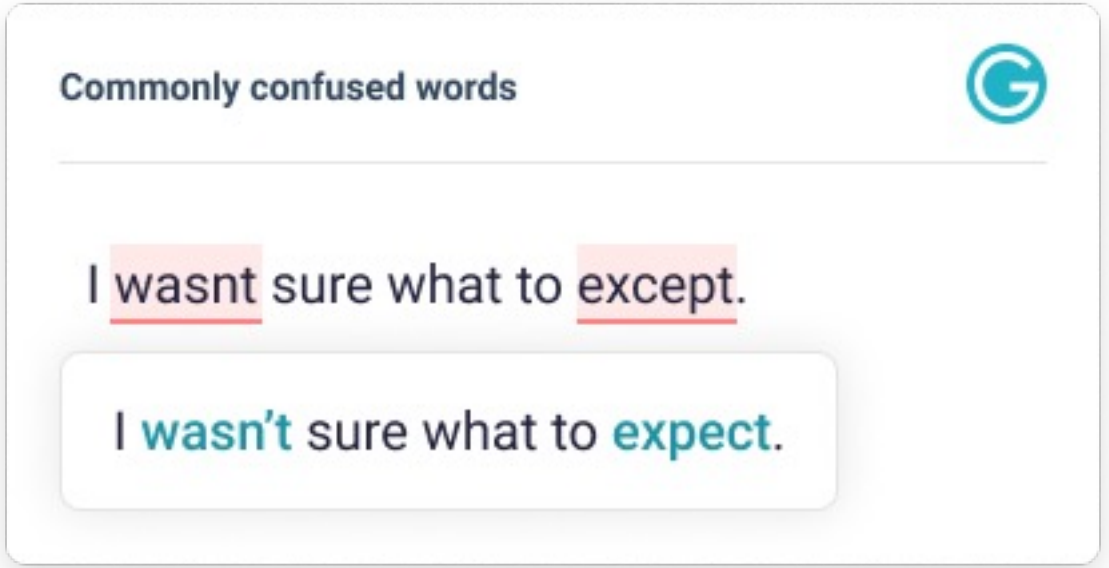
Задача исправления опечаток

План занятия

1. Постановка задачи
2. Модель Норвига
3. Модель на N-граммах
4. Трансформеры. Трюк с загрязнением данных
5. Библиотеки для решения задачи

Постановка задачи

Задача исправления опечаток в машинном обучении заключается в разработке моделей, способных автоматически обнаруживать и исправлять ошибки в тексте, вызванные опечатками. Опечатки могут включать в себя ошибки при наборе, неверное распознавание слов, перестановку букв, альтернативное написание слов и другие виды ошибок, связанных с неверным вводом текста.



Расстояние Левенштейна

Для поиска кандидатов для исправления слова часто используется расстояние Левенштейна.

Расстояние Левенштейна (или редакционное расстояние) — это метрика сходства между двумя строковыми последовательностями.

- Чем больше расстояние, тем более различны строки.
- Для двух одинаковых последовательностей расстояние равно нулю.

По сути это ***минимальное число односимвольных преобразований (удаления, вставки или замены), необходимых, чтобы превратить одну последовательность в другую***.

Б	И	Б	А
Б	О	Б	А

LEV(’БИБА’, ‘БОБА’) = 1

Расстояние Левенштейна

А	В	С	Т	Р			И	Я
А	В	С	Т	Р	А	Л	И	Я

LEV(’АВСТРИЯ’, ‘АВСТРАЛИЯ’) = 2

	К	О	Т	И	К	
С	К	О	Т	И	Н	А

LEV(’КОТИК’, ‘СКОТИНА’) = 3

Расстояние Левенштейна

Чему равно расстояние Левенштейна между словами HONDA и HYUNDAI?

Расстояние Левенштейна

Чему равно расстояние Левенштейна между словами HONDA и HYUNDAI?

H	O		N	D	A	
H	Y	U	N	D	A	I

H	Y	U	N	D	A	I
H		O	N	D	A	

Расстояние Левенштейна

Для вычисления расстояния Левенштейна используется алгоритм Вагнера-Фишера, хорошо рассказанный здесь.

