каждой пары слов существует некая целевая функция . Необходимо построить классификатор , максимально близкий к .

Для обучения классификатора используется набор предложений и пар слов в нём.

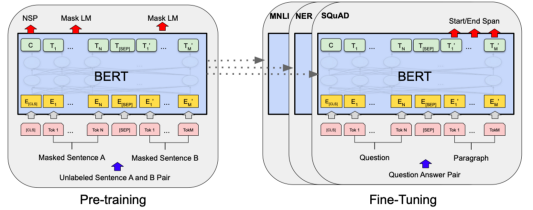
# 3 Архитектура используемой модели

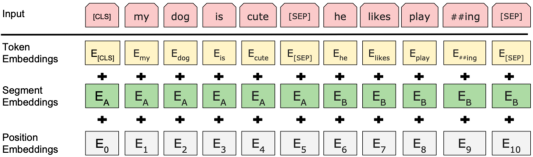
Для построения отображения будет использоваться рекуррентная нейронная сеть BERT (Bidirectional Encoder Representation Transformers), в переводе на русский “двунаправленная нейронная сеть-кодировщик” – модель представления языка, которая предназначена для предварительного обучения глубоких двунаправленных представлений на простых немаркированных текстах путем совмещения левого и правого контекстов во всех слоях.

В структуре BERT есть два этапа:

1. Предварительное обучение – модель обучается на немаркированных данных, выполняя различные задачи. Предобучение требует больших мощностей и проводится крупными компаниями, предобученные модели находятся в свободном доступе.
2. Точная настройка – модель загружается с предварительно обученными параметрами и обучается на помеченных данных из последующих задач.

На рисунках 1 и 2 представлены представления рекуррентной нейронной сети

  
Рисунок 1 – структура модели нейросети BERT до и после дообучения

  
Рисунок 2 – схема вычисления представления слова на основе его эмбеддингов токена, сегмента и позиции.

Данная архитектура является наиболее подходящей для решения задачи выделения отношений в тексте так как модель BERT считается наилучшей в сфере машинного лингвистического анализа и токенизации.

В модели, кроме предобученных слоёв BERT, также используется слой Softmax, позволяющий классифицировать полученные данные для классификации отношения между заданными словами.

# 4. Описание алгоритма обработки данных

На рисунке 3 представлен алгоритм обработки данных (pipeline), состоящий из предварительной обработки текста энкодером, модуля токенизации (BERT) и слоя классификации Softmax. Предварительная обработка нужна для очистки текста от ненужных символов или символов, которые в процессе обучения могут повлиять на обучение модели. Токенизация в алгоритме обработке это преобразование слов в численное представление, каждому уникальному слову присваивается его уникальное значение, на основе эмбеддингов. Для этого в модели использется предобученная нейросеть BERT. Последний слой – слой Softmax, принимающий представление предложения и ключевых слов в числовом виде и классифицирующий отношение между ними.

Рисунок 3 – pipeline классификации отношений

BERT

SoftmaxLayer

**…**

**…**

Класс отношения слова 1 к слову 2

Слово 1

Предложение

Слово 2

BERT-encoder

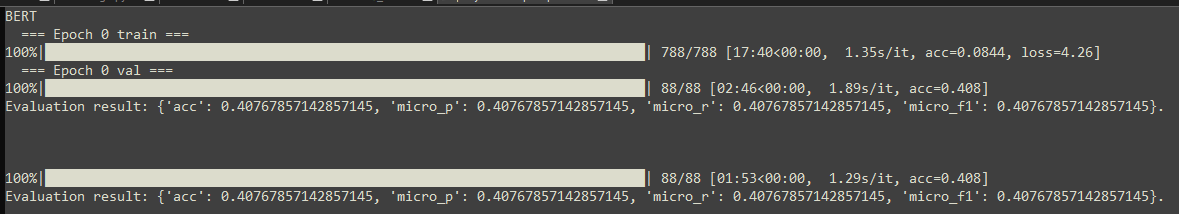
# 5 Описание экспериментов

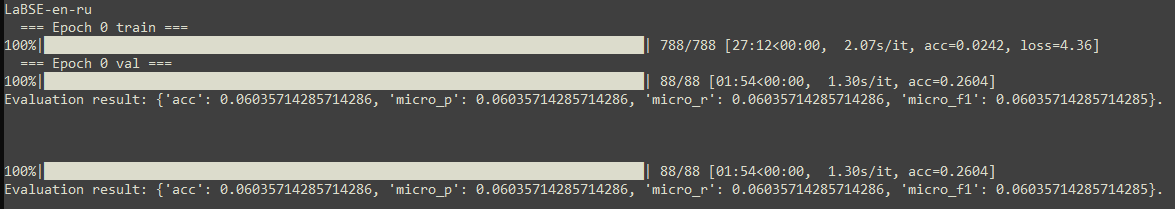
Так как обучение модели BERT – процесс слишком ресурсоёмки, будет использоваться предобученная модель BERT из тулсета openNRE, дообученная на датасете wiki80. Слой softmax будет также дообучен для решения задачи классификации заданного отношения на основе данных, полученных от выходного слоя BERT. Результаты дообучения показаны на рисунке 5.

