Annotating Named Entities in Twitter Data with Crowdsourcing

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller and Justin Martineau
Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore MD 21250
(finin,willm1,anandk1,nick6,jm1)@umbc.edu

Mark Dredze
Human Language Technology Center of Excellence
Johns Hopkins University
Baltimore MD 21211
mdredze@cs.jhu.edu

Abstract

We describe our experience using both Amazon Mechanical Turk (MTurk) and CrowdFlower to collect simple named entity annotations for Twitter status updates. Unlike most genres that have traditionally been the focus of named entity experiments, Twitter is far more informal and abbreviated. The collected annotations and annotation techniques will provide a first step towards the full study of named entity recognition in domains like Facebook and Twitter. We also briefly describe how to use MTurk to collect judgements on the quality of "word clouds."

1 Introduction and Dataset Description

Information extraction researchers commonly work on popular formal domains, such as news articles. More diverse studies have included broadcast news transcripts, blogs and emails (Strassel et al., 2008). However, extremely informal domains, such as Facebook, Twitter, YouTube or Flickr are starting to receive more attention. Any effort aimed at studying these informal genres will require at least a minimal amount of labeled data for evaluation purposes.

This work details how to efficiently annotate large volumes of data, for information extraction tasks, at low cost using MTurk (Snow et al., 2008; Callison-Burch, 2009). This paper describes a case study for information extraction tasks involving short, informal messages from Twitter. Twitter is a large multi-user site for broadcasting short informal messages. Twitter is an extreme example of an informal genre (Java et al., 2007) as users frequently abbreviate their posts to fit within the specified limit. Twitter is a good choice because it is very popular: Twitter users generate a tremendous number of status updates (tweets) every day[1]. This is a good genre to work on named entity extraction since many tweets refer to and contain updates about named entities.

Our Twitter data set has over 150 million tweets from 1.5 million users collected over a period of three years. Tweets are unlike formal text. They are limited to a maximum of 140 characters, a limit originally set to allow them to fit into an SMS message. Consequently, the use of acronyms and both standard and non-standard abbreviations (e.g., *b4* for before and *ur* for your) are very common. Tweets tend to be telegraphic and often consist of sentence fragments or other ungrammatical sequences. Normal capitalization rules (e.g., for proper names, book titles, etc.) are commonly ignored.

Furthermore, users have adopted numerous conventions including hashtags, user mentions, and retweet markers. A hashtag (e.g., #earthquake) is a token beginning with a '#' character that denotes one of the topic of a status. Hashtags can be used as pure metadata or serve both as a word and as metadata, as the following two examples show.

- EvanEcullen: #chile #earthquake #tsunami They heard nothing of a tsunami until it slammed into their house with an unearthly http://tl.gd/d798d
- LarsVonD: Know how to help #Chile after the #Earthquake

[1] Pingdom estimated that there were nearly 40 million tweets a day in January 2010 (pingdom.com, 2010).

(1) *report from the economist: #chile counts the cost of a devastating earthquake and makes plans for recovery. http://bit.ly/dwoQMD* Note: "the economist" was not recognized as an ORG.
(2) *how come when george bush wanted to take out millions for the war congress had no problem...but whe obama wants money for healthcare the ...* Note: Both "george bush" and "obama" were missed as PERs.
(3) *RT @woodmuffin: jay leno interviewing sarah palin: the seventh seal starts to show a few cracks* Note: RT (code for a re-tweet) was mistaken as a position and sarah palin missed as a person.

Table 1: Standard named entity systems trained on text from newswire articles and other well formed documents lose accuracy when applied to short status updates.

The Twitter community also has a convention where user names preceded by an @ character (known as "mentions") at the beginning of a status indicate that it is a message directed at that user. A user mention in the middle of a message is interpreted as a general reference to that user. Both uses are shown in this status:

- paulasword: @obama quit calling @johnboener a liar, you liar

The token RT is used as a marker that a person is forwarding a tweet originally sent by another user. Normally the re-tweet symbol begins the message and is immediately followed by the user mention of the original author or sometimes a chain of re-tweeters ending with the original author, as in

- politicsiswar: RT @KatyinIndy @SamiShamieh: Ghost towns on rise under Obama http://j.mp/cwJSUg #tcot #gop (Deindustrialization of U.S.- Generation Zero)

Finally, "smileys" are common in Twitter statuses to signal the users' sentiment, as in the following.

