**Chatbot Project**

**Dataset:**

**WikiQA**

[**https://www.microsoft.com/en-us/download/details.aspx?id=52419**](https://www.microsoft.com/en-us/download/details.aspx?id=52419)

* **The WikiQA corpus is a new publicly available set of question and sentence pairs, collected and annotated for research on open-domain question answering. In order to reflect the true information need of general users, we used Bing query logs as the question source. Each question is linked to a Wikipedia page that potentially has the answer. Because the summary section of a Wikipedia page provides the basic and usually most important information about the topic, we used sentences in this section as the candidate answers. With the help of crowdsourcing, we included 3,047 questions and 29,258 sentences in the dataset, where 1,473 sentences were labeled as answer sentences to their corresponding questions. More detail of this corpus can be found in our EMNLP-2015 paper, "WikiQA: A Challenge Dataset for Open-Domain Question Answering" [Yang et al. 2015]. In addition, this download also includes the experimental results in the paper, an evaluation script for judging the "answer triggering" task, as well as the answer phrases labeled by the authors of the paper.**