Can Crowdsourcing be used for Effective Annotation of Arabic?

Wajdi Zaghouani[1] **and Kais Dukes**[2]

(1) Carnegie Mellon University, Doha, Qatar
(2) School of Computing, University of Leeds, United Kingdom
E-mail: wajdiz@cmu.edu, sckd@leeds.ac.uk

Abstract

Crowdsourcing has been used recently as an alternative to traditional costly annotation by many natural language processing groups. In this paper, we explore the use of Amazon Mechanical Turk (AMT) in order to assess the feasibility of using AMT workers (also known as Turkers) to perform linguistic annotation of Arabic. We used a gold standard data set taken from the Quran corpus project annotated with part-of-speech and morphological information. An Arabic language qualification test was used to filter out potential non-qualified participants. Two experiments were performed, a part-of-speech tagging task in where the annotators were asked to choose a correct word-category from a multiple choice list and case ending identification task. The results obtained so far showed that annotating Arabic grammatical case is harder than POS tagging, and crowdsourcing for Arabic linguistic annotation requiring expert annotators could be not as effective as other crowdsourcing experiments requiring less expertise and qualifications.

Keywords: Crowdsourcing, Annotation, Arabic.

1. Introdcution

Crowdsourcing has been used recently as an alternative to traditional costly annotation by many natural language processing groups. In this paper, we explore the use of Amazon Mechanical Turk[1] (AMT) in order to assess the feasibility of using AMT workers (also known as Turkers) to perform linguistic annotation of Arabic, and in particular annotation of the Quran. Amazon's Mechanical Turk is not the only crowdsourcing framework, but since it has been widely used in research, we decided to use it in our experiment (Callison-Burch, 2009; Callison-Burch & Dredze,2010; Bloodgood, & Callison-Burch, 2010).

A key question is whether there are enough Turkers who can perform this task sufficiently well, compared to automatic annotation. In general, crowdsourcing is considered to be cheap and fast compared to more traditional approaches, but use of AMT may require careful consideration for certain linguistic tasks.

Mechanical Turk's potentials open new possibilities for annotating speech and text. In this paper, we consider an experiment that is used to evaluate the effectiveness of using Mechanical Turk to perform a specific type of annotation: simplified linguistic tagging of the Arabic text of the Quran. It is interesting to compare crowdsourcing using AMT to the more proven approach previously used to annotate the Quran, which involved a small set of dedicated volunteer Arabic experts collaborating over many months through an online message-board forum (The Quranic Arabic Corpus).[2]

The aim of the experiment was to understand the accuracy of using crowdsourcing using Mechanical Turk, and to find out if the two approaches could complement each other. Before conducting the experiment, and as described by Chamberlain et al. (2009) in his game with a purpose (GWAP) experiment, it was thought that paying for suggested corrections to part-of-speech tagging might encourage individuals with knowledge of the Arabic language to participate who might not otherwise. It may also allow for better quality of work and higher consistency over free volunteer annotation.

2. Related Work

2.1 Amazon Mechanical Turk

Mechanical Turk has been used for many natural language processing tasks in recent years (Callison-Burch, 2009; Callison-Burch & Dredze,2010; Bloodgood, & Callison-Burch, 2010). In (Su et al., 2007), Turkers performed the annotation tasks of hotel named-entity resolution and attribute extraction. The results were surprisingly accurate and close to the gold standard scores. (Snow et al., 2008) studied the accuracy of annotation by Turkers for various natural language processing tasks, including word-sense disambiguation, word similarity, temporal ordering of events and textual entailment. The results were also found to be encouraging, which suggests that the use of Mechanical Turk for such tasks may be feasible for the Arabic language.

