https://github.com/ivanrybin/kbqa_itmo_2020_summer

Летняя практика:
20 июля - 26 августа

Модификация вопросно-ответных датасетов на русском языке

Рыбин И. С.

Научный руководитель: к.т.н., к.культ.н. П.И. Браславский

Факультет информационных технологий и программирования
Магистерская программа "Разработка ПО / Software Engineering"

ИТМО
Санкт-Петербург
2020

Вопросно-ответные датасеты (QA-датасеты)

Предметная область: обработка естественного языка (NLP).
Используются в: классе задач reading comprehensions.

QA-датасет состоит из: вопроса, ответа и служебной информации для обеспечения свойств датасета.

Свойства QA-датасетов:

1. наличие привязки к структурированной базе знаний (например, Wikidata), (выделенные сущности и отношения, наличие SPARQL-запроса)
2. наличие текстов (или ссылок на них), доказывающих правильность ответа
3. сложность вопросов (число задействованных в нем сущностей и отношений)

Важные особенности:

- QA-датасет без текстового материала ограничен для обучения моделей
- выполнение свойств 1 и 2 одновременно редко в существующих датасетах

Примеры датасетов

RuBQ (2020) (В. Кораблинов, П.Браславский) - 1.500 элементов

состоит из триплетов (вопрос, ответ, SPARQL-запрос к базе знаний)

- нет текстового доказательства ответа
- есть привязка к базе знаний
- гарантирует корректность ответа базой знаний

TriviaQA (2017) - 95.000 элементов

состоит из триплетов (вопрос, ответ, ~6 ссылок на web-страницы с ответами)

- есть текстовое доказательство ответа
- нет привязки к базе знаний
- гарантирует наличие ответа по ссылкам

SQuAD (2016) - 100.000 элементов

состоит из триплетов (вопрос, ответ, несколько абзацев текста)

- есть текстовое доказательство ответа
- нет привязки к базе знаний
- не гарантирует наличие ответа в представленных абзацах текста

Результаты предыдущего проекта, цели и задачи

Два датасета, связанных с базой знаний Wikidata:

- 1400 вопросов - И.Рыбин (весна 2020)
- 1500 вопросов - В.Кораблинов
- элементы обоих вида:
  (сущность вопроса Q1, свойство P, сущность ответа Q2)

Основная цель: добавить к обоим датасетам абзацы текста с ответами

Дополнительная цель: попробовать произвести генерацию сложных вопросов

Задачи:

- доработать весенний датасет через добавление вопросов без ответа
- определить метод и место поиска абзацев текстов с ответами
- получить и верифицировать абзацы-кандидаты
- сопоставить абзацы с датасетами
- определить возможные способы генерации сложных вопросов
- произвести генерацию сложных вопросов на существующих датасетах

Поиск абзацев текста с ответами

Схема реализации:

- сущности вопросов связаны с Wikidata
- сущности Wikidata связаны со статьями Википедии
- статьи Википедии состоят из абзацев текста

  Wikidata (qid) -> Википедия (link) -> Текст -> Абзацы -> Абзацы с ответом

- получаем дамп страниц сущностей вопроса Q1
- ведем поиск ответов на этих страницах

Вопрос - как автоматически искать?

Методы автоматического поиска

Методы, выбранные для возможного решения задачи:

1. поиск по семантической близости вопроса и абзаца текста
2. поиск по ссылке сущности ответа в статье Википедии
3. строковый поиск вхождений комбинаций вопроса и ответа в текст

Конкретный метод связан с представлением статей Википедии:

