Crowdsourcing Breakthroughs for Language Technology Applications
AAAI Technical Report WS-15-24

The Effect of Text Length in Crowdsourced Multiple Choice Questions

Sarah K. K. Luger
Institute for Language, Cognition and Computation
The University of Edinburgh
Edinburgh, UK
s.k.k.luger@sms.ed.ac.uk

Abstract

Automated systems that aid in the development of Multiple Choice Questions (MCQs) have value for both educators, who spend large amounts of time creating novel questions, and students, who spend a great deal of effort both practicing for and taking tests. The current approach for measuring question difficulty in MCQs relies on models of how good pupils will perform and contrasts that with their lower-performing peers. MCQs can be difficult in many ways. This paper looks specifically at the effect of both the number of words in the question stem and in the answer options on question difficulty. This work is based on the hypothesis that questions are more difficult if the stem of the question and the answer options are semantically far apart. This hypothesis can be normalized, in part, with an analysis of the length of texts being compared. The MCQs used in the experiments were voluntarily authored by university students in biology courses. Future work includes additional experiments utilizing other aspects of this extensive crowdsourced data set.

Introduction

Crowdsourcing presents an alternate method from academic, institutional, and research-oriented document annotations for gathering useful, human judgment data. Measuring the quality of the participating workers' ouput and the difficulty of the individual tasks can be complex. Effective measurement is integral in defining worker payment and designing the annotation tasks that are often structured as pipelines of gated tasks or as decision trees. In the case of the experiments presented in this paper, MCQs will be assessed for difficulty and the amount of text in the questions and answer options will be analyzed.

The use of standardized comprehension or aptitude exams requires having access to sets of exam data, which include the questions and detailed, question-by-question results from thousands of students. Unfortunately, such ideal data is very difficult, if not impossible, to obtain. The use of crowdsourced, human-annotated, or "human-in-the-loop" data has emerged as an important resource for human judgments including answering exam questions. For example, Amazon's Mechanical Turk (Amazon 2013) and the crowdsourcing company Crowdflower (Biewald and Van Pelt 2013) both provide avenues to gather human judgments on myriad tasks (Callison-Burch 2009). As described in greater detail in the Data Section, PeerWise was chosen for this work because it had extensive metadata on the questions and answers, and also because it is open source software.

Multiple Choice Questions

To measure the usefulness of exam questions, researchers have devised methods for judging both the difficulty of the question and the differentiation power of the answer options (Patz and Junker 1999) and (Beguin and Glas 2001). One such approach is Item Analysis Theory (Gronlund 1981). The analysis of questions based on how students perform in a group with the goal of improving exam-based MCQs is broadly called Test Item Difficulty. Within the study of Test Item Difficulty are several measures including Item Analysis, which evaluate how each student performed based on which answer choice he or she made.

Comprehension and aptitude tests seek to present questions that can be correctly answered by students who understand the subject matter and to confuse all other students with seemingly viable alternate answer options (distractors). A *good* or difficult distractor is one that catches or distracts more bad students than good students.

A high-scoring student is one who answers most questions correctly, but when their answers are incorrect, chooses the best distractors. A low-scoring student will choose any of the answer options seemingly at random. A difficult question is one whose answer options are all deemed viable to a high-scoring student. With a difficult

question, the high-scoring cohort will behave like low-scoring students, with a near equal spread of multiple distractors being chosen.

A MCQ or *item* consists of several parts. The stem, or *question statement*, presents a query that is best answered by one answer option. The answer options, or *answer alternates*, include both the answer and distractors. In the following example, the item is the question statement ("Which hormone secretion pattern is directly affected from jet lag") and the four answer options (A-D). This question, from the student-authored question bank called PeerWise (Denny 2009). This question was answered 282 times and 192 students chose the correct answer A.

Example 1: Which hormone secretion pattern is directly affected from jet lag?

A. Cortisol *(correct answer)*
B. Insulin
C. Thyroid Hormone
D. Adrenaline

A question may be difficult in many ways. The stem may be confusing or poorly scoped. The topic of the stem may be from an obscure corner of a discipline or use ambiguous terminology. Further, when a question has multiple answer options, high-quality incorrect options and discriminating distractors can make a question difficult. The experiments in this paper are based on questions being difficult because the words used in the question stem are not similar to the words used in the answer options.

Using Crowdsourced Data

I procured data for two sets of MCQs from university-level introductory biology classes using the PeerWise question creation system. PeerWise consists of questions that are created by students and answered by their classmates. Instructors can review the questions or use some of the better questions for future exams. Since answering these questions may not be compulsory, the resulting data is a set of questions that have been answered by students but not all of the questions have been answered by the same students.

The process of choosing questions for the experiments consisted of automatically collecting the subset of questions that used inverse definition constructions such as "is called," "known as," and "is named". Inverse definition questions describe a term or process by providing a definition and seek the name of the process. This question format is frequently used in the sciences where mastering domain-specific concepts are a key measure of comprehension.

