Reading Wikipedia to Answer Open-Domain Questions

2017 • ACL 2017 • ACL • AI • Dataset • Machine Comprehension • NLP • QA

15 Oct 2017

Introduction

- The paper presents a new machine comprehension dataset for question answering in real life setting (say when interacting with Cortana/Siri).
- Link to the paper

Unique Aspects of the dataset

- Existing machine comprehension (MC) datasets are either too small or synthetic (with a distribution different from that or real-questions posted by humans). MARCO questions are sampled from real, anonymized user queries.
- Most datasets would provide a comparatively small and clean context to answer the question. In MARCO, the context documents (which may or may not contain the answer) are extracted using Bing from real-world documents. As such the questions and the context documents are noisy.
- In general, the answer to the questions are restricted to an entity or text span within the document. In case of MARCO, the human judges are encouraged to generate complete sentences as answers.

Dataset Description

- First release consists of 100K questions with the aim of releasing 1M questions in the future releases.
- All questions are tagged with segment information.

- A subset of questions has multiple answers and another subset has no answers at all.

- Each record in the dataset contains the following information:

    - **Query** - The actual question
    - **Passage** - Top 10 contextual passages extracted from web search engine (which may or may not contain the answer to the question).
    - **Document URLs** - URLs for the top documents (which are the source of the contextual passages).
    - **Answer** - Answer synthesised by human evaluators.
    - **Segment** - Query type, description, neumeric, entity, location, person.

Experimental Results

- Metrics

    - Accuracy and precision/recall for numeric questions
    - ROGUE-L/paraphrasing aware evaluation framework for long, textual answers.

- Among generative models, Memory Networks performed better than seq-to-seq.

- In the cloze-style test, ReasoNet achieved an accuracy of approx. 59% while Attention Sum Reader achieved an accuracy of approx 55%.

- Current QA systems (including the ones using memory and attention) derive their power from supervised data and are very different from how humans do reasoning.

- Imagenet dataset pushed the state-of-the-art performance on object classification to beyond human accuracy. Similar was the case with speech recognition dataset from DARPA which led to the advancement of speech recognition. Having a large, diverse and human-like questions dataset is a

fundamental requirement to advance the field and the paper aims to provide just the right kind of dataset.

Related Posts

Hints for Computer System Design 07 Jan 2022

Synthesized Policies for Transfer and Adaptation across Tasks and Environments 29 Mar 2021

Deep Neural Networks for YouTube Recommendations 22 Mar 2021