- sallytherose: Just wrote a 4-page paper in an hour and a half. BOiiiiii I'm getting good at this. :) Leftover Noodles for dinner as a reward. :D

The Twitter search service also uses these to retrieve tweets matching a query with positive or negative sentiment.

Typical named entity recognition systems have been trained on formal documents, such as news

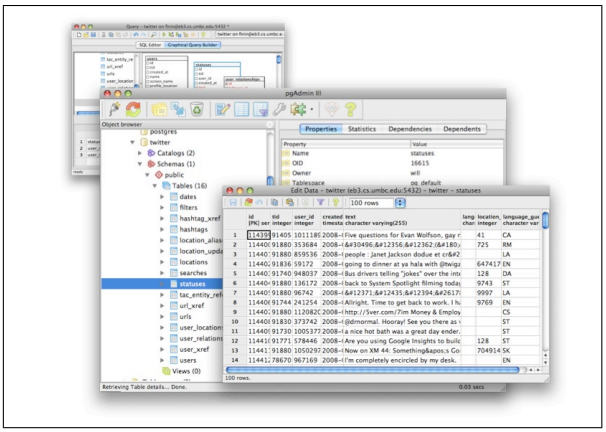

Figure 1: Our Twitter collection is stored in a relational database and also in the Lucene information retrieval system.

wire articles. Their performance on text from very different sources, especially informal genres such as Twitter tweets or Facebook status updates, is poor. In fact, "Systems analyzing correctly about 90% of the sequences from a journalistic corpus can have a decrease of performance of up to 50% on more informal texts." (Poibeau and Kosseim, 2001) However, many large scale information extraction systems require extracting and integrating useful information from online social networking sources that are informal such as Twitter, Facebook, Blogs, YouTube and Flickr.

To illustrate the problem we applied both the NLTK (Bird et al., 2009) and the Stanford named entity recognizers (Finkel et al., 2005) without retraining to a sample Twitter dataset with mixed results. We have observed many failures, both false positives and false negatives. Table 1 shows some examples of these.

2 Task design

We developed separate tasks on CrowdFlower and MTurk using a common collection of Twitter statuses and asked workers to perform the same annotation task in order to fully understand the features that each provides, and to determine the total amount of work necessary to produce a result on each service. MTurk has the advantage of using standard HTML and Javascript instead of CrowdFlower's CML. However MTurk has inferior data verification, in that the service only provides a threshold on worker agreement as a form of quality control.

This is quite poor when tasks are more complicated than a single boolean judgment, as with the case at hand. CrowdFlower works across multiple services and does verification against gold standard data, and can get more judgements to improve quality in cases where it's necessary.

3 Annotation guidelines

The task asked workers to look at Twitter individual status messages (tweets) and use a toggle button to tag each word with person (PER), organization (ORG), location (LOC), or "none of the above" (NONE). Each word also had a check box (labeled ???) to indicate that uncertainty. We provided the workers with annotation guidelines adapted from the those developed by the Linguistic Data Consortium (Linguistic Data Consortium – LCTL Team, 2006) which were in turn based on guidelines used for MUC-7 (Chinchor and Robinson, 1997).

We deliberately kept our annotation goals simple: We only asked workers to identify three basic types of named entities.

Our guidelines read:

> An entity is a object in the world like a place or person and a *named entity* is a phrase that uniquely refers to an object by its proper name (Hillary Clinton), acronym (IBM), nickname (Opra) or abbreviation (Minn.).
>
> *Person* (PER) entities are limited to humans (living, deceased, fictional, deities, ...) identified by name, nickname or alias. Don't include titles or roles (Ms., President, coach). Include suffix that are part of a name (e.g., Jr., Sr. or III).
>
> *Organization* (ORG) entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure. Some examples are businesses (Bridgestone Sports Co.), stock ticker symbols (NASDAQ), multinational organizations (European Union), political parties (GOP) non-generic government entities (the State Department), sports teams (the Yankees), and military groups (the Tamil Tigers). Do not tag 'generic' entities like "the government" since these are not unique proper names referring to a specific ORG.
>
> *Location* (LOC) entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments.

