the summary paragraph of the page as the candidate answer sentences, with labels on whether the sentence is a correct answer to the question provided by crowdsourcing workers. Among these questions, about one-third of them contain correct answers in the answer sentence set.

We implement several strong baselines to study model behaviors in the two datasets, including two previous state-of-the-art systems (Yih et al., 2013; Yu et al., 2014) on the QASENT dataset as well as simple lexical matching methods. The results show that lexical semantic methods yield better performance than sentence semantic models on QASENT, while sentence semantic approaches (e.g., convolutional neural networks) outperform lexical semantic models on WIKIQA. We propose to evaluate answer triggering using question-level precision, recall and F_1 scores. The best F_1 scores are slightly above 30%, which suggests a large room for improvement.

2 WIKIQA Dataset

In this section, we describe the process of creating our WIKIQA dataset in detail, as well as some comparisons to the QASENT dataset.

2.1 Question & Sentence Selection

In order to reflect the true information need of general users, we used Bing query logs as the question source. Taking the logs from the period of May 1st, 2010 to July 31st, 2011, we first selected question-like queries using simple heuristics, such as queries starting with a WH-word (e.g., "what" or "how") and queries ending with a question mark. In addition, we filtered out some entity queries that satisfy the rules, such as the TV show "how I met your mother." In the end, approximately 2% of the queries were selected. To focus on factoid questions and to improve the question quality, we then selected only the queries issued by at least 5 unique users and have clicks to Wikipedia. Among them, we sampled 3,050 questions based on query frequencies.

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. Figure 1 shows an example question, as well as the summary section of a linked Wikipedia page.

Question: Who wrote second Corinthians?

Second Epistle to the Corinthians The Second Epistle to the Corinthians, often referred to as Second Corinthians (and written as 2 Corinthians), is the eighth book of the New Testament of the Bible. Paul the Apostle and "Timothy our brother" wrote this epistle to "the church of God which is at Corinth, with all the saints which are in all Achaia".

Figure 1: An example question and the summary paragraph of a Wikipedia page.

2.2 Sentence Annotation

We employed crowdsourcing workers through a platform, which is similar to Amazon MTurk, to label whether the candidate answer sentences of a question are correct. We designed a cascaded Web UI that consists of two stages. The first stage shows a testing question, along with the title and the summary paragraph of the associated Wikipedia page, asking the worker "Does the short paragraph answer the question?" If the worker chooses "No", then equivalently all the sentences in this paragraph are marked incorrect and the UI moves to the next question. Otherwise, the system enters the second stage and puts a checkbox along each sentence. The worker is then asked to check the sentences that can answer the question *in isolation*, assuming coreference is resolved. To ensure the label quality, each question was labeled by three workers. Sentences with inconsistent labels would be verified by a different set of crowdsourcing workers. The final decision was based on the majority vote of all the workers. In the end, 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.

Although not used in the experiments, each of these answer sentence is associated with the *answer phrase*, which is defined as the shortest substring of the sentence that answers the question. For instance, the second sentence in the summary paragraph shown in Figure 1 is an answer sentence. Its substring "Paul the Apostle and Timothy our brother" can be treated as the answer phrase. The annotations of the answer phrases were given by the authors of this paper. Because the answer phrase boundary can be highly ambiguous, each sentence is associated with at most two answer phrases that are both acceptable, given by two different labelers. We hope this addition to the WIKIQA data can be beneficial to future researchers for building or evaluating an end-to-end question answering system.