1. Plaintext
2. HTML
3. Plaintext + Wikitext

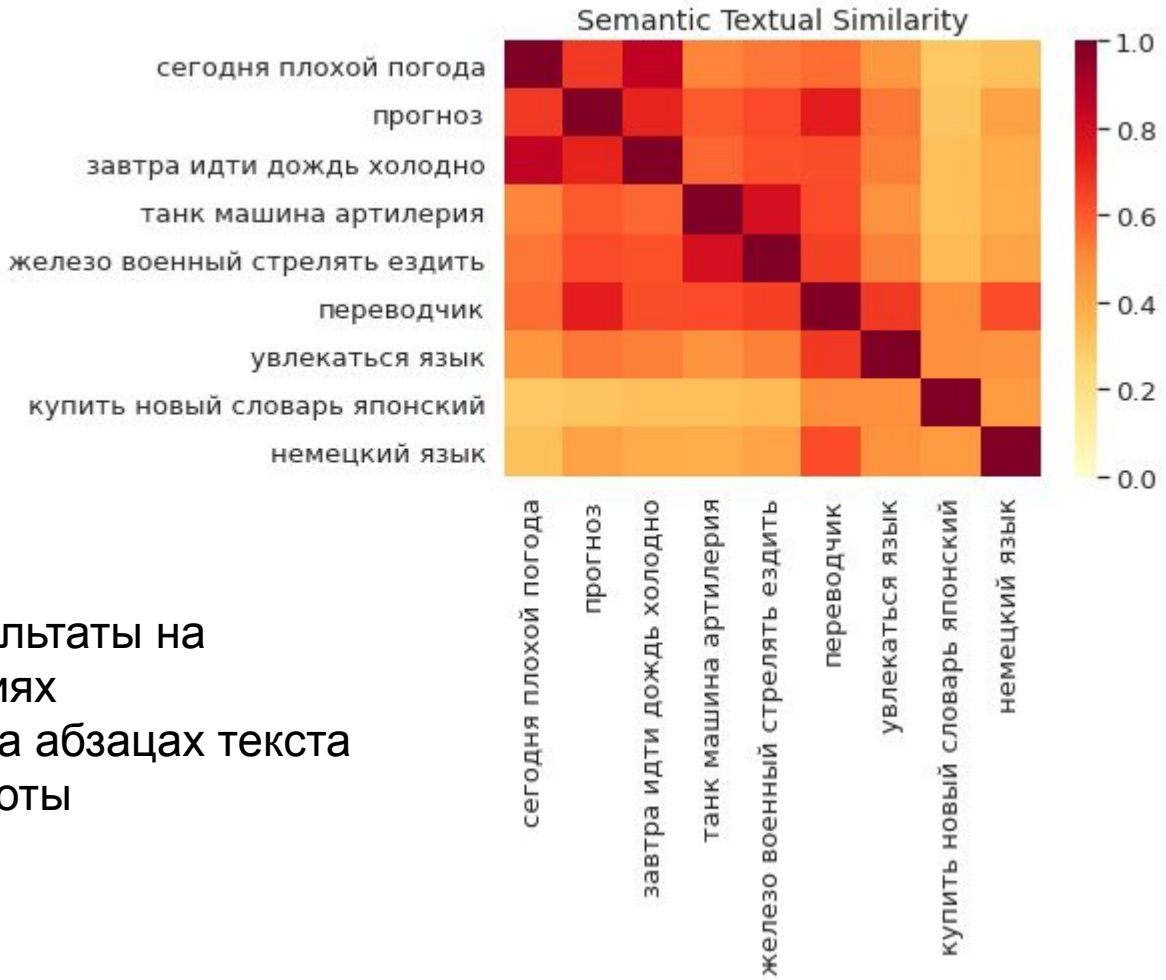
Первые проблемы:

- не все сущности Wikidata имеют статью Википедии на русском языке

Методы автоматического поиска

1. Поиск по семантической близости (библиотека TensorFlow)

Суть: матрица семантической близости вопроса и абзацев текста, выбирается абзац с наибольшим коэффициентом



Проблемы:

- показал неплохие результаты на отдельных предложениях
- оказался непригоден на абзацах текста
- длительное время работы

(отклонен)