Но как правило мы будем пользоваться готовыми его реализациями – это следующие питоновские библиотеки:

- **strsimpy**
- **python-Levenshtein**
- **NLTK**

https://colab.research.google.com/drive/1AzpUnreWsUxL_w64yEChDIvnQ9HrB2H6?usp=sharing

Модель Норвига “How to Write a Spelling Corrector” (2007)

В статье представлен простой метод исправления опечаток на основе статистики слов. В методе используется частота встречаемости слов и расстояние Левенштейна для предложения наилучшего варианта исправления опечаток.

Модель Норвига “How to Write a Spelling Corrector” (2007)

1. Корпус текста: Используется некоторый корпус текста, который предоставляет информацию о частоте встречаемости слов в естественном языке.

2. Генерация кандидатов: Для данного слова с опечаткой генерируются кандидаты исправлений (на основе расстояния Левенштейна). Кандидаты получаются из слов-оригиналов операциями вставки, удаления, замены и транспозиции букв.

3. Оценка вероятности: Для каждого кандидата вычисляется вероятность (посчитанная по большому корпусу текстов), основанная на частоте встречаемости слов в корпусе. Более вероятные слова имеют более высокую вероятность быть правильными исправлениями.

4. Выбор наилучшего кандидата: Выбирается кандидат с наивысшей вероятностью в качестве наилучшего исправления.

5. Возврат результата: Возвращается исправленное слово.

Модель на N-граммах

Эта модель – развитие модели Норвига. В ней вероятности считаются с учетом контекста.

Идея: если мы встречаем слово не из словаря, то среди слов из словаря, находящихся от него на расстоянии 1 или 2, выбираем слово с наибольшей вероятностью находиться на этой позиции, зная предыдущие несколько слов.

Модель на N-граммах

1.Создание модели N-грамм:

1. Собираем обучающий корпус текста, на котором можно обучить модель N-грамм.
2. Делим текст на N-граммы, где N - это число элементов в последовательности. Например, для биграмм N=2, для триграмм N=3.
3. Подсчитываем частоту встречаемости каждой N-граммы в корпусе.

2.Генерация кандидатов:

1. Для данного слова с опечаткой генерируем кандидатов на основе операций вставки, удаления, замены и транспозиции букв, так же, как и в методе Норвига.

3.Оценка вероятности с использованием N-грамм:

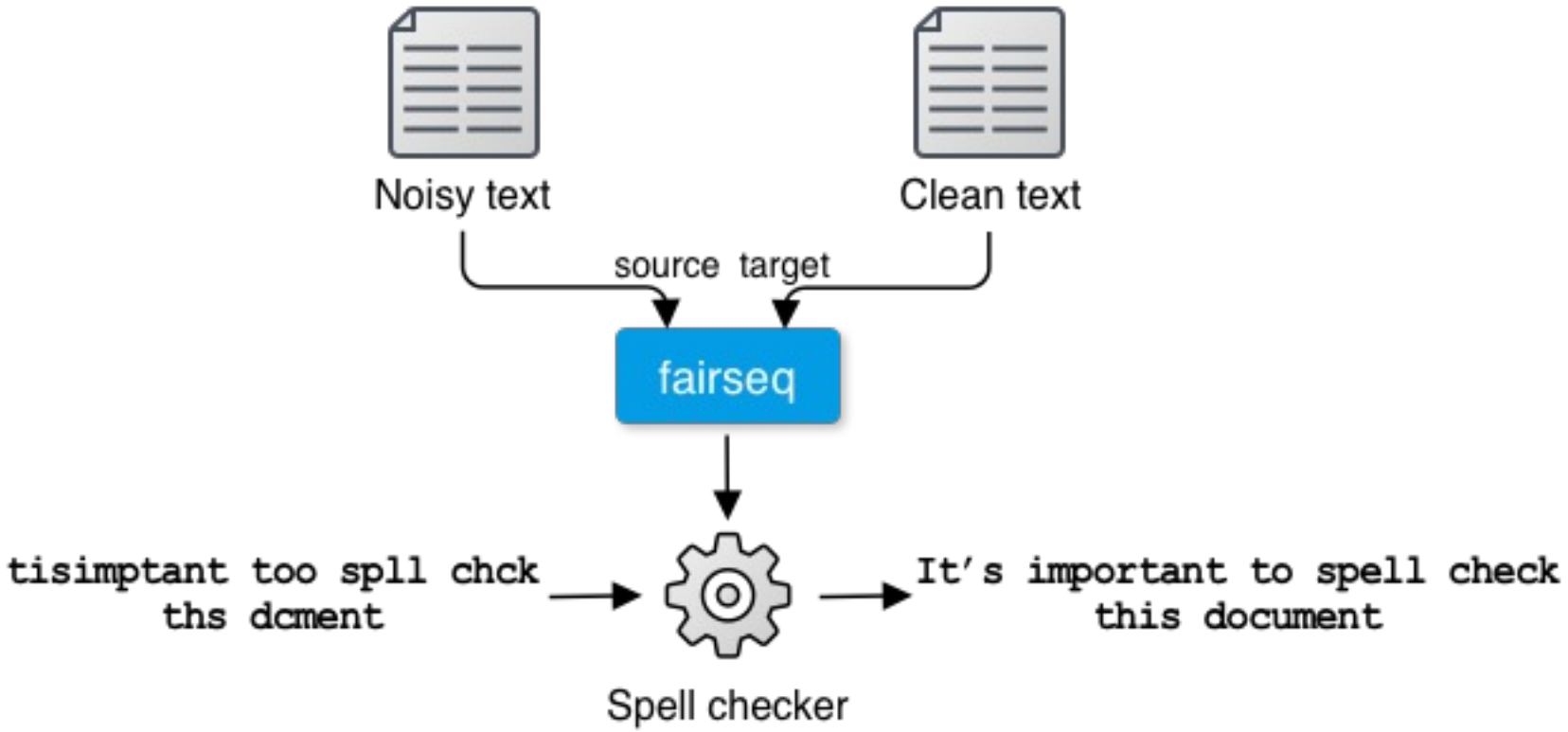
1. Для каждого кандидата оцениваем вероятность его появления в тексте с использованием модели N-грамм: это можно сделать при помощи оценки вероятности всей фразы, содержащей слово-кандидата, основанной на частоте встречаемости соответствующих N-грамм (по формуле условной вероятности).

