README

<2024-06-05 Wed>

Outline

;-- mode: Org; fill-column:120; org-src-fontify-natively:t; --

```
#+BEAMER_HEADER: \usepackage[orientation=landscape,size=custom,
#+BEAMER_HEADER: \usepackage[size=custom,width=30,height=30,sca
```

```
#+BEAMER_HEADER: \documentclass[aspectratio=169]{beamer}
```

Чем интересен хакатон X5 TECH AI HACK

- Можно видеть, что нейронные сети и языковые модели заменяют собой классические инструменты программирования, такие как регулярные выражения, Word2Vec и другие инструменты основанные на императивном анализе данных.
- облачные вычисления и сервисы являются 1) ресурсной базой для вычислений 2) обеспечивают централизованную безопасность

Поэтому маскирование приватных данных, поиск именнованных сущностей и управление языковыми моделями - это самые частые современные задачи в IT.

Маскирование: постановка

Вход: текст
Выход: этот текст с замененненными сущностями (телефоны, фамилии, адреса . . .) на похожие.
Дополнительно: иметь возможность обратной замены, устойчивой к взлому.

проблемы

- Маленький датасет с ошибками
- нет доступа к Интернету
- 8GB RAM, только CPU

Маскирование: Baseline

BERT English 110M параметров - чествительная к регистру

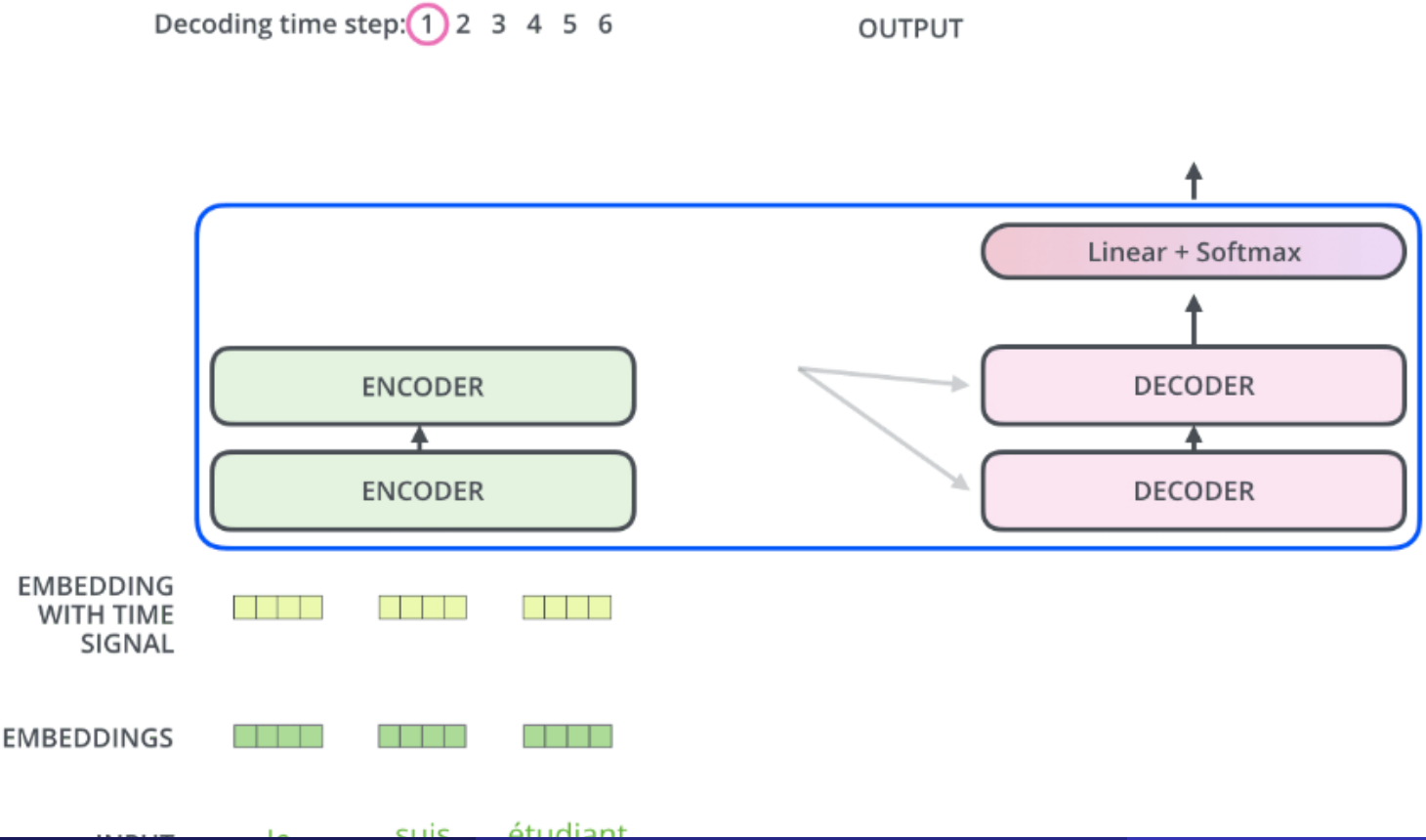
1. без токенайзера
2. Обучение NER - 400 epochs с 2e-5 lerning rate
3. неразмеченный текст подается модели посимвольно

текст разбивается на токены в 1 символ и помечается в BIO

BERT - что это?