We instructed annotators to ignore other types of named entities, e.g., events (World War II), products (iPhone), animals (Cheetah), inanimate objects and monetary units (the Euro) and gave them four principles to follow when tagging:

- Tag words according to their *meaning* in the context of the tweet.
- Only tag *names*, i.e., words that directly and uniquely refer to entities.
- Only tag names of the types *PER*,*ORG*, and *LOC*.
- Use the *???* checkbox to indicate uncertainty in your tag.

3.1 Data selection

We created a "gold standard" data set of about 400 tweets to train and screen workers on MTurk, to salt the MTurk data with worker evaluation data, for use on CrowdFlower, and to evaluate the performance of the final NER system after training on the crowdsourced annotations. We preselected tweets to annotate using the NLTK named entity recognizer to select statuses that were thought to contain named entities of the desired types (PER, ORG, LOC).

Initial experiments suggested that a worker can annotate about 400 tweets an hour. Based on this, we loaded each MTurk Human Intelligence Tasks (HIT) with five tweets, and paid workers five cents per HIT. Thus, if we require that each tweet be annotated by two workers, we would be able to produce about 4,400 raw annotated tweets with the $100 grant from Amazon, accounting for their 10% overhead price.

3.2 CrowdFlower

We also experimented with CrowdFlower, a crowdsourcing service that uses various worker channels like MTurk and SamaSource[2] and provides an enhanced set of management and analytic tools. We were interested in understanding the advantages and disadvantages compared to using MTurk directly.

[2] http://www.samasource.org/

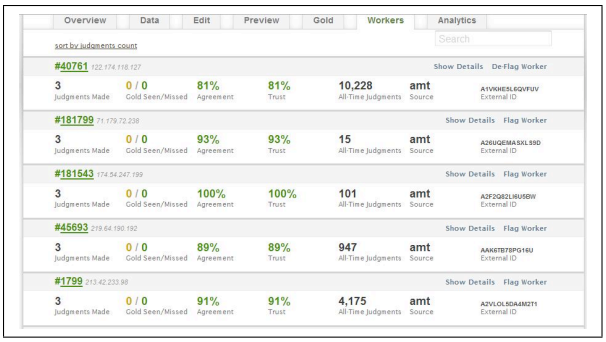

Figure 2: CrowdFlower is an enhanced service that feeds into MTurk and other crowdsourcing systems. It provides convenient management tools that show the performance of workers for a task.

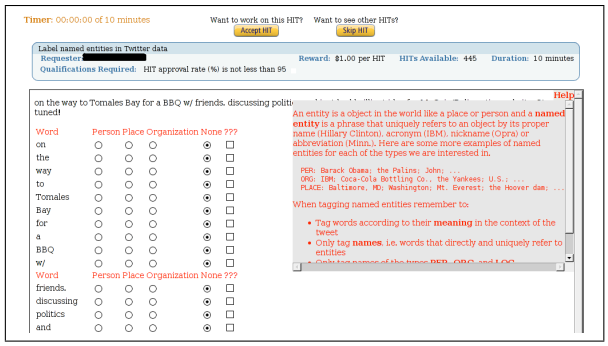

Figure 3: In the MTurk interface a tweet is shown in its entirety at the top, then a set of radio buttons and a checkbox is shown for each word of the tweet. These allow the user to pick the annotation for each word, and indicate uncertainty in labeling.

We prepared a basic front-end for our job using the CrowdFlower Markup Language (CML) and custom JavaScript. We used the CrowdFlower interface to calibrate our job and to decide the pay rate. It considers various parameters like amount of time required to complete a sample task and the desired accuracy level to come up with a pay rate.