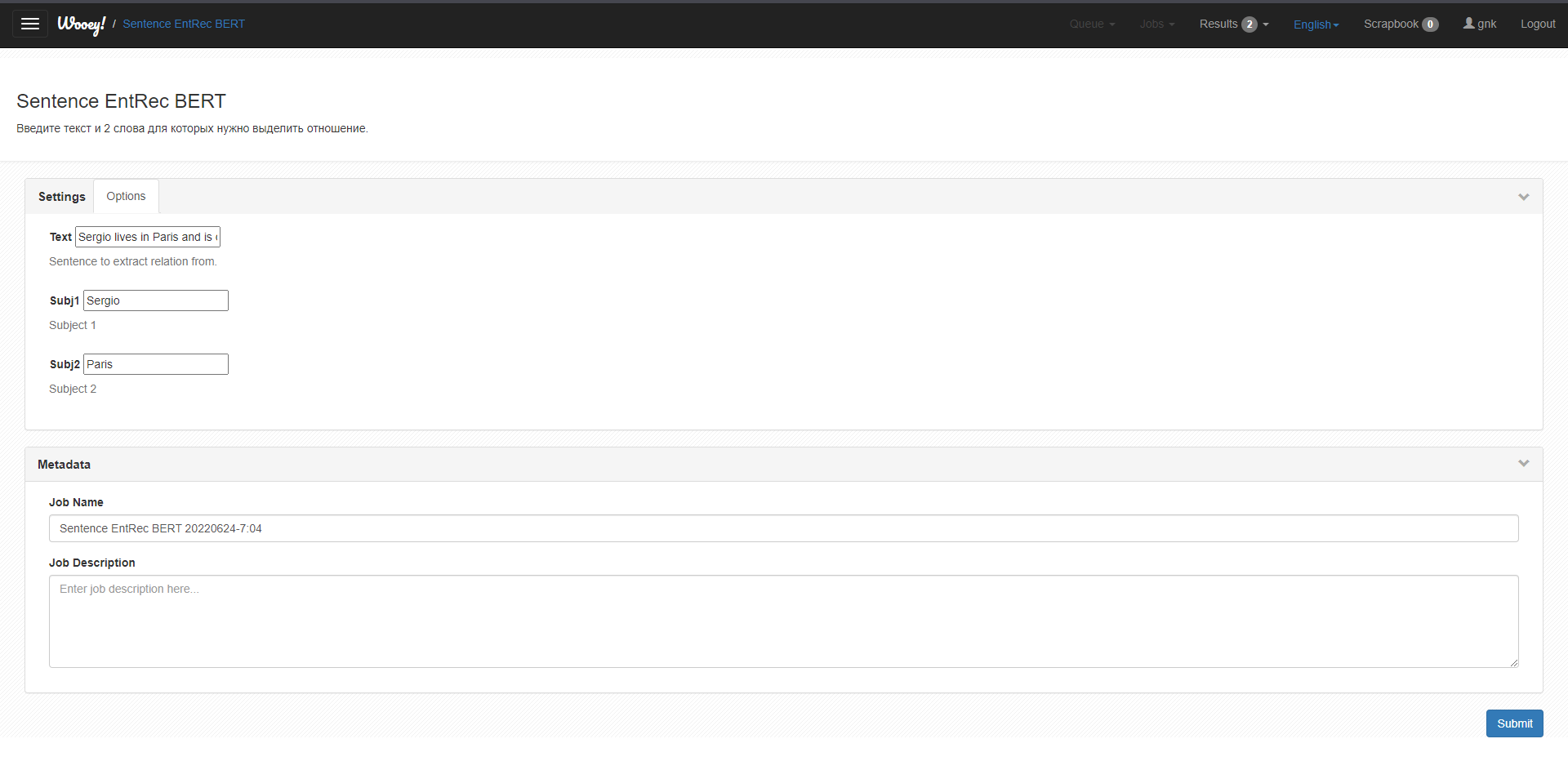
Также, для сравнения эффективности моделей, будет обучена модель LaBSE, предобученная для русского и английского языка, также использующая выходной слой softmax, дообученная на датасете wiki80. Результаты дообучения представлены на рисунке 6.

