Collecting Image Annotations Using Amazon's Mechanical Turk

Cyrus Rashtchian Peter Young Micah Hodosh Julia Hockenmaier
Department of Computer Science
University of Illinois at Urbana-Champaign
201 North Goodwin Ave, Urbana, IL 61801-2302
{crashtc2, pyoung2, mhodosh2, juliahmr}@illinois.edu

Abstract

Crowd-sourcing approaches such as Amazon's Mechanical Turk (MTurk) make it possible to annotate or collect large amounts of linguistic data at a relatively low cost and high speed. However, MTurk offers only limited control over who is allowed to particpate in a particular task. This is particularly problematic for tasks requiring free-form text entry. Unlike multiple-choice tasks there is no correct answer, and therefore control items for which the correct answer is known cannot be used. Furthermore, MTurk has no effective built-in mechanism to guarantee workers are proficient English writers. We describe our experience in creating corpora of images annotated with multiple one-sentence descriptions on MTurk and explore the effectiveness of different quality control strategies for collecting linguistic data using Mechanical MTurk. We find that the use of a qualification test provides the highest improvement of quality, whereas refining the annotations through follow-up tasks works rather poorly. Using our best setup, we construct two image corpora, totaling more than 40,000 descriptive captions for 9000 images.

1 Introduction

Although many generic NLP applications can be developed by using existing corpora or text collections as test and training data, there are many areas where NLP could be useful if there was a suitable corpus available. For example, computer vision researchers are becoming interested in developing methods that can predict not just the presence and location of certain objects in an image, but also the relations between objects, their attributes, or the actions and events they participate in. Such information can neither be obtained from standard computer vision data sets such as the COREL collection nor from the user-provided keyword tag annotations or captions on photo-sharing sites such as Flickr. Similarly, although the text near an image on a website may provide cues about the entities depicted in the image, an explicit description of the image content itself is typically only provided if it is not immediately obvious to a human what is depicted (in which case we may not expect a computer vision system to be able to recognize the image content either). We therefore set out to collect a corpus of images annotated with simple full-sentence descriptions of their content. To obtain these descriptions, we used Amazon's Mechanical Turk (MTurk).[1] MTurk is an online framework that allows researchers to post annotation tasks, called HITs ("Human Intelligence Task"), then, for a small fee, be completed by thousands of anonymous non-expert users (Turkers). Although MTurk has been used for a variety of tasks in NLP, our use of MTurk differs from other research in NLP that uses MTurk mostly for annotation of existing text. Similar to crowdsourcing-based annotation, quality control is an essential component of crowdsourcing-based data collection efforts, and needs to be factored into the overall costs. For us, the quality of the text produced by the Turkers is particularly important since we are interested in us-

[1] All of our experiments on Mechanical Turk were administered and paid for through the services offered by Dolores Labs.

Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 139–147,
Los Angeles, California, June 2010.

ing this corpus for future research at the intersection of computer vision and natural language processing. However, MTurk provides limited ways to implement such quality control directly. For example, our initial experiments yielded a data set that contained many sentences that were clearly not written by native speakers. We learned that several steps must be taken to ensure that Turkers both understand the task and produce quality data.

This paper describes our experiences with Turk (based on data collection efforts in spring and summer 2009), comparing two different approaches to quality control. Although we did not set out to run a scientific experiment comparing different strategies of how to collect linguistic data on Turk, our experience points towards certain recommendations for how to collect linguistic data on Turk.

2 The core task: image annotation

The PASCAL Data Set Every year, the Pattern Analysis, Statistical Modeling, and Computational Learning (PASCAL) organization hosts the Visual Object Classes Challenge (Everingham et al., 2008). This is a competition similar to the shared tasks familiar to the ACL community, where a common data set of images with classification and detection information is released, and computer vision researchers compete to create the best classification, detection, and segmentation systems. We chose to use this collection of images because it is a standard resource for computer vision, and will therefore faciliate further research.

The VOC2008 development and training set contains around 6000 images. It is categorized by objects that appear in the image, with some images appearing in multiple categories.[2]. The images contain a wide variety of actions and scenery. Our corpus consists of 1000 of these images, fifty randomly chosen from each of the twenty categories.