4.Выбор наилучшего кандидата:

1. Выбираем кандидата с наибольшей вероятностью в качестве наилучшего исправления.

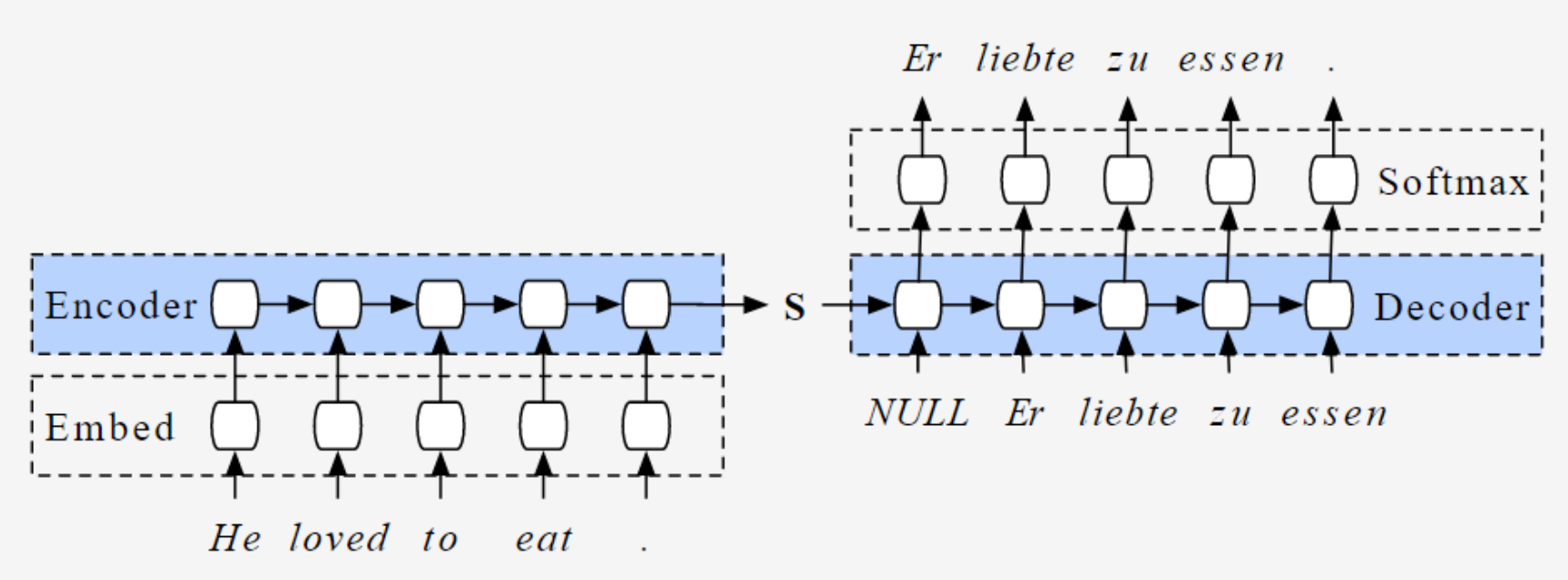
Трансформеры для задачи исправления опечаток