2.2 The Quranic Arabic Corpus

The Gold standard analysis of the Quran, annotated by Arabic experts was used as a baseline to measure the quality of the annotation in the experiment. The Quranic Arabic Corpus is an open source project organized by the

[1] https://www.mturk.com
[2] http://corpus.quran.com/

Language Research Group at the University of Leeds. The aim of the project is to provide a richly annotated linguistic resource for researchers wanting to study the language of the Quran. The Quranic corpus provides annotation which shows the Arabic grammar, syntax and morphology for each word in the Quran (Dukes and Habash, 2010; Dukes 2013). The corpus is divided in two levels of analysis: morphological annotation and a syntactic Treebank (Dukes, Atwell and Sharaf, 2010).

Currently, the gold standard annotation is provided by volunteer annotators who are typically Arabic linguists or Quranic experts. Corrections to the online corpus can be easily made online by clicking on an Arabic word and then posting the desired suggestion which will be reviewed before being included in the corpus. Moreover, a message board was created to provide a discussion space for various issues and suggestions regarding the project.
Although developed using online collaborative annotation, the Quranic Arabic Corpus has undergone several continuous stages of correction since March 2009, and the morphological and syntactic analyses in the corpus are now believed to be highly accurate. The Quranic Corpus is therefore a good baseline to use for evaluation when measuring the accuracy of crowdsourcing for annotation of Arabic. For these experiments, we chose the first few hundred words of chapter 23 of the Quran. In general, the Quranic Arabic Corpus is still undergoing continued volunteer verification. However we chose this particular section of the Quran since this data has been re-verified by an expert Arabic linguist.

3. Crowdsourcing for Arabic Annotation

While there is many crowdsourcing projects as described in Wang et al. (2010), for the purpose of this project, we limit our experiment with the Amazon Mechanical Turk since it is widely used across the research community for tasks similar to ours.

The Amazon Mechanical Turk service provides an online job market place for requesters (people offering tasks) and workers (people accepting tasks). Users on both sides are required to open an Amazon account in order to use the service. The requester can use a list of customizable pre-defined editable HTML templates in order to create Human Intelligence Tasks (HITS). The HITS can include one or more questions and are performed anonymously by one or more workers. The requester has the freedom to define the reward amount as low as $0.01.

Once a HIT has been posted online to the Mechanical Turk service, it is found by potential workers through an internal search engine using relevant keywords that were previously specified by the requester. A task can also include an optional qualification test, taken before the main task to ensure a minimum level of expected proficiency for anonymous workers. Finally, after the HITS have been submitted, the requesters have the option of accepting or rejecting the work done at their own discretion, and not paying for failed tasks.

3.1 Arabic Annotation Tasks

In order to evaluate Arabic annotation, we followed some general principles described by (Snow et al., 2008) in order to keep the tasks clear and succinct, so that they would be accessible for the non-expert workers that we were targeting. Moreover, we designed the format of task as a multiple choice test as shown in Figure 1 in the last page of this paper.

Annotating an Arabic corpus linguistically involves many tasks such as part-of-speech tagging, selecting the correct case endings and verb moods, and determining syntactic functions (Maamouri et al., 2008).

However, the majority of AMT workers are non-expert Arabic speakers. For the experiment presented in this paper, we chose reduced part-of-speech tagging (5 tags) and grammatical case endings (4 tags). Each task used a simplified tagset, which lead naturally to multiple choice questions for each word in the test corpus. It was hoped that restricting the tasks to a multiple-choice response model would improve the accuracy of results and make the tasks easier. The reward offered for both tasks was $0.01 per word, and the time allocated for each task was two hours.

3.1.1 Case ending tagging

In this task, annotators were asked to identify grammatical case endings (Habash et al., 2007; Maamouri et al., 2008). The valid responses offered were: nominative case, genitive case, accusative case or none. In Arabic, case is in general determined by the syntactic function of that particular word. For example, subjects are always in the nominative case, and objects are always in the accusative case.

The most common Arabic case endings can usually be recognized through the diacritic mark of the last letter in that Arabic word. Full accuracy of case endings can for some words be quite complex, even for expert Arabic linguists, requiring a full understanding of the Arabic sentence in order to put the word in question into context. An example of this is the sound masculine plural, where both the accusative and the genitive case have the same surface case ending (*-een*), and the correct grammatical case can only be determined through understanding the word's syntactic function. An example of this task is shown in Figure 2.