MTurk setup We asked Turkers to write one descriptive sentence for each of ten images. An example annotation screen is shown in Figure 1. We first showed the Turkers a list of instructive guidelines describing the task (Figure 6). The instructions told them to write ten complete but simple sentences, to include adjectives if possible, to describe the main characters, the setting, or the relation of the objects in the image, to pay attention to grammar and spelling, and to try to be concise. These instructions were meant to both explain the task and to prepare Turkers to write quality sentences. We then showed each Turker a set of ten images, chosen randomly from the 1000 total images, and displayed one at a time. The Turkers navigated using "Next" buttons through the ten annotation screens, each displaying one image and one text-box. We allowed Turkers ten minutes to complete one task.[3] We restricted the task to Turkers who have previously had at least 95% of their results approved. We paid $0.10 to complete one task. The total cost for all 5000 descriptions was $50 (plus Amazon's 10% fee).

Figure 1: Screenshot of the image annotation task.

2.1 Results

On average, Turkers wrote the ten sentences in a total of four minutes. The average pay rate was $1.30 per hour, and the whole experiment finished in under two days. Five different people described each image, and in the end, most of the Turkers completed the task successfully, although 2.5% of the 5000 sentences were empty strings. Turkers varied in the time they took to complete the experiment, in the length of their sentences, and in the level of detail they included about the image. An example captioned image is shown in Figure 2.

Problems with the data The quality of descriptions varied greatly. We were hoping to collect simple sentences, written in correct English, describing the entities and actions in the images. More-

[2]The twenty categories include people, various animals, vehicles and other objects: `person`, `bird`, `cat`, `cow`, `dog`, `horse`, `sheep`, `aeroplane`, `bicycle`, `boat`, `bus`, `car`, `motorbike`, `train`, `bottle`, `chair`, `dining table`, `potted plant`, `sofa`, `tv/monitor`

[3]This proved to be more than enough time for the task.

Two men playing cards at a table.
The two men are in an intense card game.
Two men playing cards on a kitchen counter are lit by a strong flash.
The men are playing cards
The scene shows two people playing cards.

Figure 2: An image along with the five captions that were written by Turkers.

over, these are explicitly the types of descriptions we asked for in the MTurk task instructions. Although we found the descriptions acceptable more than half of the time, a large number of the remaining descriptions had at least one of the following two problems:

1. Some descriptions did not mention the salient entities in the image, some were simply noun phrases (or less), and some were humorous or speculative.[4] We find all of these to be problems because future computer vision and natural language processing research will require accurate and consistent image captions.

2. A number of Turkers were not sufficiently proficient in English. Many descriptions contained grammar and spelling errors, and some included very awkward constructions. For example, the phrase "X giving pose" showed up several times in descriptions of images containing people (e.g. *"The lady and man giving pose."*). Such spelling and grammar errors will pose difficulties for any standard text-processing algorithms trained on native English.

Spell checking Due to the large number of misspellings in in the initial data set, we first ran the sentences first through our spell checker before putting them up on Turk to assess their quality. We tokenized the captions with OpenNLP, and first checked a manually created list of spelling corrections for each token. These included canonicalizations (correcting "surf board" as "surfboard"), words our automatic spell checker did not recognize ("mown"), and the most common misspellings in our data set ("shepard" to "shepherd"). If the token was not in our manual list, we passed the word to aspell. From aspell's candidate corrections, we selected the most frequent word that appeared either in other captions of the same image, of images of the same topic, or any caption in our data set.

[4]For example, some Turkers commented on the feelings of animals (e.g. *"the dog is not very happy next to the dumpster"*), and others made jokes about the content of the image (e.g. *"The goat is ready for hair cut"*)

3 Post-hoc quality control

Because our initial data collection efforts resulted in relatively noisy data, we created a new set of MTurk tasks designed to provide post-hoc quality control. Our aim was to filter out captions containing misspellings and incorrect grammar.

MTurk setup Each HIT consisted of fifty different image descriptions and asked Turkers to decide for each of them whether they contained correct grammar and spelling or not. At the beginning of each HIT, we included a brief training phase, where we showed the Turkers five example descriptions labeled as "correct" or "incorrect" (Figure 7). In the HIT itself, the fifty descriptions were displayed in blocks of five (albeit not for the same image) , and each description was followed by two radio buttons labeled "correct" and "incorrect". We did not show the corresponding images. A screenshot is shown in Figure 3. Each block of five captions contained one control item that we use for later assessment of the Turkers' spell-checking ability. We wrote these control captions ourselves, modeling them after actual image descriptions. We paid $0.08 for one task, and three people completed each task.

