# Building Resources for an Open Task on Question Generation

Vasile RUS[a], Eric WOOLLEY[a], Mihai LINTEAN[a], and Arthur C. GRAESSER[b]
[a]*Department of Computer Science*
[b]*Department of Psychology*
*The University of Memphis*
*Memphis, TN 38152*

**Abstract.** This paper contributes to the recent efforts to offer shared tasks on Question Generation by creating a data set of Question-Answer pairs collected from a community-based Question Answering repository. The data set was created with two shared tasks in mind: (1) question type generation and (2) open task. We have focused on the following six question types, which is a mix of shallow and deep questions: *how, who, what, when, where, why*.

**Keywords.** Question Generation, Data set, Open task

## Introduction

The recent Workshop on The Question Generation Shared Task and Evaluation Challenge (www.questiongeneration.org) has identified four major categories of shared tasks that could help boost research on Question Generation (QG) (see *Chapter 2* in [5]). One category is *Text-to-Question* which covers shared tasks in which the input is raw or annotated text and the goal is to generate questions for which the text contains, implies, or needs answers. Every reasonable or good question should be generated. It is beyond the scope of this paper to discuss what a good question is. A special case of this category is a task in which any, not necessarily every, good question is generated. We will call such a task an *open task*. That is, the generation of questions is not restricted in any way, it is open. An example of a more restrictive task would be to generate questions of a certain type, e.g. *how* questions. In this paper, we describe our efforts to build a data set that could be used in an open QG task as well as in a more restrictive task. The data set is in the form of question-answer pairs (Q-A pairs) collected from Yahoo!Answers, the community-based Question Answering service that Yahoo offers to its users ([3]). In Yahoo!Answers, users can freely ask questions which are then answered by other users. The answers are rated over time by various users and the expectation is that the best answer will be ranked highest after a while.

The task of selecting the *question type* implies only determining the type of question that best suits the given input (raw or annotated) without really generating the corresponding question. For instance, a fragment of text which describes a procedure, i.e. a set of steps, would most likely trigger a *how* question. Question type selection is a *subtask* of the full QG task because it is a step in the overall QG process [2, 5]. The overall QG process is usually regarded as a 4-step process: (1) deciding when to ask a

question, (2) content selection, (3) question type selection, and (4) question construction. In our work presented in this paper we collected Q-A pairs corresponding to the following six types of questions: *how, who, what, when, where,* and *why*. Thus, the data set could be used in a subtask of question type selection that focuses on these six types of questions.

**Data Collection**

The process of collecting data for an open task on Question Generation involved the following steps: (1) identifying or creating a good source for efficient collection of texts and associated questions, (2) automatically collecting Q-A pairs, (3) automatically filtering the Q-A pairs, and (4) high-quality manual filtering.

**1. Identifying or Creating a Data Source**

Yahoo!Answers is a good source for data to be used in QG shared tasks for two reasons: (1) it contains open domain/general knowledge content and (2) both questions and associated snippets of text are available. In a recent survey among QG researchers, the first author of the paper found that an open/general knowledge source of data is preferred, namely texts from Wikipedia, as opposed to other types of sources. One problem with Wikipedia is that only texts are available and no associated questions. Creating questions for Wikipedia texts would be an extremely time consuming and expensive exercise. Yahoo!Answers is a good alternative to Wikipedia as a source of open domain/general knowledge QG data. The advantage of Yahoo!Answers is the availability of questions for given text fragments which means that the expensive step of generating the questions is avoided. On the other hand, only one question is available for each answer. Ideally, multiple good questions of various types should be available for a text fragment/best answer. It should be noted that Microsoft offers a similar service to Yahoo!Answers. The only reason we started with Yahoo!Answers is our familiarity with it and the availability of a public programming interface.

**2. Automatic Collection**

The second step in creating the dataset was to collect a large number of questions and answers. We decided to collect Q-A pairs for the following six types of questions: *how, what, when, where, who,* and *why*. These types include factoid or shallow questions (*who, when*) as well as deep questions (*how, why*). Ideally, we would obtain a balanced data set in which the same number of instances will be collected for each of the six types of questions. The balanced data is best for evaluating and comparing different approaches to a given task, e.g. QG in our case. Our balanced data set could be eventually extended to reflect a real distribution of question types and thus serve other evaluation purposes. Another important aspect of our data collection efforts was to identify questions on topics and disciplines of general interest so that the data set is attractive to many research groups.