One attractive feature lets one provide a set of "gold standard" tasks that pair data items with correct responses. These are automatically mixed into the stream of regular tasks that workers process. If a worker makes errors in one of these gold standard tasks, she gets immediate feedback about her error and the correct answer is shown. CrowdFlower claims that error rates are reduced by a factor of two when gold standards are used(crowdflower.com, 2010). The interface shown in Figure 2 shows the number of gold tasks the user has seen, and how many they have gotten correct.

CrowdFlower's management tools provides a detailed analysis of the workers for a job, including the trust level, accuracy and past accuracy history associated with each worker. In addition, the output records include the geographical region associated with each worker, information that may be useful for some tasks.

3.3 MTurk

The current iteration of our MTurk interface is shown in Figure 3. Each tweet is shown at the top of the HIT interface so that it can easily be read for context. Then a table is displayed with each word of the tweet down the side, and radio buttons to pick what kind of entity each word is. Every ten rows, the header is repeated, to allow the worker to scroll down the page and still see the column labels. The interface also provides a checkbox allows the worker to indicate uncertainty in labeling a word.

We expect that our data will include some tricky cases where an annotator, even an experienced one, may be unsure whether a word is part of a named entity and/or what type it is. For example, is 'Baltimore Visionary Art Museum' a LOC followed by a three word ORG, or a four-word ORG? We considered and rejected using hierarchical named entities in order to keep the annotation task simple. Another example that might give an annotator pause is a phrase like 'White House' can be used as a LOC or ORG, depending on the context.

This measure can act as a measure of a worker's quality: if they label many things as "uncertain", we might guess that they are not producing good results in general. Also, the uncertainty allows for a finer-grained measure of how closely the results from two workers for the same tweet match: if the workers disagree on the tagging of a particular word, but agree that it is not certain, we could decide that this word is a bad example and not use it as training data.

Finally, a help screen is available. When the user mouses over the word "Help" in the upper right, the guidelines discussed in Section 3 are displayed. The screenshot in Figure 3 shows the help dialog expanded.

The MTurk interface uses hand-written Javascript to produce the table of words, radio buttons, and

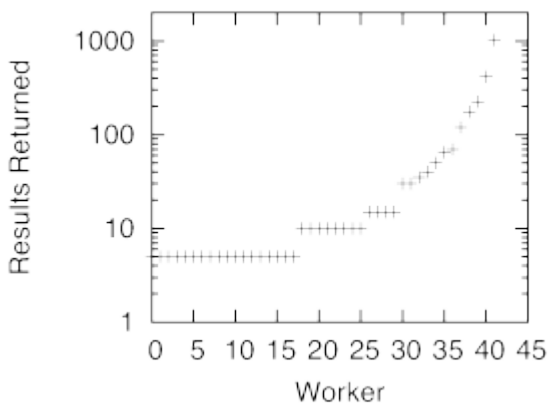

Figure 4: Only about one-third of the workers did more than three HITs and a a few prolific workers accounted for most of our data.

checkboxes. The form elements have automatically generated names, which MTurk handles neatly. Additional Javascript code collects location information from the workers, based on their IP address. A service provided by Geobytes[3] provides the location data.

4 Results from MTurk

Our dataset was broken into HITs of four previously unlabeled tweets, and one previously labeled tweet (analogous to the "gold" data used by CrowdFlower). We submitted 251 HITs, each of which was to be completed twice, and the job took about 15 hours. Total cost for this job was $27.61, for a total cost per tweet of about 2.75 cents each (although we also paid to have the gold tweets annotated again). 42 workers participated, mostly from the US and India, with Australia in a distant third place. Most workers did only a single HIT, but most HITs were done by a single worker. Figure 4 shows more detail.

After collecting results from MTurk, we had to come up with a strategy for determining which of the results (if any) were filled randomly. To do this, we implemented an algorithm much like Google's PageRank (Brin and Page, 1998) to judge the amount of inter-worker agreement. Pseudocode for our algorithm is presented in Figure 5.

This algorithm doesn't strictly measure worker quality, but rather worker agreement, so it's important to ensure that the workers it judges as having high agreement values are actually making high-quality judgements. Figure 6 shows the worker agreement values plotted against the number of results a particular worker completed. The slope of this plot (more results returned tends to give higher scores) is interpreted to be because practice makes perfect: the more HITs a worker completes, the more experience they have with the task, and the more accurate their results will be.