3.1 Results

On average, Turkers completed a HIT (judging fifty sentences) in four minutes, at an average hourly rate of $1.04. Each sentence in our data set was judged by three Turkers. The whole experiment finished

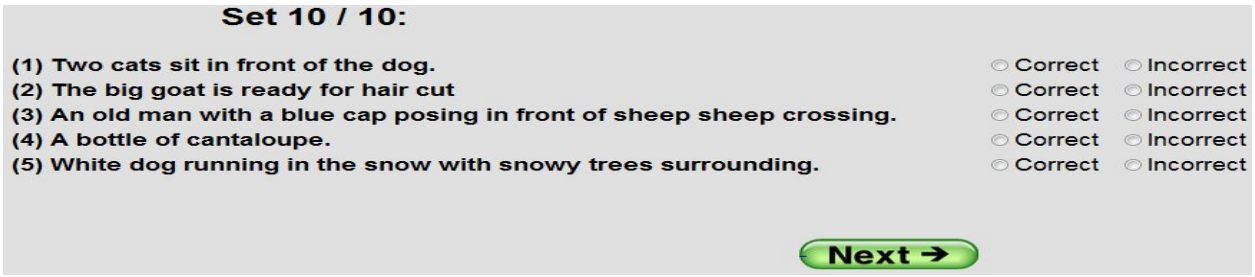

Figure 3: Screenshot from the grammar/spelling checking task. This is a block of five sentences that Turkers had to label as using correct or incorrect grammar and spelling. The first sentence is a control item that we included to monitor the Turkers' performance, and the other four are captions generated by other Turkers in a previous task.

Data set produced by...	Quality control performed by...	% Votes for "correct English" 0	1	2	3
Unqualified writers	three Turkers	18.9%	31.2%	26.4%	23.5%
Unqualified writers	three experts	11.8%	12.7%	15.3%	60.2%
Qualified writers	three experts	0.5%	2.5%	15.0%	82.0%

Table 1: Quality control by Turkers and Experts. The three experts judged 600 sentences from each data set. 565 sentences produced by unqualified workers were also judged by three Turkers.

in under two days, at a total cost of $28.80 (plus Amazon's 10% fee). We also selected randomly 600 spell-checked sentences for expert annotation. Three members of our team (all native speakers of English) judged each of these sentences in the same manner as the Turkers. Each sentence could therefore get between 0 and 3 Turker votes and between 0 and 3 expert votes for good English. The top two rows of Table 1 show the distribution of votes in each of the two groups. We also assess whether the judgments of the Turkers correlate with our own expert judgments. Table 2(a) shows the overall agreement between Turkers and expert annotators. The rest of Table 2 shows how performance of the Turkers on the control items affected agreement with expert judgments. We define the performance of a Turker in terms of the average the number of control items that they got right in each HIT they took. For each threshold in Tables 2(a)-(d), we considered only those images for which we have three quality judgments by workers whose performance is above the specified threshold.

Our results show that the effectiveness of using Turkers to filter for grammar and spelling issues is limited. Overall, the Turker judgments were overly harsh. The majority Turker vote agrees with the majority vote of the trained annotators on only 65.1% of the sentences. Manual inspection of the differences reveals that the Turkers marked many perfectly grammatical English sentences as incorrect (although they also marked a few which we had missed). Agreement with experts decreases among those Turkers that performed better on the control sentences, with only 56.7% agreement for Turkers that got all the controls right. In addition, the Turkers are significantly more likely to report false negatives over false positives and this also increases with performance on the control sentences. (Overall, the Turkers marked 29.9% of the sentences as false negatives, whereas the Turkers that scored perfectly on the controls marked 39.3% as false negatives.) Examination of the areas of high disagreement reveal that the Turkers were much more likely to vote down noun phrases than the experts were. The correct example captions provided in the instructions of the quality control test were complete sentences. Some of the control captions were noun phrases, but all of the noun phrase controls had some other error in them. Thus it was possible to either believe that noun phrases were correct or incorrect, and still be consistent with the provided examples, and provide correct judgments on the control sentences.