We were able to compare the results obtained in this task with the gold standard annotation made available in the Quranic corpus dataset. Even though it was clear that this task would require more detailed knowledge of Arabic grammar, we were curious to understand how reliable non-expert annotation could be. The first 100 words from chapter 23 of the Quran were selected for annotation of grammatical case endings.

Any Turkers wanting to participate in the case-ending task would first have to pass an Arabic screening test as shown in Figure 3, which requires a basic understanding of the Arabic language. After that, it was assumed that the Turkers were familiar with the four multiple-choice options (nominative case, genitive case, accusative case or none). These options were specified in both Arabic and English for each word in the test dataset.

3.1.2 Part-of-speech tagging

In this task, annotators were asked to choose a correct word-category (part-of-speech) for each Arabic word in the test corpus. For the part-of-speech tagging task, the same Quranic dataset was used: workers were asked to annotate data from the first 200 words from the chapter 23 of the Quran.

For the purpose of this experiment, we adapted a simplified set of Arabic part-of-speech tags limited to six choices: noun, verb, adjective, pronoun, particle and other/unknown. As with the case-ending task, a simple Arabic qualification test was presented at the start of the task in order to ensure quality and detect possible random annotation by non-native speakers of Arabic.

The annotators were asked to fill in missing personal pronouns in screening sentences from a list of five possible answers. This requires a basic level of understanding of Arabic grammar.

4. Results

The total number of interested Turkers in both experiments combined was 137. However, only 24 participants passed the qualification test and only 17 out of the 24 who qualified did some annotation work (70% of the qualified Turkers). The remaining 30% of Turkers did a very limited number of HITS (between 1 and 6 out of 100 HITS), therefore we decided to exclude their results from the reports since it could bias the report. The low participation number for both tasks can be explained by the following three factors :

1. Limited timeframe (the whole experiments was carried over a period of 1 month).
2. The low pay rate per HIT (only 0.01$).
3. The difficulty of the task (requires a certain level of Arabic).

The overall results were 50.07% accuracy for grammatical case endings (by 7 annotators) and 63.91% for POS tagging (by 10 annotators).

In the crowdsourcing experiment, it is possible to control the number of words annotated by each contributor. We did this for the case-ending task, but we left the POS-tagging task open to better understand the number of words per annotator. The results are summarized in tables 1 and 2.

Contributor	Correct HITS	Total HITS completed	Accuracy
Annotator 1	42	100 / 100	42%
Annotator 2	29	100 / 100	29%
Annotator 3	55	100 / 100	55%
Annotator 4	58	100 / 100	58%
Annotator 5	54	90 / 100	60%
Annotator 6	48	88 / 100	54.5%
Annotator 7	43	79 / 100	54.43%
Total	**329**	**657 / 700**	**50.07%**

Table 1: Accuracy for grammatical case tagging

The annotator numbers presented in the result tables are local to each task. Since the results of the crowdsourcing experiment are anonymous, it is possible that different annotators were involved in the two different tagging tasks. It's worth mentioning that only four (23%) out of the 17 Turkers finished 100% of the HITS assigned to them. Moreover, it appears that those who participated in the grammatical case ending task did complete most of their HITS (93.83%) while those who did the part-of-speech tagging tasks completed only 34.5% of the HITS.

Contributor	Correct HITS	Total HITS completed	Accuracy
Annotator 1	116	196 / 200	59.2%
Annotator 2	74	112 / 200	66.1%
Annotator 3	79	99 / 200	79.8%
Annotator 4	66	99 / 200	66.7%
Annotator 5	18	43 / 200	41.9%
Annotator 6	19	37 / 200	51.4%
Annotator 7	22	32 / 200	68.75%
Annotator 8	19	29 / 200	65.51%
Annotator 9	15	25 / 200	60%
Annotator 10	13	18 / 200	72.22%
Total	**441**	**690 / 2000**	**63.91%**

Table 2: Annotator accuracy for part-of-speech tagging.