Общая схема обучения при помощи фреймворка fairseq:



Spell Correction as Machine Translation

Архитектура модели:



Мы решаем задачу как задачу **sequence-to-sequence** путем генерации слов без опечаток.

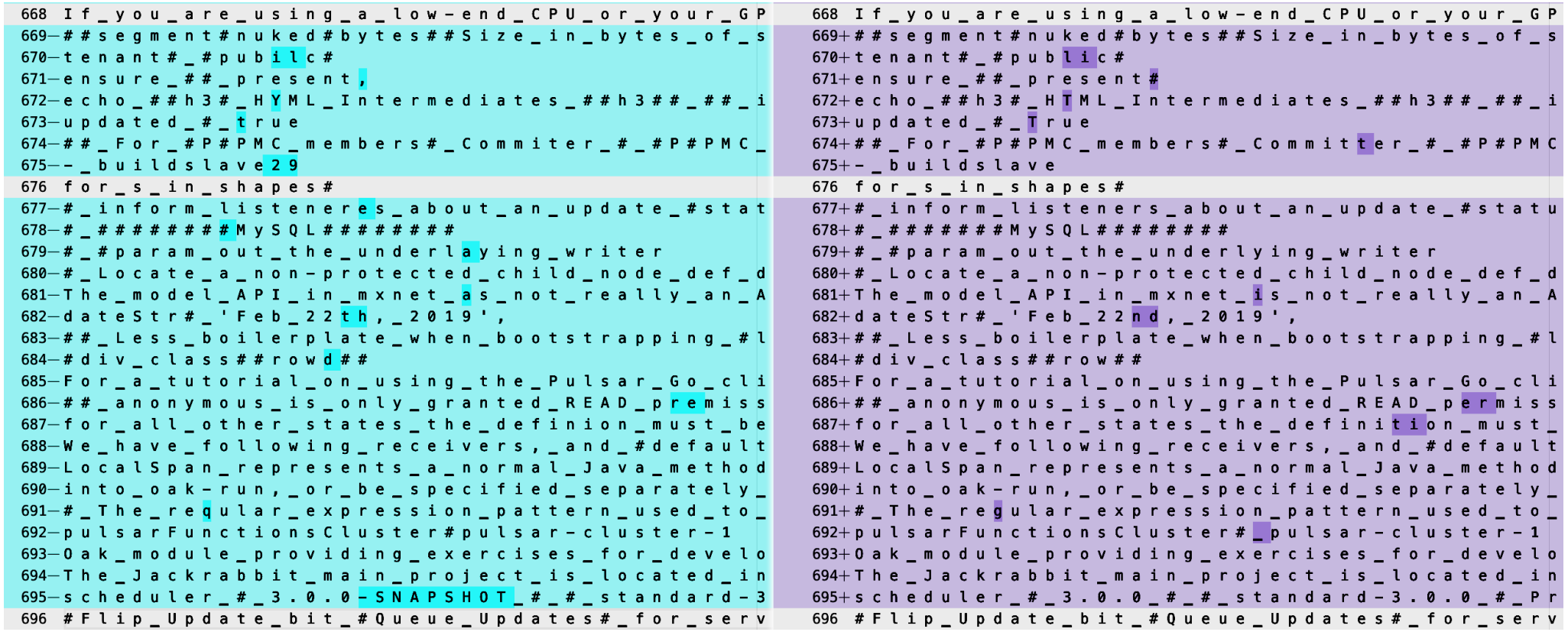
Трансформеры для задачи исправления опечаток

Для обучения модели, которую мы будем обсуждать, использовался большой размеченный корпус:

https://github.com/mhagiwara/github-typo-corpus

Трансформеры для задачи исправления опечаток

Модель обучается на буквах (то есть один токен – это один символ):



Искусственное расширение датасета

Мы можем искусственно расширить датасет, добавив опечаток. Но сложность состоит в том, чтобы опечатки были не случайными, а такими, какие обычно делают люди.