(a) ≥ 0 controls correct: 565 sentences

Turk votes	Expert votes 0	1	2	3
0	6.9%	4.4%	3.7%	3.9%
1	3.2%	5.7%	5.0%	17.3%
2	1.8%	2.8%	3.5%	18.2%
3	0.0%	0.4%	2.5%	20.7%

(b) ≥ 5 controls correct: 553 sentences

Turk votes	Expert votes 0	1	2	3
0	6.9%	4.5%	3.8%	4.0%
1	3.1%	5.4%	5.1%	17.5%
2	1.8%	2.7%	3.6%	18.4%
3	0.0%	0.4%	2.5%	20.3%

(c) ≥ 7 controls correct: 331 sentences

Turk votes	Expert votes 0	1	2	3
0	6.9%	6.3%	3.9%	5.1%
1	3.0%	4.5%	5.1%	24.5%
2	1.8%	1.8%	2.4%	15.1%
3	0.0%	0.0%	2.1%	17.2%

(d) ≥ 9 controls correct: 127 sentences

Turk votes	Expert votes 0	1	2	3
0	7.9%	6.3%	3.1%	6.3%
1	1.6%	4.7%	6.3%	23.6%
2	0.8%	3.1%	1.6%	15.7%
3	0.0%	0.0%	1.6%	17.3%

Table 2: Quality control: Agreement between Turker and Expert votes, depending on the average number of control items the Turker voters got right.

4 Quality control through pre-screening

Quality control can also be imposed through a pre-screening of the Turkers allowed to take the HIT. We collected another set of five descriptions per image, but restricted participation to Turkers residing in the US[5], and created a brief qualification test to check their English. We would like to be able to restrict our tasks to Turkers who are native speakers and competent spellers and writers of English, regardless of their country of residence. However, this seems to be difficult to verify within the current MTurk setup.

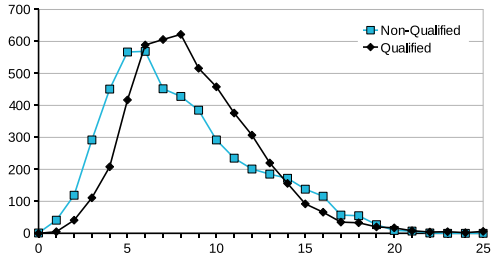

Figure 4: Average caption length (5000 images)

Qualification Test Design The qualification test consists of forty binary questions: fifteen testing spelling, fifteen testing grammar, and ten testing the ability to identify good image descriptions.

In all three cases, we started the section with a set of instructions displaying examples of positive and negative answers to the tasks. Each spelling question consisted of a single sentence, and Turkers were asked to determine if all of the words in the sentence were spelled correctly and if the correct word was being used ("lose" versus "loose"). Each grammar question consisted of a single sentence that was either correct or included a grammatical error. Both spelling and grammar checking questions were based on common mistakes made by foreign English speakers and on grammatical or spelling errors that occurred in our initial set of image captions. The grammar and spelling questions are listed in Table 3. The image description questions consisted of one image shown with two actual captions, and the Turkers were asked which caption better described the image. In order to pass the qualification test, we required each annotator to correctly answer at least twenty-four spelling and grammar questions and at least eight image description questions. To prevent Turkers from using the number of question they got correct to do a brute force search for the correct answers, we simply told them if they passed ("1") or failed ("0"). Currently, 1504 people have taken the qualification test, with a 67.2% passing rate. Since this qualification test was only required for our HITs that were restricted to US residents, we assume (but are not able to verify) that most, if not all, of the people who took this test are actually US residents.

[5]As of March 2010, 46.80% of Turkers reside in the U.S (`http://behind-the-enemy-lines.blogspot.com/` 03/09/2010)

MTurk Set-up We use the same MTurk set-up as before, but to encourage Turkers to complete the task even though they first have to pass a qualification test, we pay them $0.10 to annotate five images.

4.1 Results

We found that the Turkers who passed the qualification provided much better captions for the images. The average time spent on each image was longer (four minutes per ten images for the non-qualified workers versus five minutes per ten images for the qualified workers). On average, qualified Turkers produced slightly longer sentences (avg. 10.7 words) than non-qualified workers (avg. 10.0 words) (Figure 4), and the awkward constructions produced by the unqualified workers were mostly absent. The entire corpus was annotated in 253 hours at a cost of $100.00 (plus Amazon's 10% fee).