So, with this agreement metric established, we set out to find out how well it agreed with our expectation that it would also function as a quality metric. Consider those workers that completed only a single HIT (there are 18 of them): how well did they do their jobs, and where did they end up ranked as a result? Since each HIT is composed of five tweets,

[3] http://www.geobytes.com/

WORKER-AGREE : $results \rightarrow scores$

```
1  worker_ids ← ENUMERATE(KEYS(results))
   ▷ Initialize A
2  for worker1 ∈ worker_ids
3      do for worker2 ∈ worker_ids
4             do A[worker1, worker2]
                  ← SIMILARITY(results[worker1],
                               results[worker2])
   ▷ Normalize columns of A so that they sum to 1 (elided)
   ▷ Initialize x to be normal: each worker
     is initially trusted equally.
5  x ← ⟨1/√n, ..., 1/√n⟩
   ▷ Find the largest eigenvector of A, which
     corresponds to the agreement-with-group
     value for each worker.
6  i ← 0
7  while i < max_iter
8      do x_new ← NORMALIZE(A × x)
9         diff ← x_new − x
10        x = x_new
11        if diff < tolerance
12          then break
13        i ← i + 1
14 for workerID, workerNum ∈ worker_ids
15     do scores[workerID] ← x[workerNum]
16 return scores
```

Figure 5: Intra-worker agreement algorithm. MTurk results are stored in an associative array, with worker IDs as keys and lists of HIT results as values, and worker scores are floating point values. Worker IDs are mapped to integers to allow standard matrix notation. The Similarity function in line four just returns the fraction of HITs done by two workers where their annotations agreed.

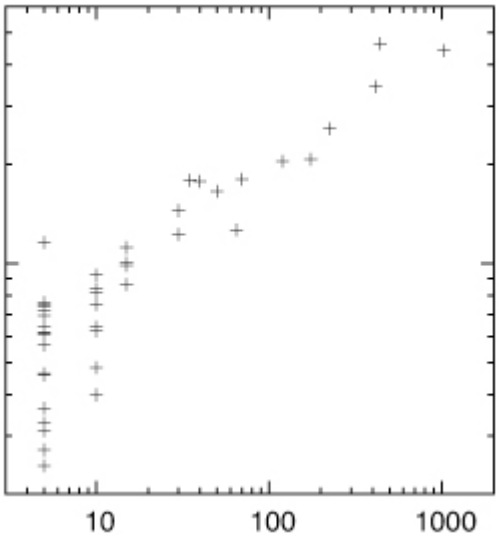

Figure 6: This log-log plot of worker agreement scores versus the number of results clearly shows that workers who have done more HITs have better inter-annotator agreement scores.

Score	0.0243	Score	0.0364	Score	0.0760
Trying	org	Newbie	person	Trying	none
to	org	here	none	out	none
decide	org	nice	none	TwittEarth	org
if	org	to	none	-	none
it's	org	meet	none	Good	none
worth	place	you	none	graphics.	none
hanging	org	all	none	Fun	none
around	org			but	none
until	org			useless.	none
the	none			(URL)	none
final	org				
implosion	org				

Figure 7: These sample annotations represent the range of worker quality for three workers who did only one HIT. The first is an apparently random annotation, the second a plausible but incorrect one, and the third a correct annotation. Our algorithm assigned these workers scores aligned with their product quality.

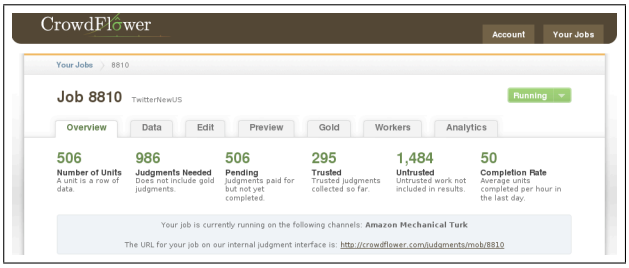

Figure 8: CrowdFlower provides good interfaces to manage crowdsourcing tasks. This view lets us to monitor the number of judgements in each category.

even such a small sample can contain a lot of data.