Following the above guidelines, we started collecting Q-A pairs from Yahoo!Answers. In Yahoo!Answers, questions are characterized by the following three parts: question summary (usually a single sentence), question content (a paragraph that details the question and its context), and the chosen answer, which can be of variable length. The length of the chosen answer varies from one single sentence to one full page of text. Questions in Yahoo!Answers are categorized based on their topic. Examples of categories of questions are *Allergies*, *Dogs*, *Garden & Landscape*, *Software, and Zoology*. In order to collect an initial set of Q-A pairs, a number of 244 Yahoo categories were queried on all six types of questions. To collect questions of each type we used the *wh*-words that define the question types (e.g. *where*, *what*, *how*, etc.) and searched for questions in Yahoo!Answers that contain those keywords in their summary. Table 1 lists some example of question summaries that were extracted from Yahoo!Answers. A maximum of 150 questions was downloaded per each category and question type, resulting in a total of maximum 150*244*6 = 219.600 number of candidate questions to be collected. Because not all categories had at least 150 answered questions of certain type, the number of collected questions is actually smaller.

**Table 1.** Examples of questions, one from each of the six categories.

| **Category** | **Type** | **Question Summary** |
|---|---|---|
| Add-ons | *How* | *How important is to have a mouse pad?* |
| Aircraft | *How* | *How do pilots of small aircraft know how far they are from an aerodrome?* |
| Economics | *Why* | *Why did the social and economic status change during the Middle Ages?* |
| Law & Ethics | *Who* | *Who wrote the final copy of the Stabilization Act of 2008?* |
| Radio | *Where* | *Where do radio stations get their digital music from?* |

## 3. Automatic Filtering

Upon review of the collected Q-A pairs in the previous step, it was clear that some of the data would not be appropriate to keep. The following criteria were applied to filter out Q-A pairs. The *first* criterion to be established was question length. In this case, length is described as the number of words in a given question. Questions in the Q-A pairs are denoted by the *Subject* tag. A minimum requirement of 3 words was proposed. This represents the smallest "common" length of a valid question. This corresponds to a typical definition questions such as *What is (object of choice)?* That is not to say that valid questions with less than 3 words do not exist. However, for the purposes of this dataset 3 words was set as the minimum. The same reasoning was applied to the answer data in the Q-A pairs. The answer data should be of a required minimum length. In our case, that length was set to 10 words. Again, valid answers can, of course, contain fewer words. However, for the purposes of this dataset and its intended uses, a minimum length of 10 words was selected.

A close review of the data also showed that it had content that was somewhat less than what might be perceived polite and courteous. Specifically, curse words, words that are and/or refer to sexual explicitness, and content that contained ethnically intolerant words. Therefore, a *second* criterion to filter out Q-A pairs was developed such that Q-A pairs with bad content were discarded. A "bad word" list was generated from the information gathered at two websites [6, 7]. When combined, the list totals

over 850 words. Unfortunately, this list is not complete, and it would be a considerable challenge to maintain a current listing of such terms given the dynamic nature of language. As such, a final decision with respect to including a Q-A pair into the final set must be made manually, i.e. by a human (see next subsection). It should also be noted that the "bad word" list does not include any misspellings of the terms therein, as it would be very difficult to anticipate all of the incorrect ways to spell these terms.

Other criteria were evaluated and rejected, e.g. Q-A pairs that were answered by the highly rated Yahoo!Answers users or the total number of answers per given question is another method of filtering that was considered.

The total reduction due to filtering in the dataset was about 55%.

## 4. High-Quality Manual Filtering

In the fourth and final step, we used three human raters to filter out the Q-A pairs resulted from the previous automatic filtering phase. The raters have worked on different subsets of the collected and automatically filtered Q-A pairs. This is because the goal was to collect as many Q-A pairs as possible. We will use different raters to judge same Q-A pairs once we have large enough initial data set. Inter-rater agreement scores will be reported.

