

Отбор релевантных предложений в задаче построения вопросно-ответных систем

Сайранов Данил

Научный руководитель: д.ф.-м.н., профессор В. А. Серебряков

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Факультет управления и прикладной математики
Кафедра прикладных проблем теоретической и математической физики

Москва, 2018

Задача

Предложить алгоритмы выделения предложений, которые могут содержать ответ на вопрос, в задаче построения вопросно-ответных систем.

Проблема

В реальных задачах документы чаще всего представляют собой достаточно большие тексты (более 10 предложений).

Методы решения

- Построить алгоритм, который учитывает тип вопросов и генерирует ответ, соответствующий этому типу
- Построить алгоритм, который ранжирует предложения относительно вопросов.

Дано

- Пары (D_i, q_j) , где D_i - i -й документ, q_j - j -й вопрос.

Найти

- Документы $D_i^* : |D_i^*| < |D_i|$ и в D_i^* содержится ответ на вопрос q_j

Схема исходной системы

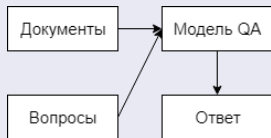
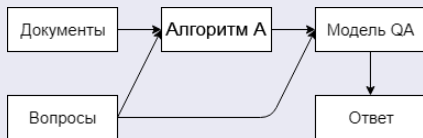
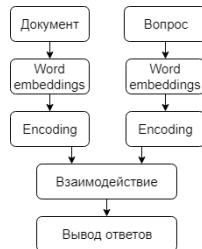


Схема модифицированной системы



Этапы работы

- На вход подаются документ $D = \{w_i\}_{i=1}^n$ и вопрос $q = \{q_j\}_{j=1}^m$, где w_i - i -е слово в документе D , q_j - j -е слово в вопросе.
- Каждое слово w_i в документе и q_j в вопросе заменяется его векторным представлением \vec{w}_i и \vec{q}_j соответственно.
- Построение контекстно-зависимого вектора для каждого вектора-слова
- Обогащение полученных векторов документа информацией о вопросе
- Вывод ответа на вопрос



Архитектура
нейросетевых
вопросно-ответных
систем

Гипотеза

Предложение, сгенерированное только на основе вопроса, по типу и структуре похоже на предложения в документе, которые могут содержать ответ.

Примеры:

- **In the range between 1980 and 1990, what did demand for grow?**

Throughout the 1980s and 1990s, demand for a **Scottish Parliament** grew, in part because the government of the United Kingdom was controlled by the Conservative Party, while Scotland itself elected relatively few Conservative MPs.

- **What doesn't change from being at rest to movement at a constant velocity?**

For instance, while traveling in a moving vehicle at a constant velocity, the **laws of physics** do not change from being at rest.

Этапы решения:

- Предобработка данных.
- Обучение генератора на предобработанных данных.
- Генерация предложения S на основе полученного на вход вопроса q .
- Сравнение сгенерированного предложения S с каждым предложением из документа D .
- Выбор k -наиболее близких к S предложений.

Данные

- В качестве обучающих выборок для генератора взяты датасеты MS Marco, SQuAD, SelQA.
- В качестве данных для проверки качества работы генератора взят датасет SQuAD.

Примеры пар "вопрос - предложение, содержащее ответ"

- MS Marco:

What is the Meiji emperors name?

- The Meiji emperor's name is Mutsuhito.

What does subside mean?

- Subside means to sink to a low or lower level.

- SQuAD:

When was levi's stadium awarded the right to host super bowl 50?

- On may 21, 2013, nfl owners at their spring meetings in boston voted and awarded the game to levi's stadium

- SelQA:

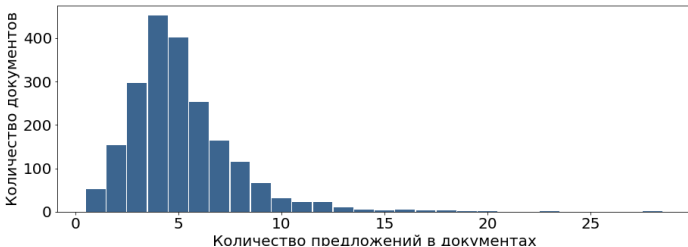
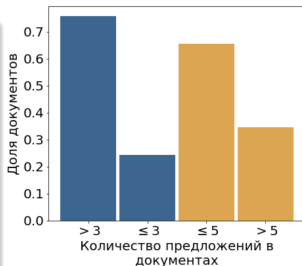
In what year was the chilean national museum of fine arts built?