Figure 7 shows a sample annotation for three tweets, each from a worker who did only one HIT, and the ranking that the worker received for doing that annotation. The worst scoring one is apparently a random fill: there's no correlation at all between the answers and the correct ones. The middle tweet is improved: "Newbie" isn't a person in this context, but it's a mistake a non-native speaker might make, and everything else is right, and the score is higher. The last tweet is correctly labeled within our parameters, and scores the highest. This experiment shows that our agreement metric functions well as a correctness metric.

Also of interest is the raw effectiveness of MTurk workers; did they manage to tag tweets as well as our experts? After investigating the data, our verdict is that the answer is not quite—but by carefully combining the tags that two people give the same tweet it is possible to get good answers nevertheless, at much lower cost than employing a single expert.

5 Results from CrowdFlower

Our CrowdFlower task involved 30 tweets. Each tweet was further split into tokens resulting in 506 units as interpreted by CrowdFlower's system. We required a total 986 judgments. In addition, we were required to generate thirteen "gold" data, which is the minimum required by the service. Every gold answer has an optional text with it to inform workers why we believe our answer is the correct one and theirs is incorrect. This facilitates gradually training workers up to the point where they can provide reliably correct results. Figure 8 shows the interface CrowdFlower provides to monitor the number of judgements in each category.

We used the calibration interface that CrowdFlower provides to fix the price for our task (Figure 9). It considers various parameters like the time required per unit and desired accuracy level, and also adds a flat 33% markup on the actual labor costs. We divided the task into a set of assignments where each assignment had three tweets and was paid five cents. We set the time per unit as 30 seconds, so, based on the desired accuracy level and markup overhead, our job's cost was $2.19. This comes to $2 hourly pay per worker, assuming they take the whole 30 sec-

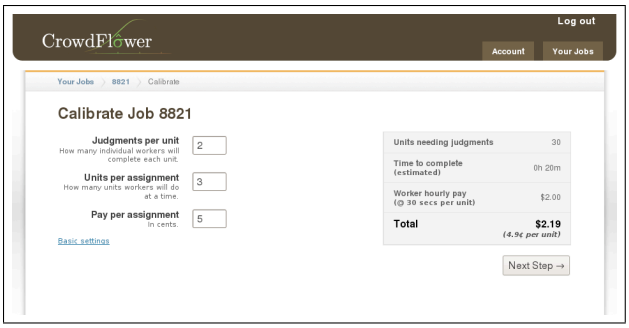

Figure 9: CrowdFlower has an interface that makes it easy to select an appropriate price for a task.

onds to complete the task.

6 Cloud Comparison

MTurk can also be used to efficiently evaluate results requiring human judgments. We implemented an additional HIT to evaluate a new technique we developed to generate "word clouds." In this task workers choose which of two word clouds generated from query results by two different algorithms provides a more useful high level description that can highlight important features and opinions about the query topic.

Evaluating how well a set of words describes and highlights the important features and opinions pertaining to the subject of the query is subjective, which necessitates human evaluations. MTurk workers were given two word clouds, one from our technique and the other from a baseline relevance feedback technique (Rocchio (Rocchio, 1971)), for each query. Queries were shown with a short descriptive blurb to disambiguate it from possible alternatives, reveal the intent of the user who created the query, and provide a short description of it for workers who were unfamiliar with the query subject. Wikipedia links were provided, when applicable, for anyone needing further information about the query subject. Workers were asked to use a slider to determine which cloud better represented the key concepts related to the query. The slider would snap into one of eleven positions, which were labeled with value judgments they represented. The center value indicates that the two clouds were equally good. Figure 10 shows the final query interface.

