MTurk Crowdsourcing: A Viable Method for Rapid Discovery of Arabic Nicknames?

Chiara Higgins
George Mason University
Fairfax, VA. 22030, USA
chiara.higgins@gmail.com

Elizabeth McGrath
MITRE
McLean, VA. 20112, USA
emcgrath@mitre.org

Lailla Moretto
MITRE
McLean, VA. 20112, USA
lmoretto@mitre.org

Abstract

This paper presents findings on using crowdsourcing via Amazon Mechanical Turk (MTurk) to obtain Arabic nicknames as a contribution to exiting Named Entity (NE) lexicons. It demonstrates a strategy for increasing MTurk participation from Arab countries. The researchers validate the nicknames using experts, MTurk workers, and Google search and then compare them against the Database of Arabic Names (DAN). Additionally, the experiment looks at the effect of pay rate on speed of nickname collection and documents an advertising effect where MTurk workers respond to existing work batches, called Human Intelligence Tasks (HITs), more quickly once similar higher paying HITs are posted.

1 Introduction

The question this experiment investigates is: can MTurk crowdsourcing add undocumented nicknames to existing Named Entity (NE) lexicons?

This experiment seeks to produce nicknames to add to DAN Version 1.1, which contains 147,739 lines of names. While DAN does not list nicknames as a metadata type, it does include some commonly known nicknames.

1.1 Traditional collection methods are costly

According to DAN's website, administrators collect nicknames using a team of software engineers and native speakers. They also draw on a "large variety of sources including websites, corpora, books, phone directories, dictionaries, encyclopedias, and university rosters" (Halpern, 2009). Collecting names by searching various media sources or employing linguists and native speakers is a massive effort requiring significant expenditure of time and money.

1.2 Crowdsourcing might work better

The experiment uses crowdsourcing via MTurk since it offers a web-based problem-solving model and quickly engages a large number of international workers at low cost. Furthermore, previous research shows the effectiveness of crowdsourcing as a method of accomplishing labor intensive natural language processing tasks (Callison-Burch, 2009) and the effectiveness of using MTurk for a variety of natural language automation tasks (Snow, Jurafsy, & O'Connor, 2008).

The experiment answers the following questions:

- Can we discover valid nicknames not currently in DAN?
- What do we need to pay workers to gather nicknames rapidly?
- How do we convey the task to guide non-experts and increase participation from Arab countries?

Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 89–92,
Los Angeles, California, June 2010.

2 Experiment Design

The experiment contains three main phases. First, nicknames are gathered from MTurk workers. Second, the collected names are validated via MTurk, internet searches, and expert opinion. Finally, the verified names are compared against the available list of names in the DAN.

2.1 Collecting nicknames on MTurk

In this phase, we open HITs on MTurk requesting workers to enter an Arabic nickname they have heard. In addition to writing a nickname, the workers input where they heard the name and their country of residence.

HIT instructions are kept simple and written in short sentences to guide non-experts and include a basic definition of a nickname. To encourage participation of native Arabic speakers, the instructions and search words are in Arabic as well as English. Workers are asked to input names in the Arabic alphabet, thus eliminating any worker who does not use Arabic often enough to warrant having an Arabic keyboard. Further clarifying the task, words highlighted in red, "Arabic alphabet", emphasize what the worker needs to do.

While seeking to encourage participation from Arab countries, we choose not to block participation from other countries since there are Arabic speakers and immigrants in many countries where Arabic is not the main language.

To evaluate the effect of pay rate on nickname collection rate, HITs have a variety of pay rates. HITs paying \$0.03 per HIT are kept up throughout the experiment, while HITs paying \$0.05 and finally \$0.25 are added later.

2.2 Nickname validation phase

Vetting the nicknames, involves a Google check and asking 3 experts and 5 MTurk workers to rate each name that is submitted in a valid format.

Each expert and MTurk worker has the opportunity to rate the likelihood the nickname would occur in the Arab world on a Likert scale (Strongly Agree, Agree, Neither Agree nor Disagree, Disagree, Strongly Disagree).

The entire validation process is completed twice, once paying the workers \$.01 per validation and once paying \$.05 per validation to allow us to further research the effect of pay on HIT collection rate.

The Google check vets the names to see if they occur on the web thus eliminating, any nicknames that are nowhere in print and therefore not currently necessary additions to NE lexicons.

2.3 Compare data to ground truth in DAN

The third phase is a search for exact matches for the validated nicknames in DAN to determine if they represent new additions to the lexicon.

3 Results

MTurk workers generated 332 nicknames during the course of this experiment. Because the initial collection rate was slower than expected, we validated and compared only the first 108 names to report results related to the usefulness of MTurk in nickname collection. Results involving pay and collection rate draw on the full data.

Based on self-reported data, approximately 35% of the respondents came from the Arabic speaking countries of Morocco, Egypt, Lebanon, Jordan, UAE, and Dubai. 46% were submitted from India, 13% from the U.S. and 5% elsewhere.

