Reading Wikipedia to Answer Open-Domain Questions, ACL, 2017, Danqi Chen, Adam Fisch, Jason Weston, Antoine Bordes

https://qdata.github.io/deep2Read

Presenter: Chao Jiang

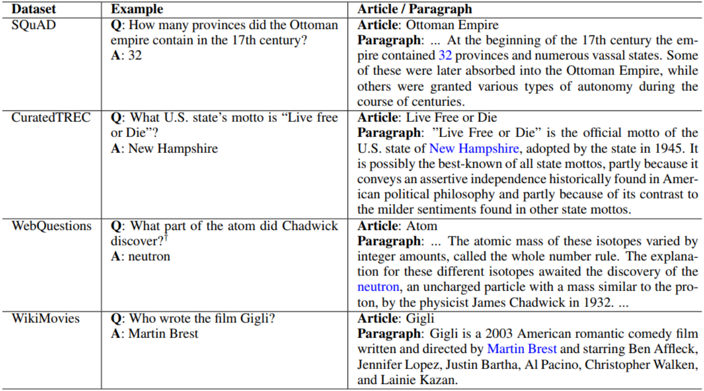
Spring 2018

Outline

1. Task
2. Method
   - Document Retriever
   - Document Reader

Task

Opendomain question answering using Wikipedia as the unique knowledge source: the answer to any factoid question is a text span in a Wikipedia article.



Dataset	Example	Article / Paragraph
SQuAD	**Q**: How many provinces did the Ottoman empire contain in the 17th century? **A**: 32	**Article**: Ottoman Empire **Paragraph**: ... At the beginning of the 17th century the empire contained 32 provinces and numerous vassal states. Some of these were later absorbed into the Ottoman Empire, while others were granted various types of autonomy during the course of centuries.
CuratedTREC	**Q**: What U.S. state's motto is "Live free or Die"? **A**: New Hampshire	**Article**: Live Free or Die **Paragraph**: "Live Free or Die" is the official motto of the U.S. state of New Hampshire, adopted by the state in 1945. It is possibly the best-known of all state mottos, partly because it conveys an assertive independence historically found in American political philosophy and partly because of its contrast to the milder sentiments found in other state mottos.
WebQuestions	**Q**: What part of the atom did Chadwick discover?† **A**: neutron	**Article**: Atom **Paragraph**: ... The atomic mass of these isotopes varied by integer amounts, called the whole number rule. The explanation for these different isotopes awaited the discovery of the neutron, an uncharged particle with a mass similar to the proton, by the physicist James Chadwick in 1932. ...
WikiMovies	**Q**: Who wrote the film Gigli? **A**: Martin Brest	**Article**: Gigli **Paragraph**: Gigli is a 2003 American romantic comedy film written and directed by Martin Brest and starring Ben Affleck, Jennifer Lopez, Justin Bartha, Al Pacino, Christopher Walken, and Lainie Kazan.

Table 1: Example training data from each QA dataset. In each case we show an associated paragraph where distant supervision (DS) correctly identified the answer within it, which is highlighted.

System Overview

DrQA

1. Document Retriever, a module using bigram hashing and TF-IDF matching designed to, given a question, efficiently return a subset of relevant articles
2. Document Reader, a multi-layer recurrent neural network machine comprehension model trained to detect answer spans in those few returned documents.

System Overview

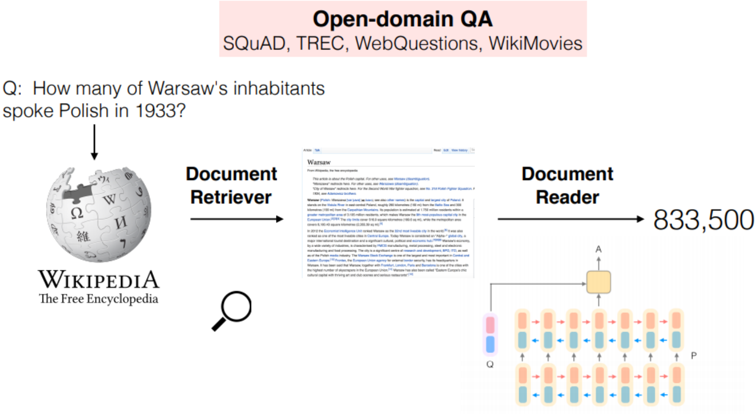

Open-domain QA
SQuAD, TREC, WebQuestions, WikiMovies
Q: How many of Warsaw's inhabitants spoke Polish in 1933?
WIKIPEDIA
The Free Encyclopedia
Document Retriever
Warsaw
Document Reader
833,500
A
Q
P

Figure 1: An overview of our question answering system DrQA.

Outline

1. Task

2. Method
   - Document Retriever
   - Document Reader

Document Retriever

- Document Retriever is the first part of their full model, by setting it to return 5 Wikipedia articles given any question. Those articles are then processed by Document Reader.
- Articles and questions are compared as TF-IDF weighted bag-of-word vectors.
- They further improve our system by taking local word order into account with n-gram features.

Outline

1. Task

2. Method
   - Document Retriever
   - Document Reader

Document Reader

Given a question q consisting of l tokens $\{q_1, \cdots, q_l\}$ and a document or a small set of documents of n paragraphs where a single paragraph p consists of m tokens $\{p_1, \cdots, p_m\}$, they develop an RNN model to predict two ends of the answer span