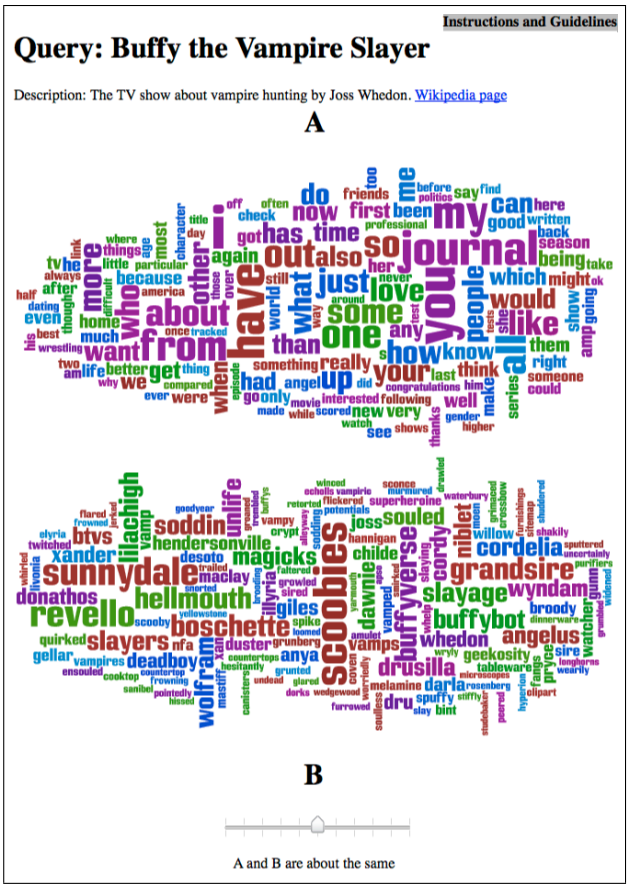

Figure 10: MTurk workers were asked which word cloud they thought best represented returned the results of a query, in this case "Buffy the Vampire Slayer".

6.1 Results

Since MTurk workers are paid per task they complete, there is an incentive to do low quality work and even to randomly guess to get tasks done as fast as possible. To ensure a high quality evaluation we included in every batch of five queries a quality control question. Quality control questions were designed to look exactly like the regular cloud comparisons, but only one of the two clouds displayed was actually from the query in the description. The other word cloud was generated from a different query with no relation to the real query, and hand checked to make sure that anyone who was doing a respectable job would agree that the off-topic word cloud was a poor result for the query. If a worker's response indicated that the off topic cloud was as good as or better than the real cloud then they failed that control question, otherwise they passed.

We asked that twelve workers label each set of questions. We only used results from workers that answered at least seven control questions with an average accuracy rating of at least 75%. This left us with a pool of eight reliable workers with an average accuracy on control questions of about 91%. Every question was labeled by at least five different workers with a mode of seven.

Workers were not told which technique produced which cloud. Techniques were randomly assigned to either cloud A or B to prevent people from entering into a "cloud A is always better" mentality. The position of the quality control questions were randomly assigned in each set of five cloud comparisons. The links to the cloud images were anonymized to random numbers followed by the letter A or B for their position to prevent workers from guessing anything about either the query or the technique that generated the cloud.

We applied a filter to remove the query words from all word clouds. First of all, it would be a dead giveaway on the control questions. Second, the query words are already known and thus provide no extra information about the query to the user while simultaneously taking up the space that could be used to represent other more interesting words. Third, their presence and relative size compared to the baseline could cause users to ignore other features especially when doing a quick scan.

The slider scores were converted into numerical scores ranging from -5 to +5, with zero representing that the two clouds were equal. We averaged the score for each cloud comparison, and determined that for 44 out of 55 clouds workers found our technique to be better than the baseline approach.

6.2 Issues

We faced some issues with the CrowdFlower system. These included incorrect calibration for jobs, errors downloading results from completed jobs, price displayed on MTurk being different that what was set through CrowdFlower and gold standard data not getting stored on CrowdFlower system. Another problem was with the system's 10-token limit on gold standards, which is not yet resolved at the time of this writing. On the whole, the CrowdFlower team has been very quick to respond to our problems and able to correct the problems we encountered.