5. Conclusion

A key question is whether there are enough Arabic language expert Turkers who can perform this task sufficiently well, compared to automatic annotation. From the results it would appear that: (a) annotating Arabic grammatical case is harder than POS tagging, and (b) for the two tasks, crowdsourcing for Arabic annotation is not as effective as other methods.

We believe that the Turkers did understand the two tasks, given that a screening test was performed. However, based on the low results for case-endings (50.07%) it would appear that the task was understood, but not easily carried out by the Turkers, who did not have a sufficient deep understanding of Arabic grammar. We can conclude that the task was understood based on the fact the completely random selection of the four cases would have resulted in a far lower score of around 25%.

A possible improvement for a future repeat experiment might be to give a harder initial screening test to ensure that annotators performing the task have a higher level of Arabic linguistic knowledge. By comparing this experiment to the methodology used to annotate the Quranic Arabic Corpus, results could have potentially been improved by supplying more detailed annotation guidelines and examples to further explain the tag-sets for the two tasks.

We can conclude that based on these results, using crowdsourcing for Arabic annotation is not straightforward, and there may be better techniques as suggested by Adda et al. (2011). An alternative approach could involve automatic tagging instead of using non-Arabic linguists to perform manual tagging of Arabic using a system such as Amazon Mechanical Turk. Automatic Arabic morphological taggers have a higher accuracy for case selection and part-of-speech tagging, with some systems reporting over 80-90% accuracy (Habash and Rambow, 2005).

It is also interesting to note that as hypothesized before the start of the experiment, Arabic annotation tasks such as POS-tagging and grammatical case-ending require linguistic expertise (Alkuhlani, Habash and Roth, 2013; Habash et al., 2007). The case-ending experiment also demonstrates that strong linguistic knowledge is required for Arabic case classification – a task sometimes found difficult even by well-read native speakers of Arabic.

6. References

Adda, G., Sagot, B., Fort, K., Mariani, J. (2011). Crowdsourcing for Language Resource Development: Critical Analysis of Amazon Mechanical Turk Overpowering Use. In *Proceedings of the Language and Technology Conference*, Poznan, Poland.

Alkuhlani, S., Habash, N. & Roth, R. (2013). Automatic Morphological Enrichment of a Morphologically Underspecified Treebank. In *Proceedings of Conference of the North American Association for Computational Linguistics (NAACL)*, Atlanta, Georgia.

Bloodgood, M., & Callison-Burch, C. (2010). Using mechanical turk to build machine translation evaluation sets. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*. Los Angeles, Ca.

Callison-Burch, C., & Dredze, M. (2010). Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon'smechanical Turk*. Los Angeles, Ca.

Callison-Burch, C. (2009). Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP 2009)*. Singapore, Singapore.

Chamberlain, J.; Poesio, M. & Kruschwitz. (2009). A new life for a dead parrot: Incentive structures in the Phrase Detectives game. In *Proceedings of the Webcentives Workshop at WWW'09*. Madrid, Spain.

Dukes, K. (2013). Statistical Parsing by Machine Learning from a Classical Arabic Treebank. PhD Thesis, University of Leeds.

Dukes, K. & Buckwalter, T. (2010). A Dependency Treebank of the Quran using Traditional Arabic Grammar. In *Proceedings of the Seventh International conference on Informatics and Systems*. Cairo, Egypt.

Dukes, K. & Habash, N. (2010). Morphological Annotation of Quranic Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. Valletta, Malta.

Dukes, K., Atwell, E. & Sharaf, A-B. (2010). Syntactic Annotation Guidelines for the Quranic Arabic Treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. Valletta, Malta.