An exam is a set of students who have answered the same questions. The PeerWise data sets consist of students who have answered some questions, but not necessarily the same questions from a set. In a graph, these may be viewed as shared edges as noted in Table 1. Thus, the data contains an incomplete, or sparse exam, as seen in Table 1.

Course	1	2
Total number of students	1055	887
Total number of questions	148	132
Shared edges between the questions and students	31019	31314

Table 1: Data characteristics from the two academic courses. A shared edge represents a question answered by a student.

One potential research problem is the hypothesis that PeerWise attracts the better-performing students to practice and build their expertise in a field. The better students may tend to both author and answer more questions than their lower-performing peers. Thus, the PeerWise system may skew Item Analysis from a more conventional normal distribution across performance cohorts to a tight cluster of top-scoring students versus a long tail of the middle- and lowest-performing students. This hypothesis was tested and revealed that a sufficiently large group of potential students participated in the exam for meaningful Item Analysis (Luger and Bowles 2013). Consequently, the bell curve composed of the three performance cohorts–lowest, middle, and highest-scoring students–did shift to the left, but the three cohorts were distinctly discriminated.

The PeerWise data has an exhaustive amount of detailed information that covers the authoring and taking of course-related questions. Much of this information is outside the scope of analyzing question difficulty. There is sufficient additional question information for myriad research projects, but there are a few characteristics of the data in general that reflect its value for this and additional research. These characteristics are as follows:

- The fewest number of questions answered by any of the students was 1. 101 students in course 1 answered only 1 question; 152 students in course 2 answered only 1 question.
- 112 was the most questions answered by a student in course 1; 11 students answered 112.
- The average number of questions answered in course 1 was 26.6 and in course 2 was 35.8.
- No student answered a question more than once.
- None of the questions were so easy that all of the students answered them correctly, nor so hard that none of them got them correct.
- The most times a single question was answered in course 1 was 439; the least was 89.
- In course 2, 331 was the maximum number of times a question was answered, 132 the minimum.
- There were 62,333 distinct answers or questions that were answered in total.

- There were 20,532 incorrect answers of the total 62,333 answers, or 32.7%.
- There were 14,094 question ratings; each of the sets' 280 total questions were rated at least once.

Analysis of the student exam results suggests a method for classifying the output of the automated system that parallels the human results. This supports my underlying belief that the closeness of answer options to the question stem and to one other as well is reflected in student results. When a student is choosing an answer to an inverse definition question, he or she is in some manner choosing the closest concept, process, or term that sufficiently fulfills the description given in the question stem. Based on this perception, a difficult question is one where the answer options are not only closely related terms but also meaningfully linked to the question stem. In other words, my research compared how students selected answer options with how the automated system found shared terms in the definitions of the answer options and the original question stem.

Mitkov et al. introduces distractor classes, which are a way of grouping how well an answer option "distracted" low-performing students as opposed to their higher- performing peers (Mitkov et al. 2006). The four distractor classes are "good" or useful distractors, "poor" distractors, not useful distractors, and distractors that confer no discriminating power.

Good distractors are negative numbers indicating that they attracted more low-scoring students than high-scoring ones. Poor distractors are positive numbers because they are chosen by more high-performing students than low-performing ones. The goal of exams is to differentiate performance groups of students and distract less-prepared students.

My hypothesis for generating a model of question difficulty is focused on how many distractors with negative (good) values there are for a given question. One way that MCQs can be hard is when there are several strong answer options. When multiple answer options have positive numbers, there are answer options with strong distraction power. As noted earlier, in this case, good students act like their lower-performing peers conventionally do and spread their answer choice among several answer options. Difficulty can be measured by the total number of students who answered a question correctly divided by the number of students who took the exam. Difficulty can also be measured by how many good distractors are in a question.

Methodology and Results

There are two questions that arose from the early data analysis: What method for answering inverse definition questions best controlled for stem word bag size and what method best controlled for answer option length? The concern was a larger stem bag size might support more correct answers.

The "Define: X" experiment used an API to automatically return the top 50 definitions for each answer option. Then, the text of the original question stem was compared to each definition's text using Lucene (Apache 2013), incorporating both bag-of-words and WordNet weights (Princeton University 2010). The answer option producing the individual definition with the most text overlap was deemed correct.

The subsequent query type experiment used sampling methods, where the sample size was justified according to appropriate statistical criteria (Feller 1950). Random sampling selected questions representing the three stem size bags present in the question set. There were questions with 5 and fewer content words, 6 to 9 content words, and 10 or more content words.

As Table 2 shows, questions with 6 or more words in the stem were answered correctly more often than those with shorter stems. This suggests that the greater the amount of information provided in the query, the more likely the chance of getting the answer correct in the experiments performed. In the "Define: X" experiment, the search engine shortcut "Define: X" is utilized where "X" is the term being defined. In this experiment, each of the answer options is automatically sent to a search engine using this "Define: X" pattern and the results of the query are compared to the text of the question.

Close analysis of the question answer options for concept identification showed three answer types: 1 word, 2 words (usually a colocation), and more than 2 words. I randomly selected questions based on these answer alternatives and tested how the "Define: X" query performed on these classes of answers. The results are in Table 3.