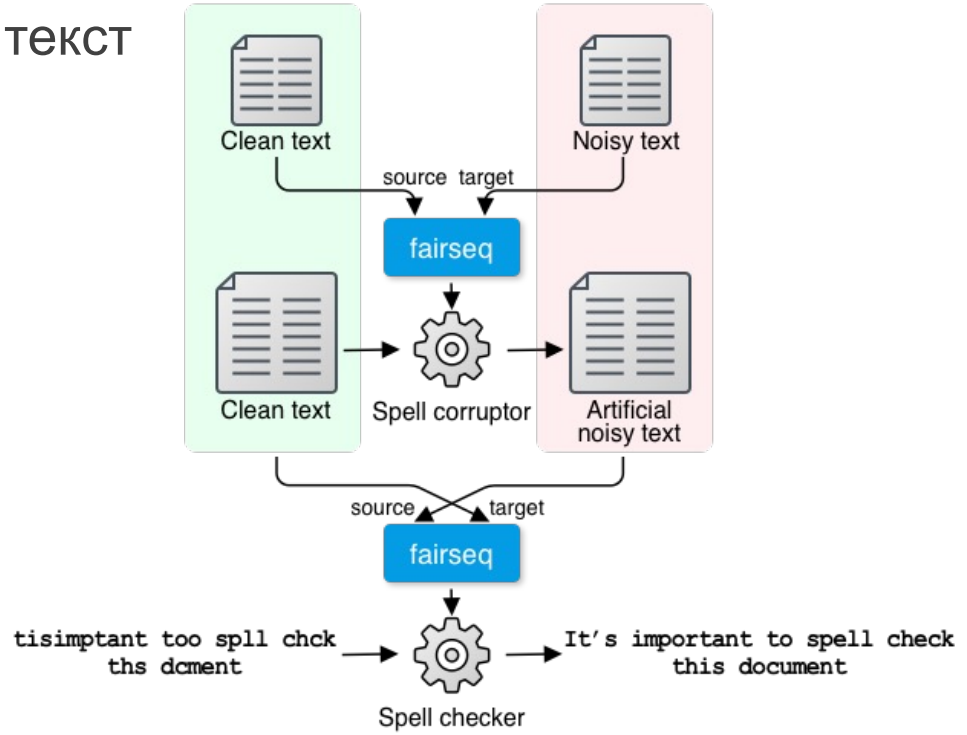
Идея:

- Обучим Spell Corruptor, поменяв местами Clean и Noisy текст

из наших размеченных данных.