Habash, N., Gabbard, R., Rambow, O., Kulick, S. & Marcus, M. (2007). Determining Case in Arabic: Learning Complex Linguistic Behavior Requires Complex Linguistic Features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),* Prague, Czech Republic.

Habash, N. & Rambow, O. (2005). Arabic Tokenization, Morphological Analysis, and Part-of-Speech Tagging in One Fell Swoop. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, MI.

Maamouri, M., Bies, A., Kulick, S. (2008). Enhanced Annotation and Parsing of the Arabic Treebank. In *Proceedings of the INFOS International Conference*, Cairo, Egypt.

Maamouri, M., Kulick, S. & Bies, A. (2008). Diacritic Annotation in the Arabic Treebank and Its Impact on Parser Evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC),* Marrakech, Morocco.

Ryding, K-C. (2008). *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press.

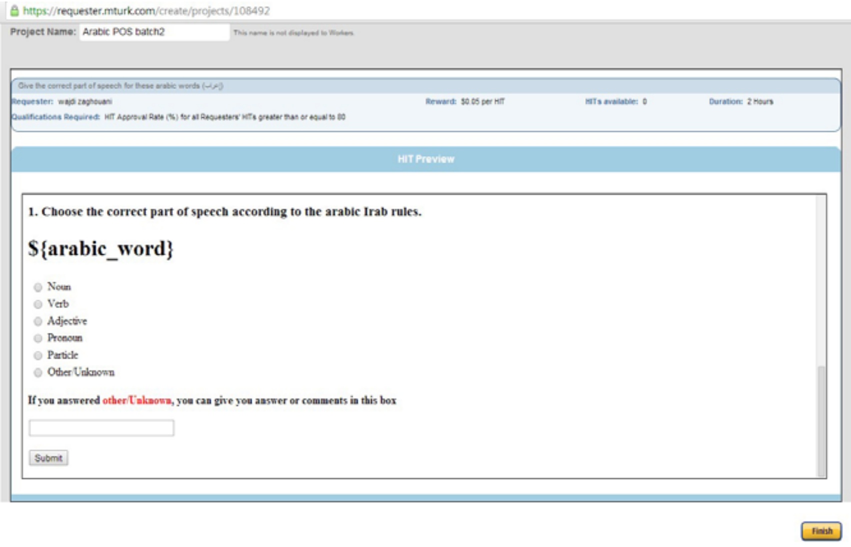
Siddiqui, A-R. (2008). *Quranic Keywords: A Reference Guide*. The Islamic Foundation.

Smrž, O. & Hajič, J. (2006). The Other Arabic Treebank: Prague Dependencies and Functions. In *Arabic Computational Linguistics: Current Implementations*, CSLI Publications.

Snow, R., O'Connor, B., Jurafsky, D. and.Ng, A-Y. (2008). Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP,* Waikiki, Honolulu, Hawaii.

Su, Q., Pavlov, D., Chow, J-H & Baker, W-C. (2007). Internet-Scale Collection of Human-Reviewed Data. In *Proceedings of the 16th international conference on World Wide Web*. Banff, Alberta, Canada.

Wang, A., Hoang, C.D-V., & Kan, M-Y. (2010). Perspectives on Crowdsourcing Annotations for *Natural Language Processing. Language Resources and Evaluation*,47:1.


https://requester.mturk.com/create/projects/108492

Project Name: Arabic POS batch2 This name is not displayed to Workers.

