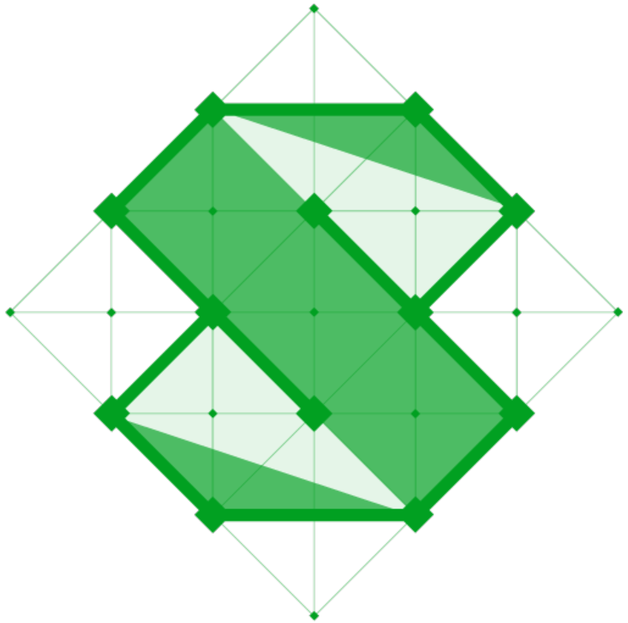


Sberbank

Data Science Journey

Обзор базовых решений задачи B

или

Как научить нейросеть отвечать на вопросы и выиграть 1,000,000 рублей

Алексей Натекин, Петр Ромов, Михаил Гавриков, Андрей Киселев

Задача B:

- SQuAD на русском:
  100,543 троек параграф-вопрос-ответ

- Формат:
  prediction скрипты
  решения запускаются в изолированном docker

- Призовой фонд задачи **1,400,000** рублей (+250,000)
  повод разобраться в DL NLP за 23+ дня

Пример данных:

paragraph_id: 14754

question_id: 60544

paragraph: Первые упоминания о строении человеческого тела встречаются **в** Древнем Египте. В XXVII веке до н. э. египетский врач Имхотеп описал некоторые органы и их функции, в частности головной мозг, деятельность сердца, распространение крови по сосудам. В древнекитайской книге Нейцзин (XI—VII вв. до н. э.) упоминаются сердце, печень, лёгкие и другие органы тела человека. В индийской книге Аюрведа (Знание жизни , IX-III вв. до н. э.) содержится большой объём анатомических данных о мышцах, нервах, типах

question: Где встречаются первые упоминания о строении человеческого тела?

answer: в Древнем Египте

Simple baseline:

- Идея: в качестве ответа выбирается целое предложение из параграфа, такое что оно сильнее всего пересекается со словами из вопроса

F1: 0.25108

https://github.com/sberbank-ai/data-science-journey-2017/tree/master/problem_B/simple-baseline

Simple ML baseline:

- Идея решения:
  1. Генерируем кандидатов для каждой пары вопроса/параграфа
  2. Строим модель, которая по кандидату предсказывает F_1 score
  3. Применение модели: arg max по всем кандидатам, которые есть в параграфе
  4. Используются максимально простые фичи

- Идея улучшений:
  1. Добавить POS
  2. Добавить word2vec + tfidf

F1: 0.31896

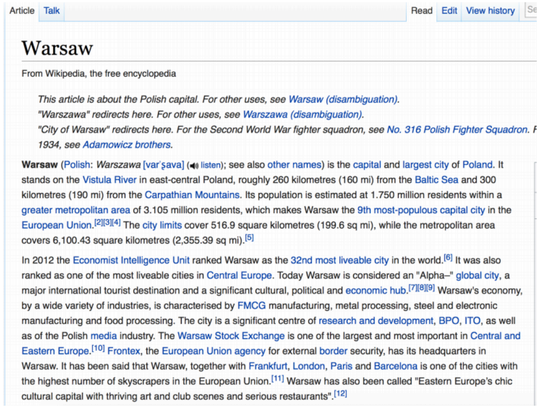
https://github.com/sberbank-ai/data-science-journey-2017/tree/master/problem_B/ml-baseline

DrQA baseline:

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

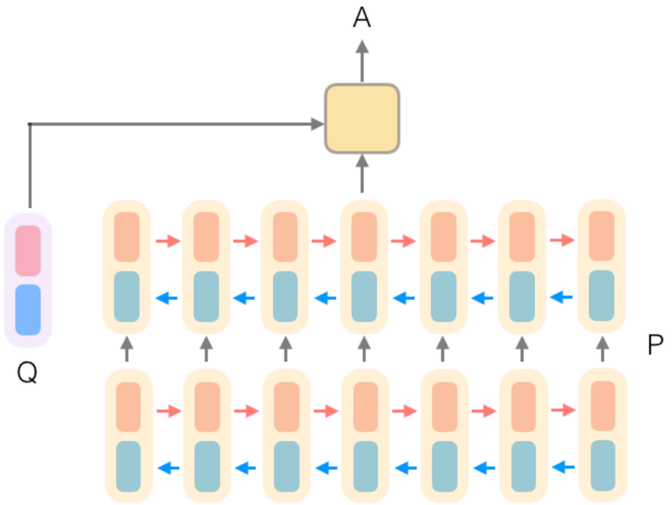
WIKIPEDIA
The Free Encyclopedia

Document Retriever

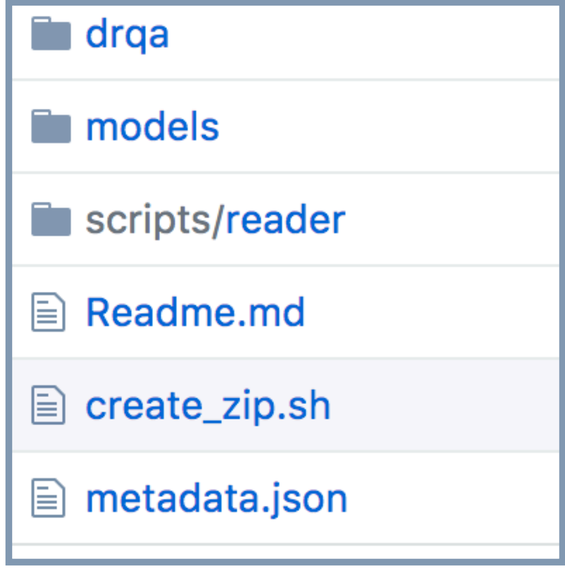


Document Reader

833,500



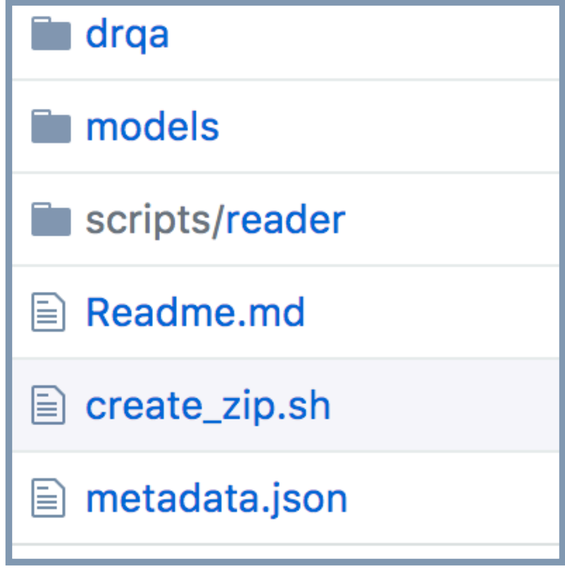
DrQA baseline:



Обучение модели

1. Скачайте файл с векторным представлением слов, готовые модели вы можете найти на сайте http://rusvectores.org/ru/models/ или выполните `scripts/reader/download_w2v.sh`
2. Трансформируйте файл с представлениями слов в текстовый формат: для этого выполните все ячейке в ноутбуке scripts/reader/BinaryW2VToSpaceSepartor.ipynb
3. Конвертируйте данные в формат, подходящий для обучения: `PYTHONPATH=.:$PYTHONPATH python3 scripts/reader/preprocess.py --tokenizer SimpleTokenizer train.csv data/datasets/output_filename.json`
4. Разделите файл на обучающую выборку и валидационную
5. В scripts/reader/train.sh вы можете найти пример запуска обучения модели
6. После обучения можете делать сабмит: [scripts/reader/train.sh](`sh create_zip.sh`) положит все необходимые файлы (убедитесь, что среди них есть модель, если вы переименовали модель незабудьте)
7. Также вы можете запустить сессию в интерактивном режиме `PYTHONPATH=.:$PYTHONPATH python3 scripts/reader/interactive.py --model models/20171007-1ce20c3f.mdl`

DrQA baseline:



Параметры обученной модели

Текст разбивается на токены с помощью простейшего регулярного выражения (см `drqa/reader/simple_tokenizer.py`) Все слова приводятся к леммам с помощью `pymorphy2`, переводятся в lowercase и кодируются соответствующими `word2vec`-представлениями. Информация о частях речи, именованных сущностях и т.д. не используется.

Слова, которых нет в предобученном `word2vec` игнорируются.

В качестве валидационной метрики используется `exact_match` - число полностью верных ответов на вопросы.

На этапе применения модели учитываются только те слова, что встречались в тренировочных данных.

F1: 0.59148

https://github.com/kiselev1189/SberBDrQAReader в ближайшее время появится на платформе

P.S.

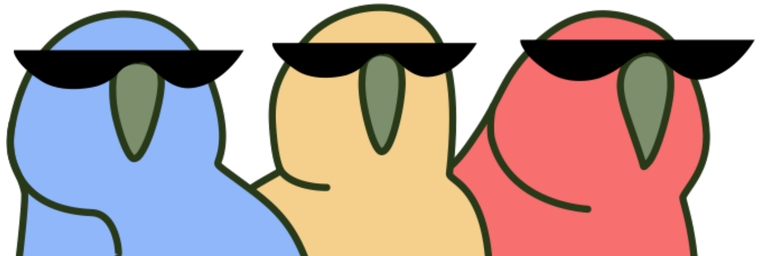
- Добавили pymystem3, pymorphy2[fast], tqdm
- Лимиты увеличены до 120 минут и 20 минут на check (5 предсказаний)
- Много идей для улучшений DrQA baseline: дообучить и отвалидировать, настроить гиперпараметры и архитектуру, поправить w2v, дообучить w2v на топ-N слов, добавить pos и другие фичи оригинальной статьи, …

Summary

- DrQA достаточно хорошо работает в самом наивном и не оптимизированном варианте
- Лучшие решения будут использовать DL, повод разобраться в DL для NLP
- Вопросное-ответные системы это круто и можно самим играться с ними (в т.ч. с оригинальным retriever DrQA)

Спасибо!

@natekin
@peter
@gavrmike
@kiselev1189



contest.sdsj.ru