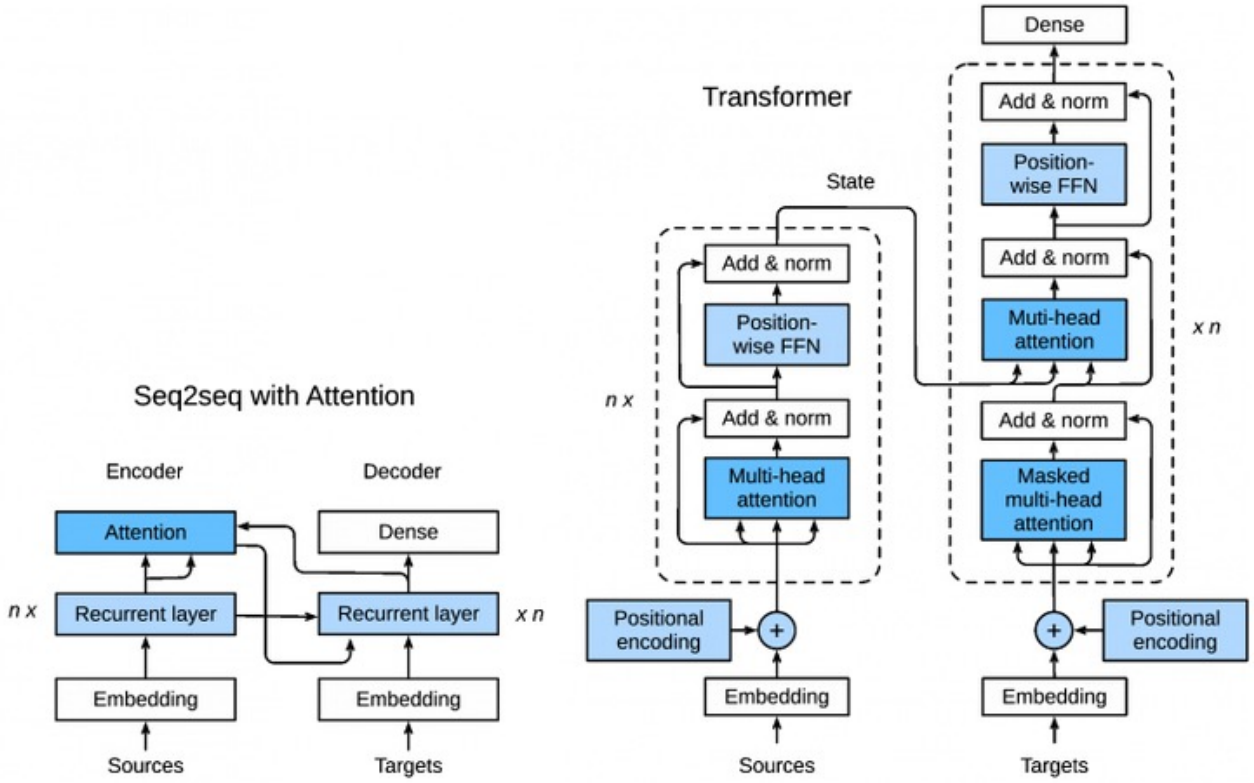
- Затем мы можем “испортить” любой чистый текст,

и получим дополнительные данные для обучения.