Give the correct part of speech for these arabic words (إعراب)

Requester: wajdi zaghouani Reward: $0.05 per HIT HITs available: 0 Duration: 2 Hours

Qualifications Required: HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 80

HIT Preview

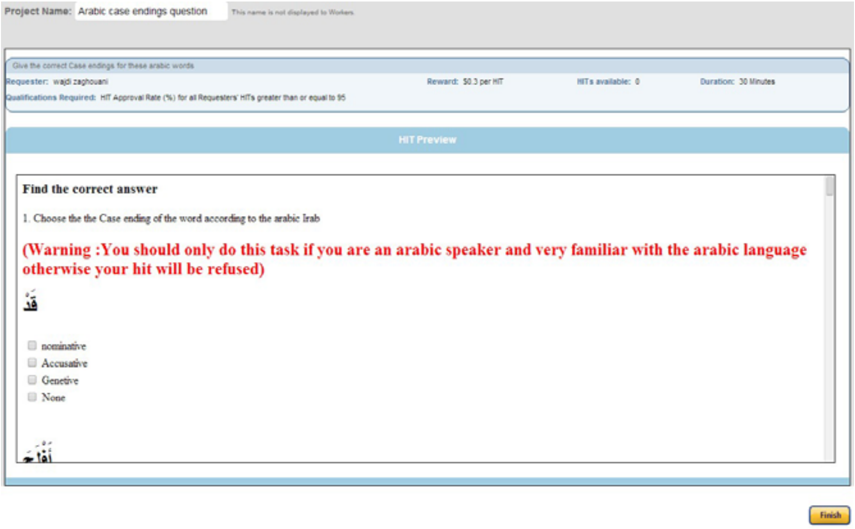
1. Choose the correct part of speech according to the arabic Irab rules.

${arabic_word}

○ Noun
○ Verb
○ Adjective
○ Pronoun
○ Particle
○ Other/Unknown

If you answered other/Unknown, you can give you answer or comments in this box

Figure 1: Task creation sample.


Project Name: Arabic case endings question This name is not displayed to Workers.

Give the correct Case endings for these arabic words

Requester: wajdi zaghouani Reward: $0.3 per HIT HITs available: 0 Duration: 30 Minutes

Qualifications Required: HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 95

HIT Preview

Find the correct answer

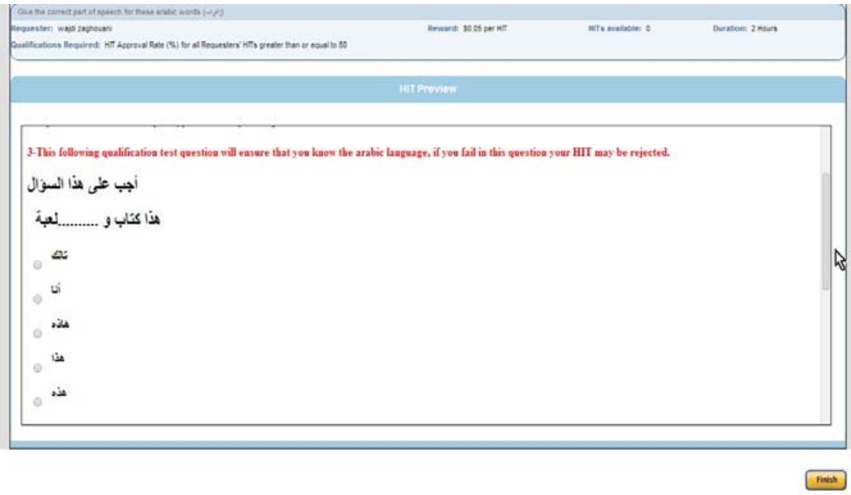
1. Choose the the Case ending of the word according to the arabic Irab

(Warning :You should only do this task if you are an arabic speaker and very familiar with the arabic language otherwise your hit will be refused)

قَدْ

□ nominative
□ Accusative
□ Genetive
□ None

Figure 2: A sample case ending question.


Give the correct part of speech for these arabic words (إعراب)

Requester: wajdi zaghouani Reward: $0.05 per HIT HITs available: 0 Duration: 2 Hours

Qualifications Required: HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 80

HIT Preview

3-This following qualification test question will ensure that you know the arabic language, if you fail in this question your HIT may be rejected.

أجب على هذا السؤال

هذا كتاب و ....لعبة

○ تلك
○ أنا
○ هاذه
○ هذا
○ هذه

Figure 3: A sample screening test question.