Appendix D

Description of selected gold standards

MSMARCO was created by sampling real user queries from the log of a search engine and presenting the search results to experts in order to select relevant passages. Those passages were then shown to crowd workers in order to generate a free-form answer that answers the question or mark if the question is not answerable from the given context. While the released dataset can be used for a plethora of tasks we focus on the MRC aspect where the task is to predict an expected answer (if existent), given a question and ten passages that are extracted from web documents.

HOTPOTQA is a dataset and benchmark that focuses on "multi-hop" reasoning, i.e. information integration from different sources. To that end the authors build a graph from a where nodes represent first paragraphs of Wikipedia articles and edges represent the hyperlinks between them. They present pairs of adjacent articles from that graph or from lists of similar entities to crowd-workers and request them to formulate questions based on the information from both articles and also mark the supporting facts. The benchmark comes in two settings: We focus on the *distractor* setting, where question and answer are accompanied by a context comprised of the two answer source articles and eight similar articles retrieved by a information retrieval system.

RECORD is automatically generated from news articles, as an attempt to reduce bias introduced by human annotators. The benchmark entries are comprised of an abstractive summary of a news article and a close-style query. The query is generated by sampling from a set of sentences of the full article that share any entity mention with the abstract and by removing that entity. In a final step, the machine-generated examples were presented to crowd workers to remove noisy data. The task is to predict the correct entity given the Cloze-style query and the summary.

MULTIRC features passages from various domains such as news, (children) stories, or textbooks. Those passages are presented to crowd workers that are required to perform the following four tasks: *(i)* produce questions based multiple sentences from a given paragraph, *(ii)* ensure that a question cannot be answered from any single sentence, *(iii)* generate a variable number of correct and incorrect answers and *(iv)* verify the correctness of produced question and answers. This results in a benchmark where the task is to predict a variable number of correct natural language answers from a variable number of choices, given a paragraph and a question.

NEWSQA is generated from news articles, similarly to RECORD, however by employing a crowd-sourcing pipeline. Question producing crowd workers were asked to formulate questions given headlines and bullet-point summaries. A different set of answer producing crowd workers was tasked to highlight the answer from the article full text or mark a question as unanswerable. A third set of crowd workers selected the best answer per question. The resulting task is, given a question and a news article to predict a span-based answer from the article.

DROP introduces explicit discrete operations to the realm of machine reading comprehension as models are expected to solve simple arithmetic tasks (such as addition, comparison, counting, etc) in order to produce the correct answer. The authors collected passages with a high density of numbers, NFL game summaries and history articles and presented them to crowd workers in order to produce questions and answers that fall in one of the aforementioned categories. A submission was only accepted, if the question was not answered correctly by a pre-trained model that was employed on-line during the annotation process, acting as an adversary. The final task is, given question and a passage to predict an answer, either as a single or multiple spans from the passage or question, generate an integer or a date.