Query type	"Define: X"
Total percent correct	57%
Correct with stems of 10 words and greater	100%
Correct with stems of 6-9 words	40%
Correct with stems of 5 and fewer	33%

Table 2: Results controlling for stem length from the "Define: X" experiment.

The nature of inverse definition questions is to present a definition and seek the concept being defined. In the biology domain, the answer options are concepts and they are primarily 1 word terms or 2 word noun phrases that colocate or are paired to indicate one concept. In the experiments, the web query with the most word overlap with the question stem was used as the "choice" of the system.

Answer option length	1 word	2 words	>2 words
Query type "Define: X"	83%	50%	100%

Table 3: Results controlling for answer option length from the "Define: X" experiment.

Table 3 shows that the "Define: X" query type performed best with varied answer option length. Perhaps, this is related to the underlying query methodology that uses the "Define: X" shortcut to search specifically for definitions. Thus, multi-word definitions are maintained as colocated words when they are sent to the web for results. Concepts such as "active transport" are defined as a unit and not individually with the definition of "active" and "transport" combined. The returned definitions were compared to the question stem and this approach does moderately well when stem length is controlled.

Example 2: What type of glia cells engulfs and destroys micro-organisms and debris?

A. Astrocytes
B. Microglia *(correct answer)*
C. Ependymal Cells
D. Oligodendrocytes
E. Schwann Cells

Some of these data analysis issues are exemplified by Example 2. The two highest scoring "Define: X" results in Example 2 supported "Ependymal Cells" and "Schwann Cells." Neither result is the correct answer but each shares the term "cells" with the question stem. "Glia" is the more important word in the noun phrase "glia cells" and it appears as the root of the correct answer "Microglia." Unfortunately, this match was not caught because "microglia" was not separated into its constituent parts and there was no weighting of the word component matches that would have supported a different, correct result.

The biology domain uses specific scientific terminology that often adheres to Latin and Greek-based word templates. "Micro" is a prefix that could be used with other similar biology prefixes to identify or decompose words for possible matching. Domain-specific spelling variants should also be incorporated. "Micro-organism," which appears as it was written by a student, could match "micro," "organism," "micro-organism," "micro organism," but most importantly, it should first match "microorganism" because that is the correct spelling of the word in full. Similarly, the suffixes "-cytes" and "-cyto" mean "cells." Incorporating a table of biological terms and their variants would increase successful matching. In addition, this example shows a limit on the "Define: X" component of the experiment pipeline and area for future system improvements.

Future Work

The next steps in this research focus on comparing these results to those from MCQs generated by professional educators. In addition, more sophisticated natural language processing techniques could be introduced to extend the text comparisons. For example, it would be interesting to observe both how the "Define: X" pattern worked on different types of distractor classes and the role that text length played in those results. Finally, there is additional information on the students and questions that could lead to deeper analysis of the dataset.

References

Amazon. 2013. Amazon's mechanical turk. http://www.mturk.com/.

Apache. 2013. Lucene. http://lucene.apache.org/.

Beguin, A. A., and Glas, C. 2001. Mcmc estimation and some model-fit analysis of multidimensional irt models. In *Psychometrika, Vol. 66, No. 4, pp. 541-562*.

Biewald, L., and Van Pelt, C. 2013. Crowdflower. http://www.crowdflower.com/.

Callison-Burch, C. 2009. Fast, cheap, and creative: evaluating translation quality using amazons mechanical turk. In *EMNLP 09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Denny, P. 2009. Peerwise. http://peerwise.cs.auckland.ac.nz/.

Feller, W. 1950. *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, Inc.

Gronlund, N. E. 1981. *Measurement and Evaluation in Teaching*. Macmillan, 4 edition.

Luger, S. 2011. A graph theory approach for generating multiple choice exams. In *2011 AAAI Fall Symposium on Question Generation*.

Luger, S. and Bowles, J. 2013. An analysis of question quality and user performance in crowdsourced exams. In *Proceedings of the 2013 Workshop on Data-Driven User Behavioral Modelling and Mining from Social Media at CIKM 2013. pp. 29-32.*

Mitkov, R., Ha, L. A., and Karamanis, N. 2006. A computer-aided environment for generating multiple choice test items. In *Natural Language Engineering 12(2) pp. 177-194*.

Patz, R. J., and Junker, B. W. 1999. Applications and extensions of mcmc in irt: Multiple item types, missing data, and rated responses. In *Journal of Educational and Behavioral Statistics, Vol. 24, No. 4 (Winter, 1999), pp. 342-366*.

Princeton University. 2010. About WordNet. http://www.wordnet.princeton.edu/.

Raykar, V.C.; Yu, S.; Zhao, L.H.; Valadez, G.H.; Florin, C.; Bogoni, L.; Moy, L. 2010. Learning from crowds. *Journal of Machine Learning Research* Vol. 11 pp. 1297-1322.

Whitehall, J.; Ruvolo, P.; Wu, T.; Bergsma, J., and Movellan, J. 2009. Whose vote should count more: optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems.*