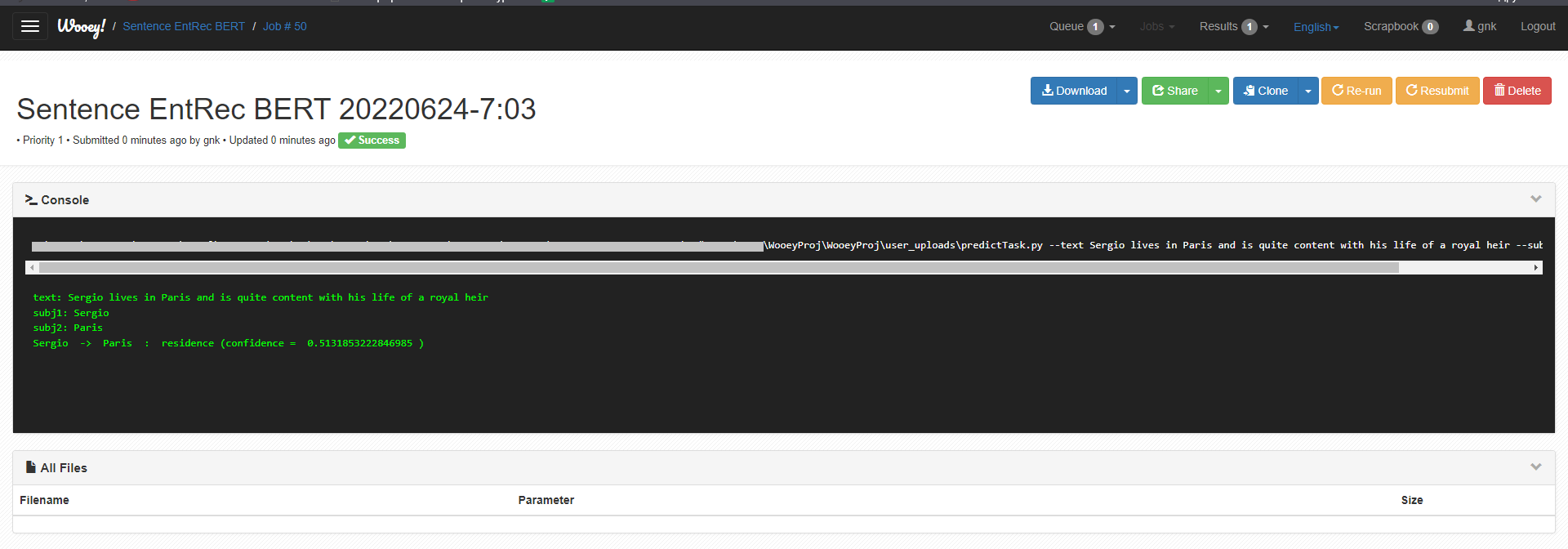
В результате, модель, использующая сеть BERT для токенизации предложения, показывает очень внушительные результаты, успешно классифицируя отношения в 40% случаев, а модель, использующая LaBSE нуждается в чрезмерном дообучении т.к. имеет лишь точность 6%.

Для создания веб-интерфейса был использован веб-интерфейс Wooey, использующий фреймворк Django. Wooey позволяет максимально быстро и просто создать и настроить веб-интерфейс для облачного выполнения программ Python. На рисунках 6 и 7 представлены скриншоты работы системы.

  
Рисунок 4 – результат дообучения для модели BERT+Softmax

  
Рисунок 5 – результат дообучения для модели LaBSE+Softmax

   
Рисунок 6 – окно веб-интерфейса задания предложения и двух слов для классификации их отношения

  
Рисунок 7 – результат работы веб-интерфейса системы для определения отношений в тексте

# ВЫВОДЫ

В данной работе была рассмотрена нейросеть BERT, являющаяся наилучшей в задачах машинной обработки речи, позволяющая наиболее точно классифицировать отношения между заданными словами в предложении. Также была предпринята попытка использования нейросети LaBSM, но опытным путём было выяснено, что она нуждается в чрезмерном дообучении для выполнения задачи классификации отношений.

Итоговая нейросеть была использована для создания веб-интерфейса, способного классифицировать отношения между двумя заданными словами в заданном предложении.

# ПРИЛОЖЕНИЕ А

Код доступен для просмотра по следующей ссылке: <https://github.com/algorithm-ssau/SmirnovMatyshkina6131courseProject>.

Клонировать репозиторий: *git clone*