We also looked at the rate of misspellings (approximated by how often our spell-checker indicated a misspelling). Without the qualification test, Out of the 600 sentences produced without the qualification test, 78 contained misspellings, whereas only 25 sentences out of the 600 produced by the qualified workers contained misspellings. Furthermore, misspellings in the no-qualification group include many genuine errors (*"the boys are playing in tabel"*, *"bycycles"*, *"eatting"*), whereas misspellings in the qualification group are largely typos (e.g. *Ywo* for *Two*, *tableclothe*, *chari* for *chair*). Furthermore, the spell checker corrected all 25 misspellings in the qualified data set to the intended word, but 27 out of the 78 misspellings in the data produced by the unqualified workers got changed to some other word.

The same three members of our team rated again the English of 600 randomly selected sentences written by Turkers residing in the US who passed our test. We found a significant improvement in quality (Table 1, bottom row), with the majority expert vote accepting over 97% of the sentences. This is also corroborated by qualitative analysis of the data (see Figure 5 for examples). Inspection reveals that sentences that are deemed ungrammatical by the experts typically contain some undetected typo, and would be correct if these typos could be fixed. Without a qualification test, there is a significantly greater percentage of nonsensical responses such as: "Is this a bird squirrel?" and "thecentury". In addition, grammatically correct but useless fragments such as "very dark" and "peace" only appear without a test. After requiring the qualification test, the major reasons for rejection by Turkers are typos such as in "The two dogs blend in with the stuff animals" or missing determiners such as in "a train on tracks in town".

Overall cost effectiveness Using the no qualification test approach, we first paid $50.00 to get 5000 sentences written by unqualified Turkers (which resulted in 4851 non-empty sentences). This resulted in low-quality data which required further verification. Since this is too time-consuming for expert annotators, we then paid another $28.80 to get each of these sentences subsequently checked by three Turkers for grammaticality, resulting in 2222 sentences which received at least two positive votes for grammaticality. With the qualification test approach, we paid $100.00 to get 5000 sentences written. Based on our experiments on the set of 600 sentences, experts would judge over 97% of these sentences as correct, thus obviating the immediate need for further control. That is, it effectively costs more for non-qualified Turkers to produce sentences that are judged to be good than for qualified Turkers. Furthermore, their sentences will probably be of lower quality even after they have been judged acceptable.

5 A corpus of captions for Flickr photos

Encouraged by the success of the qualification test approach, we extended our corpus to contain 8000 images collected from Flickr. We again paid the Turkers $0.10 to annotate five images. Our data set consists of 8108 hand-selected images from Flickr, depicting actions and events (rather than images depicting scenery and mood). These images are more likely to require full sentence descriptions than the PASCAL images. We chose six large Flickr groups[6] and downloaded a few thousand images from each, giving us a total of 15,000 candidate images. We removed all black and white or sepia images as well as images containing photographer signatures or seals. Next, we manually identified pictures that depicted the actions of people or animals. For example, we kept images of people walking in parks, but not of

[6]The groups: strangers!, Wild-Child (Kids in Action), Dogs in Action (Read the Rules), Outdoor Activities, Action Photography and Flickr-Social (two or more people in the photo)

Without qualification test
(1) lady with birds
(2) Some parrots are have speaking skill.
(3) A lady in their dining table with birds on her shoulder and head.
(4) Asian woman with two cockatiels, on shoulder
head, room with oak cabinets.,
(5) The lady loves the parrot
With qualification test
(1) A woman has a bird on her shoulder, and another bird on her head
(2) A woman with a bird on her head and a bird on her shoulder.
(3) A women sitting at a dining table with two small birds sitting on her.
(4) A young Asian woman sitting at a kitchen
table with a bird on her head and another on her shoulder.
(5) Two birds are perched on a woman sitting in a kitchen.

Figure 5: Comparison of captions written by Turkers with and without qualification test

empty parks; we kept several people posing, but not a close-up of a single person.[7] Each HIT asked Turkers to describe five images. We required the qualification test and US residency. Average completion time was a little above 3 minutes for 5 sentences. The corpus was annotated in 284 hours[8], at a total cost of $812.00 (plus Amazon's 10% fee).