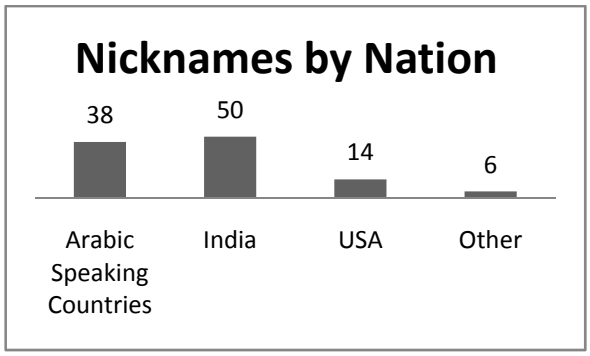

Figure 1. Nicknames by nation

3.1 Validation results

Each of the nicknames was verified by MTurk workers and three experts. On a five-point Likert scale with 1 representing strong disagreement and 5 showing strong agreement, we accepted 51 of the names as valid because the majority (3 of 5 MTurk workers and 2 of 3 experts) scored the name as 3 or higher.

One of the 51 names accepted by other means could not be found in a Google search leaving us with 50 valid nicknames.

Comparing the 50 remaining names to DAN we found that 11 of the valid names were already in the lexicon.

3.2 Effect of increased pay on responses

Holding everything else constant, we increased the worker's pay during nickname collection. On average, $0.03 delivered 9.8 names a day, for $0.05 we collected 25 names a day and for $0.25 we collected 100 names in a day.

We also posted one of our MTurk verification files two times, once at $0.01 per HIT and once at $0.05 per HIT, holding everything constant except the pay. Figure 2 shows the speed with which the two batches of HITs were completed. The results show not only an increased collection speed for the higher paying HITs, but also an increased collection speed for the existing lower paying HIT once the higher paying HITs were posted.

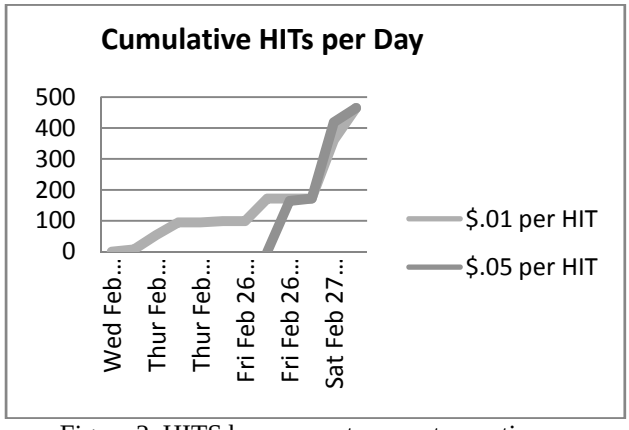

Figure 2. HITS by payment amount over time

4 Conclusions

As our most significant goal, we sought to investigate whether MTurk crowdsourcing could successfully collect undiscovered nicknames to add to an existing NE lexicon.

The results indicate that MTurk is a viable method for collecting nicknames; in the course of the experiment, we successfully produced 39 verified nicknames that we recommend adding to the DAN.

Another goal was to explore the effect of worker pay on HIT completion rate. Our initial collection rate, at $0.03 per HIT, was only 9.8 names per day. By increasing pay, we were able to speed up the process. At $0.05 per name, we increased the daily collection rate from 9.8 to 25, and by making the pay rate $0.25 we collected 100 names in a day. So increasing pay significantly improved collection speed.

While working with pricing for the verification HITs, we were able to quantify an "advertising effect" we had noticed previously where the posting of a higher paying HIT causes existing similar lower paying HITs to be completed more quickly as well. Further research could be conducted to determine a mix of pay rates that maximizes collection rate while minimizing cost.

Furthermore, the experiment shows that by using bilingual directions and requiring typing in Arabic, we were able to increase the participation from Arabic speaking countries. Based on our previous experience where we posted Arabic language related HITs in English only, Arab country participation on MTurk is minimal. Other researchers have also found little MTurk participation from Arabic speaking countries (Ross, Zaldivar, Irani, & Tomlinson, 2009). In this experiment, however, we received more than 35% participation from workers in Arabic speaking countries.

Acknowledgments

Thanks to the MITRE Corporation for providing the ground truth DAN data under their research

license agreement with the CJK Dictionary Institute. Also thanks to Trent Rockwood of MITRE for providing expert assistance in the Arabic language and on some technical issues.

References

Callison-Burch, C. (2009). Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. *Proceedings of the EMNLP*. Singapore.

Halpern, J. (2009). Lexicon-Driven Approach to the Recognition of Arabic Named Entities. *Second International Conference on Arabic Language Resources and Tools.* Cairo.

Ross, J., Zaldivar, A., Irani, L., & Tomlinson, B. (2009). *Who are the Turkers? Worker Demographics in Amazon.* Department of Informatics, University of California, Irvine.

Snow, R., Jurafsy, D., & O'Connor, B. (2008). Cheap and fast – but is it good?: Evaluating Non-Expert Annotations for Natural Language Tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* (pp. 254-263). Honolulu.