In order to facilitate the human filtering step, we built a software tool that allows a quick analysis of each question and the corresponding content and answer by the human rater. Additionally, the tool allows easy relabeling and removal of the corresponding Q-A pair in case it is deemed unacceptable (incorrect, improper, or too difficult for the purposes of the data set which are question type detection and generating the best question for a given answer).

Even with the aid of the tool, manual filtering proved to be a time-consuming process. On average, it takes about 10 hours to select about 100 questions. The advantage is that it results in high quality Q-A pairs for the two QG tasks for which we developed the data set. Similar to automatic filtering, manual filtering further reduces the collected data set. On average, out of all the Q-A pairs judged by humans about 10% of them are retained. For some types of questions and topics, such as *when* questions and *Camcorders*, the retaining rate is even lower than 10% at 2%. This is because the majority of *when* questions in the Camcorders category are not really *when* questions. The keyword *when* is used frequently to describe a context while the actual main question is usually of a different type as illustrated by the following example: *When I transfer footage from my video camera to my computer why can't I get sound?* In this example, *when* introduces the context of the main *why* question. The only category with a high retaining rate for *when* types of questions is the *History* category. A similar problem of low retaining rate during human filtering was noticed for *where* questions. *Where* can be used to ask for a source of information, e.g. a website or book, and not necessarily for a real place. An example of a *where* question which asks for an information source is the following: *Where can I find information about the Industrial Revolution in USA?* While such *where* questions could be retained we opted not to.

We list below examples of reasons for which Q-A pairs were discarded during manual filtering.

1. **The question is a compound question.** For instance, the following question contains both a *how* and an *who* question: *How do you figure out who your video card manufacturer is?* Another example is the question *How long has*

*the computer mouse been around, and who is credited with its invention ?*

2. **The question is not in interrogative form.** An example of a non-interrogative question is the following *I want a webcam and headset etc to chat to my friends who moved away?*
3. **Poor grammar or spelling.** The following questions are examples of questions with poor grammar and spelling: *Yo peeps who kno about comps take a look?* and *Who the memory eaten is bigger?*
4. **The question does not solicit a reasonable answer for our purposes.** An example of such a question is *Who knows something about digital cameras?*
5. **The question is ill-posed**. For instance, for the question *When did the ancient city of Mesopotamia flourish?* the answer is *Mesopotamia wasn't a city.*

## Future Work and Conclusions

The initial data set we collected so far contains about 500 Question-Answer clean pairs (automatically and then manually filtered) selected out of a set of almost 5000 candidate Q-A pairs. The goal is to increase the size of the clean dataset to 5000. We already have enough raw Q-A pairs collected and automatically filtered. The only challenge to increase the size of the clean dataset is to manually filter more Q-A pairs. Once we have achieved our goal of 5000 clean Q-A pairs we plan to further validate the Q-A pairs and to annotate the answers with a deep representation. For further validation, we envision an experiment in which human subjects are shown first the answer and then the corresponding question and asked to rate how good they consider the question given the answer. We are considering three options for annotating the answers with a deep representation: FrameNet [1], PropBank [4], or Unified Annotation Language [9]. In a recent survey among QG researchers, PropBank seems to be the preferred annotation language.

One important conclusion to draw from our work is that collecting data from community Question Answering sources is more challenging than it seems. The alternative of explicitly collecting Q-A pairs from sources of general interest such as Wikipedia through target experiments with human subjects may be a comparable rather than a much costlier effort.

## References


[1] C. Fillmore, C.F. Baker & H. Sato. The FrameNet Database and Software Tools. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*. Las Palmas. 1157-1160, 2002.

[2] R. Nielsen. (2008). Question Generation: Proposed challenge tasks and their evaluation. *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

[3] T. Marciniak, Language generation in the context of Yahoo! answers. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA, 2008.

[4] M. Palmer, P. Kingsbury, D. Gildea. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* **31** (1): 71–106, 2005.

[5] V. Rus & A.C. Graesser, *Workshop Report: The Question Generation Task and Evaluation Challenge*, Institute for Intelligent Systems, Memphis, TN, ISBN: 978-0-615-27428-7.

[6] http://www.noswearing.com/list.php

[7] http://carbonize.co.uk/Old/sexglos.php

[8] http://www.cs.brandeis.edu/~jamesp/ula2007/index.html