6 Related work and conclusions

Related work MTurk has been used for many different NLP and vision tasks (Tietze et al., 2009; Zaidan and Callison-Burch, 2009; Snow et al., 2008; Sorokin and Forsyth, 2008). Due to the noise inherent in non-expert annotations, many other attempts at quality control have been made. Kittur et al. (2008) solicit ratings about different aspects of Wikipedia articles. At first they receive very noisy results, due to Turkers' not paying attention when completing the task or specifically trying to cheat the requester. They remade the task, this time starting by asking the Turkers verifiable questions, speculating that the users would produce better quality responses when they suspect their answers will be checked. They also added a question that required the Turkers to comprehend the content of the Wikipedia article. With this new setup, they find that the quality greatly increases and carelessness is reduced. Kaisser and Lowe (2008) collected question and answer pairs by presenting Turkers with a question and telling them to copy and paste from a document of text they know to contain the answer. They achieve a good but far from perfect interannotator agreement based on the extracted answers. We speculate that the quality would be much worse if the Turkers wrote the sentences themselves. Callison-Burch (2009) asks Turkers to produce translations when given reference sentences in other languages. Overall, he finds find that Turkers produce better translations than machine translation systems. To eliminate translations from Turkers who simply put the reference sentence into an online translation website, he performs a follow-up task, where he asks other Turkers to vote on if they believe that sentences were generated using an online translation system. Mihalcea and Strapparava (2009) ask Turkers to produce 4-5 sentence opinion paragraphs about the death penalty, about abortion and describing a friend. They report that aside from a small number of invalid responses, all of the paragraphs were of good quality and followed their instructions. Their success is surprising to us because they do not report using a qualification test, and when we did this our responses contained a large amount of incorrect English spelling and grammar.

The TurKit toolkit (Little et al., 2009) provides another approach to improving the quality of MTurk annotations. Their iterative framework allows the requester to set up a series of tasks that first solicits text annotations from Turkers and then asks other Turkers to improve the annotations. They report successful results using this methodology, but we chose

[7] Our final data set consists of 1482 pictures from action photography, 1904 from dogs, 776 from flickr-social, 916 from outdoor, 1257 from strangers and 1773 from wild-child.

[8] Note that the annotation process scaled pretty well, considering that annotating more than eight times the number of images took only 31 hours longer.

to stick with simply using the qualification test because it achieves the desired results already. Furthermore, although using TurKit would have probably done away with our few remaining grammar and spelling mistakes, it may have caused the captions for an image to be a little too similar, and we value a diversity in the use of words and points of view.

Our experiences We have described our experiences in using Amazon's Mechanical Turk in the first half of 2009 to create a corpus of images annotated with descriptive sentences. We implemented two different approaches to quality control: first, we did not impose any restrictions on who could write image descriptions. This was then followed by a second set of MTurk tasks where Turkers had to judge the quality of the sentences generated in our initial Turk experiments. This approach to quality control would be cost-effective if the initial data were not too noisy and the subsequent judgments were accurate and cheap. However, this was not the case, and quality control on the judgments in the form of control items turned out to result in even lower accuracy. We then repeated our data collection effort, but required that Turkers live in the US and take a brief qualification test that we created to test their English. This is cost-effective if English proficiency can be accurately assessed in such a brief qualification test. We found that the latter approach was indeed far cheaper, and produced significantly better data. We did not set out to run a scientific experiment comparing different strategies of how to collect linguistic data on Turk, and therefore there may be multiple explanations for the effects we observe. Nevertheless, our experience indicates strongly that even very simple prescreening measures can provide very effective quality control.

We also extended our corpus to include 8000 images collected from Flickr. We hope to release this data to the public for future natural language processing and computer vision research.

Recommended practices for using MTurk in NLP Our experience indicates that with simple prescreening, linguistic data can be elicited fairly cheaply and rapidly from crowd-sourcing services such as Mechanical Turk. However, many applications may require more control over where the data comes from. Even though NLP data collection differs fundamentally from psycholinguistic experiments that may elicit production data, our community will typically also need to know whether data was produced by native speakers or not. Until MTurk provides a better mechanism to check the native language of its workers, linguistic data collection on MTurk will have to rely on potentially very noisy input.

Acknowledgements

This research was funded by NSF grant IIS 08-03603 *INT2-Medium: Understanding the Meaning of Images*. We are grateful for David Forsyth's advice and for Alex Sorokin's support with MTurk.

References

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of EMNLP 2009*.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2008. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/.

Michael Kaisser and John Lowe. 2008. Creating a research collection of question answer sentence pairs with amazons mechanical turk. In *LREC 2008*.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of SIGCHI 2008*.

Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. 2009. Turkit: tools for iterative tasks on mechanical turk. In *HCOMP '09: Proceedings of the ACM SIGKDD Workshop on Human Computation*.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*.

