Leveraging Crowdsourcing for the Thematic Annotation of the Qur'an

Amna Basharat, I. Budak Arpinar, Khaled Rasheed
Dept. of Computer Science, University of Georgia, Athens, GA, 30605 USA
amnabash,budak,khaled@uga.edu

ABSTRACT

In this paper, we illustrate how we leverage crowdsourcing to create workflows for knowledge engineering in specialized and knowledge intensive domains. We undertake the special case of the Arabic script of the Qur'an, a widely studied manuscript, and attempt to employ crowdsourcing methods for its thematic annotation at the sub-verse level, for which, there is no standardized knowledge model available to date. We demonstrate that our proposed method presents feasibility to achieve reliable annotations in an efficient and scalable manner. The proposed methodology and framework is meant to be generalizable to other knowledge intensive and specialized domains.

Categories and Subject Descriptors

H.4 [**Information Systems**]: Crowdsourcing

Keywords

thematic annotation; disambiguation; crowdsourcing; Qur'an; knowledge engineering; ontology; semantic web

1. INTRODUCTION

Efforts towards semantic annotation and knowledge engineering in specialized and knowledge intensive domains continue to present challenges to the semantic web researchers. The Qur'an is one of the most widely read and studied books. Its original script is in the Arabic language, which is rich in both its morphology and semantics.

As part of this research, we undertook the task of thematic disambiguation and annotation (as part of formal and standardized knowledge modelling and ontology engineering activities) of the Qur'anic verses by augmenting the traditional information extraction and text mining techniques with crowdsourcing methods. The need for crowdsourcing stems from the fact that for the Qur'anic knowledge, a high level of accuracy and reliability is desired, given the sensitivity of the knowledge at hand. Pure computational approaches fail to meet this standard and the contribution from human experts is indispensable. However, finding experts is greatly time consuming and this makes the process of obtaining semantic annotations non-scalable. Motivated by the success of human-computation and crowdsourcing methods, we therefore attempt to investigate the usefulness of these approaches for knowledge intensive and specialized domains, such as the one encompassed by the Qur'an.

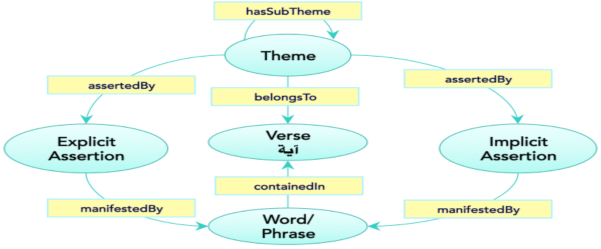

Figure 1: A segment of the thematic annotation knowledge model populated via the crowd

2. PROBLEM DEFINITION: THEMATIC ANNOTATION IN THE QUR'AN

As part of this paper, we aimed to augment the available thematic hierarchies of the Qur'an to include annotations for the various themes at sub-verse level, a level deeper than what existing thematic hierarchies provide. The motivation for this comes from the diversity of thematic coverage that Qur'anic verses provide at an individual and collective level. The Qur'anic verses span different lengths; while some verses may be as short as a word or few words, others may span half a page or an entire page of a standard sized book. This inspires the need for increasing the level of granularity at which the *thematic assertions* for each verse are classified. We take our initial hierarchy from QuranyTopics datasource[1], which contains a hierarchy of themes, hand-crafted from a classical source. This is one of the only data sources, that provides an authentic, concept driven, thematic classification of the Qur'anic verses. However, the thematic classification in this resource is not only limited only to the verse level, its coverage of the concepts is not exhaustive.

As part of this research, not only do we propose sub-verse level annotation, we also make the distinction between *explicit* and *implicit assertions* of a *theme* as shown by a seg-

[1] http://quranytopics.appspot.com

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889409.

ment of the ontology schema designed for obtaining thematic annotations in Figure 1. Explicit assertions are amenable to be obtained through text mining and NLP techniques, such as, the direct occurrence of the word or a phrase that directly indicates the *manifestation* of a particular theme. However, even when a word or a phrase explicitly appears in a verse, in the Arabic language, it may manifest different meanings in different contextual settings. When modelling such thematic assertions, at sub-verse level, disambiguation by a human expert becomes indispensable. To add to the challenge, themes may also exhibit implicitly; whereby, the same theme may be manifested using not just a mere difference of word morphology, rather, a difference in expression or rhetoric. It is extremely difficult to extract such implicit thematic assertions via automated techniques, therefore, it becomes imperative that human contribution be sought.