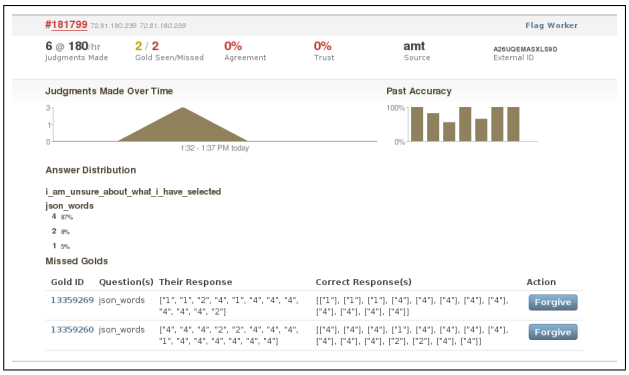

Figure 11: Statistics for worker #181799. The interface has an option to "forgive" the worker for missing gold and an option to "flag" the worker so that the answers are excluded while returning the final set of judgments. It also displays workers ID, past accuracy and source, e.g. MTurk.

6.3 Live Analytics

CrowdFlower'a analytics panel facilitates viewing the live responses. The trust associated with each worker can be seen under the workers panel. Workers who do a large amount of work with low trust are likely scammers or automated bots. Good gold data ensures that their work is rejected. The system automatically pauses a job when the ratio of untrusted to trusted judgments exceeds a certain mark. This was particularly helpful for us to rectify some of our gold data. Currently, the job is being completed with 61% accuracy for gold data. This could be due to the current issue we are facing as described above. It's also possible to view statistics for individual workers, as shown in Figure 11.

7 Conclusion

Crowdsourcing is an effective way to collect annotations for natural language and information retrieval research. We found both MTurk and CrowdFlower to be flexible, relatively easy to use, capable of producing usable data, and very cost effective.

Some of the extra features and interface options that CrowdFlower provided were very useful, but did their were problems with their "gold standard" agreement evaluation tools. Their support staff was very responsive and helpful, mitigating some of these problems. We were able to duplicate some of the "gold standard" functionality on MTurk directly by generating our own mix of regular and quality control queries. We did not attempt to provide im-

mediate feedback to workers who enter a wrong answer for the "gold standard" queries, however.

With these labeled tweets, we plan to train an entity recognizer using the Stanford named entity recognizer[4], and run it on our dataset. After using this trained entity recognizer to find the entities in our data, we will compare its accuracy to the existing recognized entities, which were recognized by an ER trained on newswire articles. We will also attempt to do named entity linking and entity resolution on the entire corpus.

We look forward to making use of the data we collected in our research and expect that we will use these services in the future when we need human judgements.

Acknowledgments

This work was done with partial support from the Office of Naval Research and the Johns Hopkins University Human Language Technology Center of Excellence. We thank both Amazon and Dolores Labs for grants that allowed us to use their systems for the experiments.

References

S. Bird, E. Klein, and E. Loper. 2009. *Natural language processing with Python*. Oreilly & Associates Inc.

S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference (WWW 1998)*.

C. Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazons Mechanical Turk. In *Proceedings of EMNLP 2009*.

N. Chinchor and P. Robinson. 1997. MUC-7 named entity task definition. In *Proceedings of the 7th Message Understanding Conference*. NIST.

crowdflower.com. 2010. The error rates without the gold standard is more than twice as high as when we do use a gold standard. http://crowdflower.com/general/examples. Accessed on April 11, 2010.

J.R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, volume 100, pages 363–370.

[4]http://nlp.stanford.edu/software/CRF-NER.shtml

A. Java, X. Song, T. Finin, and B. Tseng. 2007. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM.

Linguistic Data Consortium – LCTL Team. 2006. Simple named entity guidelines for less commonly taught languages, March. Version 6.5.

pingdom.com. 2010. Twitter: Now more than 1 billion tweets per month. http://royal.pingdom.com-/2010/02/10/twitter-now-more-than-1-billion-tweets-per-month/, February. Accessed on February 15, 2010.

T. Poibeau and L. Kosseim. 2001. Proper Name Extraction from Non-Journalistic Texts. In *Computational linguistics in the Netherlands 2000: selected papers from the eleventh CLIN Meeting*, page 144. Rodopi.

J. Rocchio. 1971. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall.

R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

Stephanie Strassel, Mark Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda. 2008. Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.