Alexander Sorokin and David Forsyth. 2008. Utility data annotation with amazon mechanical turk. In *Computer Vision and Pattern Recognition Workshop*.

Martin I. Tietze, Andi Winterboer, and Johanna D. Moore. 2009. The effect of linguistic devices in information presentation messages on comprehension and recall. In *Proceedings of ENLG 2009*.

Omar F. Zaidan and Chris Callison-Burch. 2009. Feasibility of human-in-the-loop minimum error rate training. In *Proceedings of EMNLP 2009*.

Are all of the words correctly spelled and correctly used?	Is the sentence grammatically correct?
A group of children playing with thier toys (N)	A man giving pose to camera. (N)
He accepts the crowd's praise graciously. (Y)	The white sheep walks on the grass. (Y)
The coffee is kept at a very hot temperture. (N)	She is good woman. (N)
A green car is parked in front of a resturant. (N)	He should have talk to him. (N)
An orange cat sleeping with a dog that is much larger then it. (N)	He has many wonderful toy. (N)
I ate a tasty desert after lunch. (N)	He sended the children home to their parents. (N)
A group of people getting ready for a surprise party. (Y)	The passage through the hills was narrow. (Y)
A small refrigerator filled with colorful fruits and vegetables. (Y)	A sleeping dog. (Y)
Two men fly by in a red plain. (N)	The questions on the test was difficult. (N)
A causal picture of a man and a woman. (N)	In Finland, we are used to live in a cold climate. (N)
Three men are going out for a special occasion. (Y)	Three white sheeps graze on the grassy field. (N)
Woman eatting lots of food. (N)	Between you and me, this is wrong. (Y)
Dyning room with chairs. (N)	They are living there during six months. (N)
A woman recieving a package. (N)	I was given lots of advices about buying new furnitures. (N)
This is a relatively uncommon occurance. (Y)	A horse being led back to it's stall. (N)

Table 3: The spelling and grammar portions of the qualification test. The test may be found on MTurk by searching for the qualification entitled "Image Annotation Qualification".

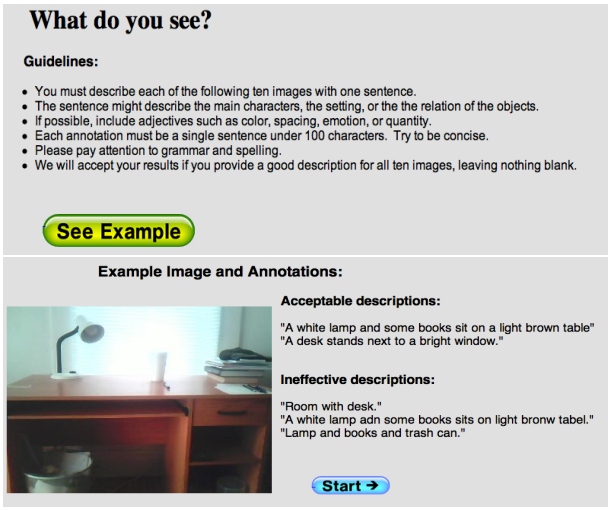

Figure 6: Screenshot of the image annotation instructions: guidelines (top) and examples (bottom).

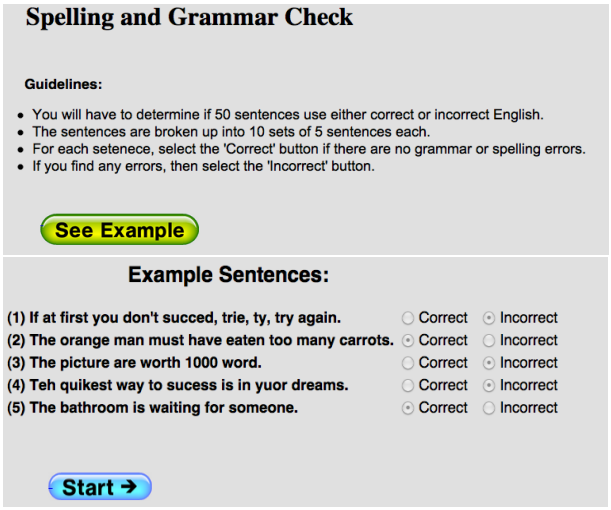

Figure 7: Screenshot of the quality control test instructions: guidelines (top) and examples (bottom).