- Museums in chile such as the chilean national museum of fine arts built in 1880 feature works by chilean artists

Информация о датасете SQuAD

Доля документов с количеством предложений

- не более 3: 25%
- не более 5: 65%
- более 3: 75%
- более 5: 35%



Генератор ответов на вопрос

- Архитектура сети:
 - BiRNN + dotAttention
 - GloVe embeddings

Ранжирование предложений

Для ранжирования предложений в документе относительно сгенерированного ответа применим следующие алгоритмы:

- TF-IDF
- InferSent + cosine

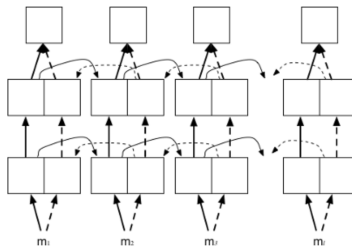


схема BiRNN

Примеры сгенерированных предложений

When is groundhog day?

- A groundhog day is on may 28, 2016.

What is the smallest us state in terms of land area?

- Alaska is the smallest us state in terms of land area.

What is nightwatch?

- Nightwatch is a type of nightwatch what what and what

Definition of sacrament?

- The definition of sacrament is definition

Should you eat before or after a workout?

- You should eat before a workout.

$$\text{Recall}@k = \frac{tp}{total},$$

где tp – число вхождений предложений, содержащих ответ, в k -наилучших предложениях, $total$ – общее число предложений, содержащих ответ.

| Выборка | Recall@1 | Recall@3 | Recall@5 |
|-----------------|----------|----------|----------|
| SQuAD | 0.567 | 0.888 | 0.971 |
| MS Marco | 0.660 | 0.918 | 0.980 |
| SelQA | 0.278 | 0.703 | 0.910 |

Качество работы генератора, обученного на различных выборках

Ранжирование предложений в документах

Для уменьшения размера документов обучается ранжирующий алгоритм, который сортирует предложения в документе на основе их соответствия вопросу.

Этапы решения

- Обучение алгоритма на парах «вопрос - предложение, содержащее ответ».
- Поддача на вход алгоритма пары (D_i, q_j) .
- Вывод отсортированного по релевантности списка предложений из документа.

Данные

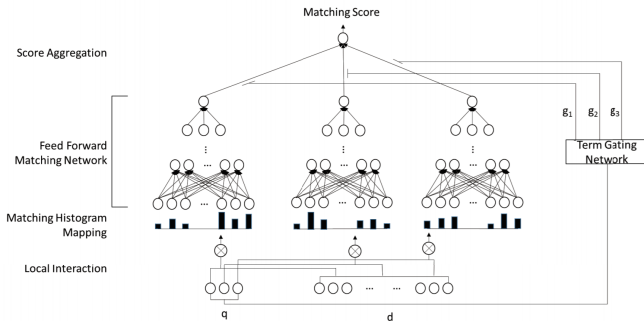
- Датасет SQuAD

Дополнительно

TF-IDF, InferSent

Ранжирование предложений относительно вопроса

- Архитектура сети: MatchZoo DRMM



DRMM (Deep Relevance Matching Model)

Качество уменьшения документов определяется значением Recall at k

$$\text{Recall}@k = \frac{tp}{total},$$

где tp – число вхождений предложений, содержащих ответ, в k -наилучших предложениях, $total$ – общее число предложений, содержащих ответ.

| Алгоритм | Recall@1 | Recall@3 | Recall@5 |
|------------------|----------|----------|----------|
| TF-IDF | 0.627 | 0.882 | 0.962 |
| Infersent | 0.682 | 0.920 | 0.982 |
| DRMM | 0.423 | 0.510 | 0.561 |

Качество ранжирующего алгоритма

| Алгоритм | Exact Match | F1 |
|--------------------------------|-------------|--------|
| R-Net (без модификаций) | 70.9% | 79.64% |
| Генератор (SQuAD) | 62.4% | 71.2% |
| Генератор (MS Marco) | 65.1% | 74.2% |
| Генератор (SelQA) | 44.4% | 52.2% |
| TF-IDF | 63.9% | 72.5% |
| InferSent | 62.1% | 70.7% |
| DRMM | 62.7% | 71.2% |

Качество работы системы R-Net на уменьшенных документах при выборе 3 наиболее релевантных предложений

- Подтверждена гипотеза о схожести сгенерированных ответов с предложениями, содержащими ответ на вопрос.
- Ранжирование предложений в документе относительно вопроса выгоднее, чем ранжирование предложений относительно сгенерированных ответов.
- Несмотря на небольшое понижение качества работы вопросно-ответной системы R-Net, алгоритмы могут быть использованы для уменьшения достаточно больших документов.