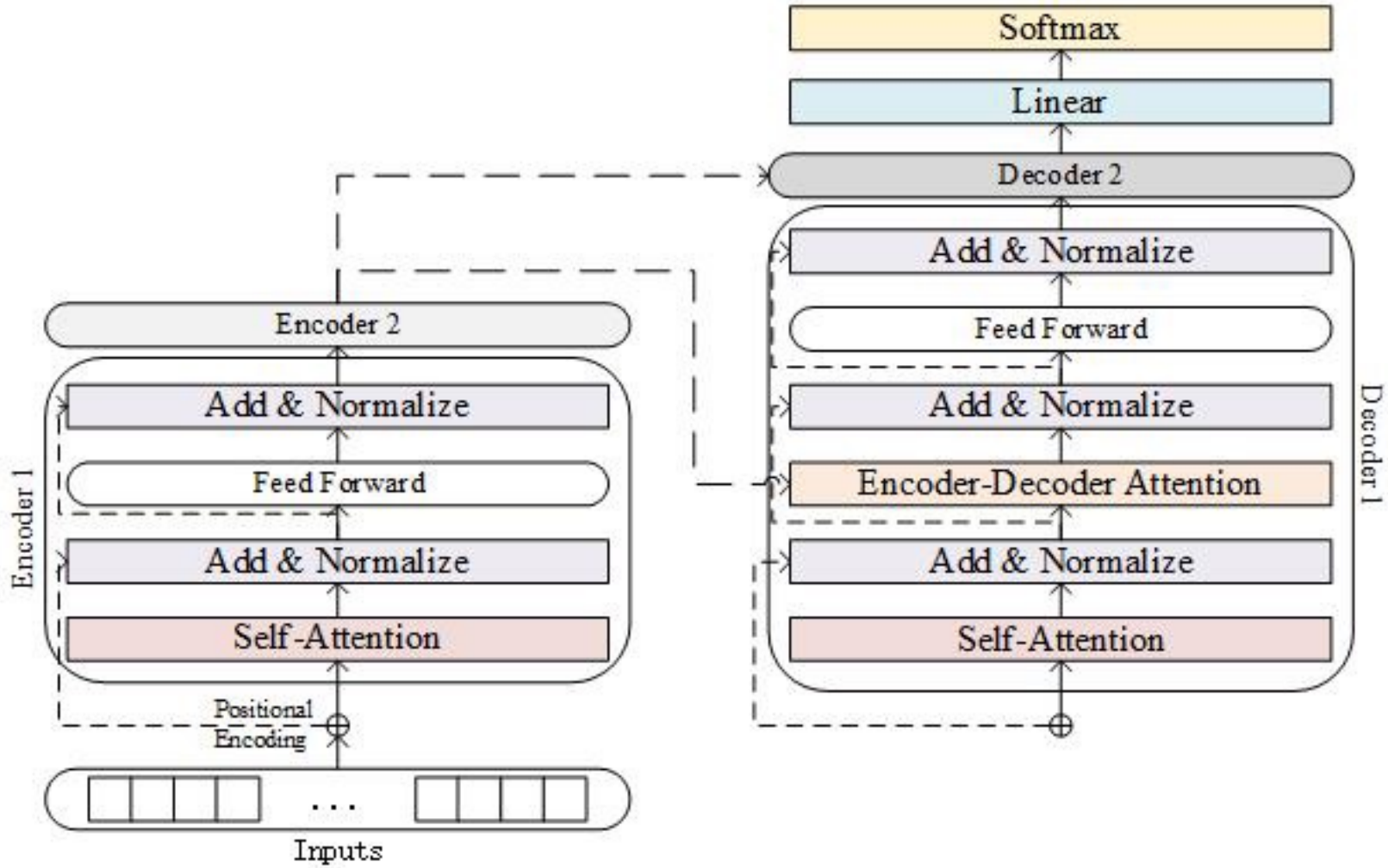


M2M100-модель

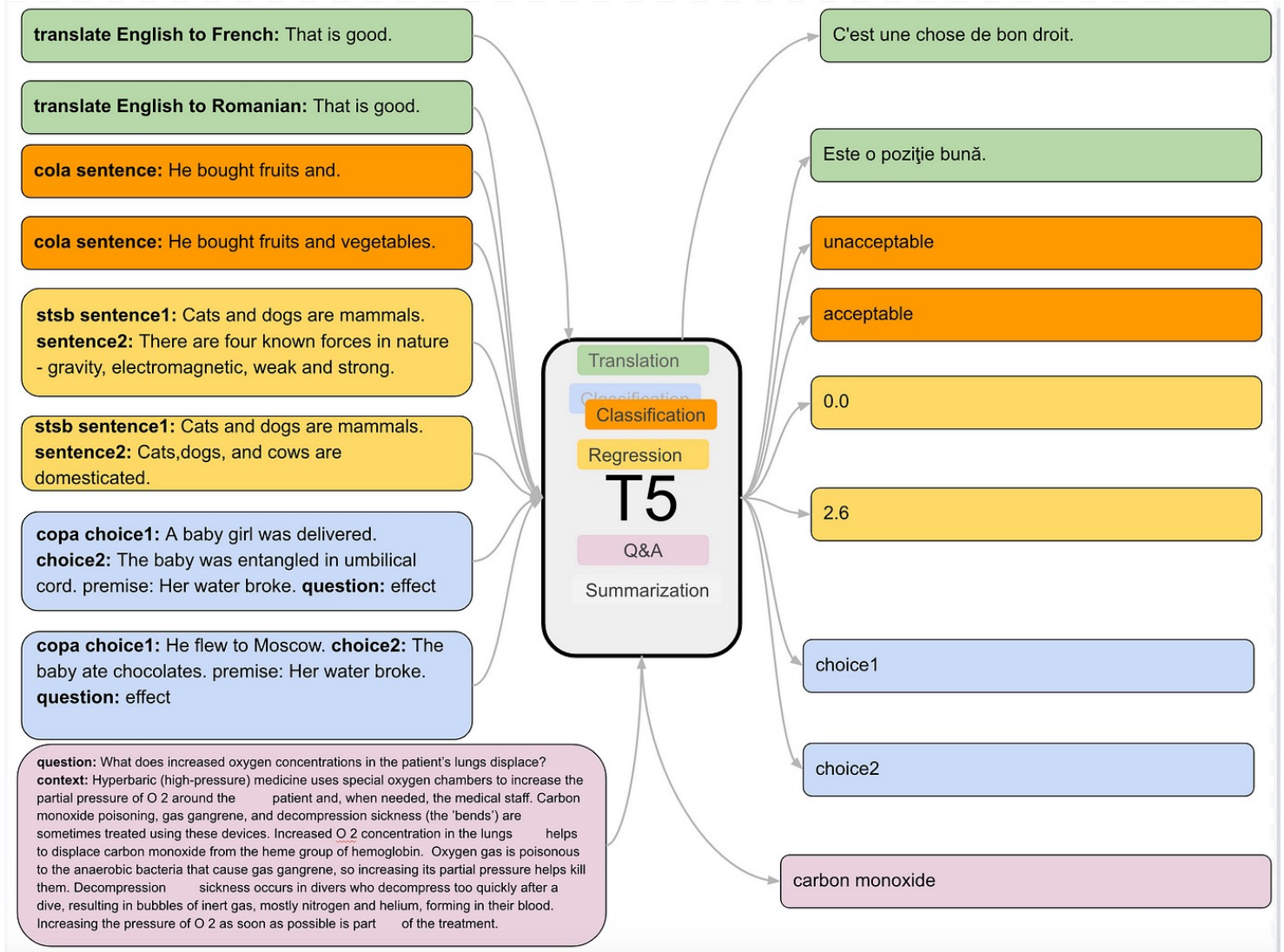
M2M100 is a multilingual encoder-decoder (seq-to-seq) model primarily intended for translation tasks. As the model is multilingual it expects the sequences in a certain format: A special language id token is used as prefix in both the source and target text. The source text format is [lang_code] X [eos], where lang_code is source language id for source text and target language id for target text, with X being the source or target text.