2. Поиск по ссылке сущности ответа:

Суть: упоминание объекта в статье Википедии сопровождается ссылкой на статью этого объекта

Проблемы:

-ссылка есть только у первого вхождения сущности
-абзац с первым упоминанием не обязательно отвечает на вопрос
-найдя абзац в HTML трудно конвертировать его в чистый текст

(отклонен)

3. Поиск по строковому вхождению слов вопроса и ответа в текст

Суть: проводим стемминг и очистку абзацев, вопросов и ответов, ищем абзацы с наибольшим вхождением слов вопроса и ответа в plaintext представлении статей, разбив перед этим их на секции и абзацы

Проблемы:

- датасет не включает в себя иные названия/имена сущностей, по которым ведется поиск

Решение - собираем все псевдонимы из:

- Wikidata
- из wikitext представления статей Wikipedia, которое хранит текстовые ссылки на любые формы упоминания сущностей в тексте вида:

[[Пушкин|А.С. Пушкин]]
[[Александр|А.С. Пушкин]]
[[молодой поэт|А.С. Пушкин]]

Верификация и результаты

Верификация через Яндекс.Толока:

- создаем интерфейс, задание, обучение, инструкцию, пулы на Толоке
- задание вида: вопрос и 1-4 абзаца текста, в которых может быть ответ
- надо выбрать абзацы,в которых есть ответ (не зная реального ответа)
- один вопрос показывается 3 толокерам
- абзац считается хорошим, если за него проголосовало 2-3 человека

Результаты

- датасет из 57.000 абзацев, его линковка с вопросами двух датасетов

	датасет И. Рыбин	датасет В. Кораблинов
всего элементов	**1410**	**1500**
покрыты хорошими абзацами	**1076 (76%)**	**1055 (70%)**
не покрыты	**334 (24%)**	**445 (30%)**
среднее число абзацев на вопрос	24	19
среднее число хороших абзацев на вопрос	**1.64**	**1.76**

Генерация вопросов

Каждый элемент датасета имеет id сущности ответа, вопроса и отношения, которое их связывает

На основе логических, числовых и иных комбинаций с использованием id можно формально выразить более сложные формы вопросов

Из двух статей ([1] и [2]) были выбраны две доступные для реализации на имеющихся датасетах схемы:

1. Дополнение одного вопроса другим

2. Конъюнкция двух вопросов по ответу

Генерация вопросов

Пусть:
(x, P, y) - вопрос, в котором сущность вопроса x связана с ответом y отношением P

1. Дополнение одного вопроса другим:
   (x, P1, y) + (z, P2, x)

   Пример:
   Как называется денежная единица Макао?
   \+ Какой стране принадлежит домен высшего уровня .mo? (ответ Макао)
   = Как называется денежная единица страны, которой принадлежит домен высшего уровня .mo?

2. Конъюнкция двух вопросов по ответу:
   (x, P1, y) + (z, P2, y)

   Пример:
   Где образовалась группа Кино? (Петербург)
   \+ Где погребен Кутузов? (Петербург)
   = Где образовалась группа Кино и погребен Кутузов?

Генерация вопросов: результаты

Результаты по генерации:

датасет И.Рыбин
- **259** конъюнкции (из них только 43 с разными отношениями)
- **90** дополнения из них 84 с разными отношениями)

датасет В.Кораблинов
- **432** конъюнкции (из них только 192 с разными отношениями)
- **178** дополнения (из них 155 с разными отношениями)

Генерация не обработана человеком, вопросы выглядят так:

“На чем играл Моцарт, **[который, кто]** сочинил симфонию Юпитер?”
“Где жил Пушкин **AND** погребен Кутузов?”

Задел на будущее

- расширение датасетов путем подстановки в существующие вопросы сущностей из того же класса, что и сущность вопроса

- генерация иных и более сложных форм вопросов

- приведение сложных вопросов в естественную форму