3. APPROACH: CROWDSOURCING WORKFLOW

We devised the generic workflow that a typical crowdsourcing driven method for obtaining semantic annotations will entail. This is shown in Figure 2. There are several key stages and components. *Ontology Design:* An ontology schema such as the one given in Figure 1 guides the semantic annotation process. This serves as input to the *Task Generation and Design* stage, which creates an annotation or disambiguation relevant task to be crowdsourced based on the nature of the entity, relation or both, as specified in the *task specification*. The relevant task input is generated by retrieving relevant candidate verses from the available data sources such as the Semantic Qur'an [2] dataset or the quranontology[2]. The tasks are published on the Amazon Mechanical Turk (AMT)[3] platform. A complete workflow management system is implemented (a derivative of a workflow model for Linked Data Management presented in [1]), which includes means for generating dynamic tasks from a range of task profiles. The task generation module creates the required input, question and parameter files needed by the AMT API for publishing the task. The AMT crowd performs the *disambiguation and annotation tasks*. Both tasks are based on the Arabic script of the Qur'an, therefore, requiring the crowd workers to be familiar with the Arabic language of the Qur'an. For the disambiguation task, a question is presented to the crowd, which includes a verse, along with a highlighted, candidate explicit assertion for the given theme, and the crowd responds by declaring this assertion as either a positive or negative by determining if the occurrence is a true occurrence of the given theme. The annotation tasks require deeper knowledge and understanding of the Arabic text. The crowd determines whether the given verse contains any implicit reference to the given theme. If their response is positive, then they are also required to provide the portion of the verse (a meaningful phrase or a word) that implies the presence of the theme. Each task is required to have atleast five responses for it to be marked as complete. As a form of a quality measure, the crowd is also required to provide a confidence level (ranging from Very High to Very low), to indicate their confidence in their response. There is a *response collection and aggregation* module that collects and aggregates the responses based on statistical measures of aggregation. Weighted confidence measures and thresholds are applied. Based on this aggregation, the completed tasks are marked as either *Approved* or *Reviewable*. A high confidence and aggregation threshold is applied for the approved tasks. An *expert review and validation* is conducted as a followup for the reviewable tasks that have a degree of disagreement more than the baseline threshold. The approved and validated annotations are passed on for *Ontology Population* and linked with existing data sources.

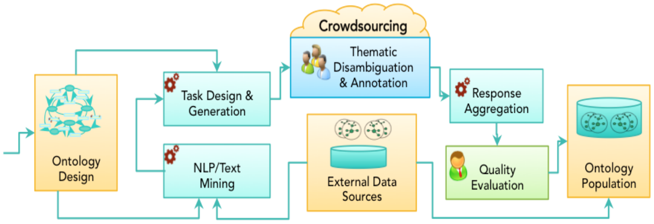

Figure 2: Generic workflow for Crowdsourcing semantic annotations for ontology population

4. RESULTS AND DISCUSSION

For the initial pilot study, we obtained 12,000 (appx) submissions for the two types of tasks, for 70 distinct themes for each type. Of these, 1300 key verse-theme pairs were disambiguated, with explicit assertions, while 2200 pairs were disambiguated/annotated with implicit ones. Of the completed submissions, 96% tasks were approved for explicit assertions, while, for the implicit ones, 81% were approved, without applying any expert validation. While the results are promising, a considerable number of tasks remained incomplete for the implicit annotations. While these were not included in the results compilation, this indicated lower crowd engagement, which may have been due to that fact that the task statements were presented only in the Arabic language or since only the AMT sandbox was employed for the pilot study.

The results of the ongoing study suggest that the crowd can significantly assist in scaling the knowledge engineering activities such as knolwedge formalization, and semantic annotation with reasonable reliability. Our ongoing efforts are focused towards augmenting this workflow to obtain expert validations for the reviewable tasks (through our own custom web framework) so that high quality annotations may be obtained. In future, we plan to increase the size of the study, and experiment with a range of other task designs of varying complexity. The study clearly shows the potential benefit of crowdsourcing that can be harnessed by knowledge intensive and expertise driven domains.

5. REFERENCES

[1] A. Basharat, I. B. Arpinar, S. Dastgheib, U. Kursuncu, K. Kochut, and E. Dogdu. Semantically enriched task and workflow automation in crowdsourcing for linked data management. *International Journal of Semantic Computing*, 8(04):415–439, 2014.

[2] M. A. Sherif and A.-C. N. Ngomo. Semantic Quran - a multilingual resource for natural-language processing. *Semantic Web*, 6(4):339–345, 2015.

[2] http://quranontology.com

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