T5-модель


Softmax
Linear
Decoder 2
Add & Normalize
Feed Forward
Add & Normalize
Encoder-Decoder Attention
Add & Normalize
Self-Attention
Decoder 1
Encoder 2
Add & Normalize
Feed Forward
Add & Normalize
Self-Attention
Encoder 1
Positional Encoding
Inputs

T5-модель


translate English to French: That is good.
translate English to Romanian: That is good.
cola sentence: He bought fruits and.
cola sentence: He bought fruits and vegetables.
stsb sentence1: Cats and dogs are mammals. sentence2: There are four known forces in nature - gravity, electromagnetic, weak and strong.
stsb sentence1: Cats and dogs are mammals. sentence2: Cats,dogs, and cows are domesticated.
copa choice1: A baby girl was delivered. choice2: The baby was entangled in umbilical cord. premise: Her water broke. question: effect
copa choice1: He flew to Moscow. choice2: The baby ate chocolates. premise: Her water broke. question: effect
question: What does increased oxygen concentrations in the patient's lungs displace? context: Hyperbaric (high-pressure) medicine uses special oxygen chambers to increase the partial pressure of O 2 around the patient and, when needed, the medical staff. Carbon monoxide poisoning, gas gangrene, and decompression sickness (the 'bends') are sometimes treated using these devices. Increased O 2 concentration in the lungs helps to displace carbon monoxide from the heme group of hemoglobin. Oxygen gas is poisonous to the anaerobic bacteria that cause gas gangrene, so increasing its partial pressure helps kill them. Decompression sickness occurs in divers who decompress too quickly after a dive, resulting in bubbles of inert gas, mostly nitrogen and helium, forming in their blood. Increasing the pressure of O 2 as soon as possible is part of the treatment.
Translation
Classification
Regression
T5
Q&A
Summarization
C'est une chose de bon droit.
Este o poziţie bună.
unacceptable
acceptable
0.0
2.6
choice1
choice2
carbon monoxide

Библиотеки для Spell Correction

Статистические подходы к исправлению опечаток:

- TextBlob
- Pyspellchecker (Norwig model)
- Spylls (N-gram model)

Языковые модели для исправления опечаток:

- JamSpell
- SpaCy
- Модели из HuggingFace

Практика! Библиотеки для Spell Correction

- https://colab.research.google.com/drive/1N3UK004WVR1-vhF4GP35ms06L4blhHaM?usp=sharing
- https://colab.research.google.com/drive/1dd-IcVXXY1OrTtB5xKPF_cZgP0Nl1Q7l?usp=sharing