BERT - языковая модель на основе Transformer на одном кодировщике.

- Вход - фиксированная строка, выход - фиксированная строка.
- Tokenizer с WordPiece - обученный отдельно.
- предобучен на Masked LM и Next Sentence Prediction (NSP)



Маскирование: простые решения

1. Использование слов, а не символов - предобученного токенизатора
2. Обучение Tokenizer на словах
3. Использование предобученной модели и токенайзера на русском корпусе
4. DataCollatorForTokenClassification вместо самописного
5. при обучени устранение дисбаланса классов

Маскирование: победившие решения

- использование bert-base-multilingual-cased
- регулярные выражения + LLM NER + поиск по словарю
  - найденные позиции помечаются
- xml-roberta-large-ner-russian
- удаление лишних пробелов знако пунктуцаии улучшает NER.

Маскирование: наше решение

- без дообучения DeepPavlov/ner_rus_bert + regex выражения

Результатирующая точность: 0.41 - низкий. Времени не хватило на выяснение причин.

```
link_pattern = r'https?://\w*\.\w*/'
phone_pattern = r"((8|\+7)[\- ]?)?(\(?\d{3}\)?[\- ]?)?[\d\- ]{7
email_pattern = r'\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]
date_pattern = r'\b\d{2}\.\d{2}\.\d{4}\b'
num_pattern = r'\b\w*[0-9]+\w*\b'
acr_pattern = r'\b[A-Z]{3}\b'
```

Галлюцинации: постановка

Вход: контектс, вопрос, ответ.
Выход: метка 0/1 ответ правильный или нет.
Дополнительно: сделать из решения качественный программный продукт.

Галлюцинации: Baseline

BERT English 110M параметров - нечувствительная к регистру

1. токенайзер - huggingface.TFBertTokenizer
2. дополнительный слой с выходом на 2 нейрона
3. loss = nn.CrossEntropyLoss() - бинарная классификация
   - Вход: "summary: ” | question: ” | answer: ”
   - Выход: следующее слово - метка

Галлюцинации: победившие решения

- [CLS] + summary + [SEP] + question + [SEP] + answer + [SEP].
- token_type_ids mask = 1 для ответа
- Стеккинг нескольких LLM и простой классификатор для объединения
- генерация датасета на базе RussianNLP/wikiomnia
- выделение признаков - сомнительно
- применение Saiga_8b_q4 и DeepPavlov/rubert-base-cased
- проверка выхода Baseline решения и добавление второй LLM

`https://huggingface.co/docs/transformers/glossary#token-type-ids`

Галлюцинации: наши решения

1. Saiga Llama3 8B + IPEX квантование - простой prompt engineering
2. Knewledge Distilation 0.902 - Малая модель учится повторять большую
   - cross-entropy loss function между парамтртризованным ответом учителя и студента
   - студент: cointegrated/rubert-tiny2
   - учитель: DeepPavlov/rubert-base-cased

a small model is trained to mimic a pre-trained, larger model (or ensemble of models)

Недостатки хакатонов

- датасеты с ошибками, нужно повторить ошибки чтобы победить
- организаторы дают свой подход и если не следовать ему это почти 100% самоубийство, так как временя ограничено
- заходить на хакатон нужно только с полной коммандой и в первые дни после объявления
- важна только скорость любой ценой, чем не контер страйк?
- в угоду скорости приходится жертвовать безопасностью, а это имеет долгосрочный характер.
- главная сложность это понять что вообще организаторы ожидают, что должно быть сделано.
- напряжения сил требуется для победы больше, что приз.
- залог победы - хорошая большая команда

Достоинства и возможности хакатонов

- найти команду и партнеров
- отбросить медленные неэффективные подходы
- попробовать командную работы
- узнать новое и современное
- узнать эффективные подходы от других команд

командная работа

- Общий чат без созвонов - один из лучших форматов.
- Любые напоминания о необходимости работать убивают желание работать.
- Письменный отчет каждый день о проделанной работе как средство проверки на бездельника. Но дополнительная нагрузка.
- Бездельникам нужно раздавать четкие задачи раньше
- Нет отчета - либо бездельник, либо загнал себя и не успевает.
- Правила которые ты ждешь от других лучше доносить персонально с подтверждением и всеми возможными вариантами событий.
- Со временем люди работают меньше, а не больше. Поэтому нужно оценивать по первичной работоспособности.
- Человек с пустым гитхаб аккаунтом не программист, а аналитик или ученый.

Допущенные ошибки

- Маленькая команда из недосаточно свободных людей
- Использование масштабных подходов с полой заменой Baseline
- Отсутствие подготовленного GPU у каждого в команде
- Дообучение и finetuning и ансамблирование, это главные навыки всех хакатонов, кооторыми нужно владеть в совершенстве

- предобработка текста для LLM улучшает качество
- можно использовать ансамбли из малых языковых моделей
- Knewledge distillation как эффективный метод дообучения малых языковых моделей
- галлюцинации это не факт чекинг.
- языковые модели эффективнее регулярных выражений, потому что на практике риск ошибки и взлома не критичен.