An Analysis of Crowdsourcing Relevance Assessments in Spanish

Omar Alonso¹ and Ricardo Baeza-Yates²

¹ Microsoft Corp.
1065 La Avenida, Mountain View, CA 94043
omar.alonso@microsoft.com

² Yahoo! Research Barcelona
Avinguda Diagonal 177, 8th Floor 08018 Barcelona
rbaeza@acm.org

Abstract. Recently Amazon Mechanical Turk has emerged as a viable platform for conducting relevance assessments. It is possible to conduct experiments extremely fast, with good results and at a very low cost. However, most of the previous work on crowdsourcing has been done in English. In this paper we present the results of a series of experiments using the Spanish part of CLEF, demonstrating that crowdsourcing platforms do work for other languages than English. Our findings reinforce the wisdom of the crowd for this kind of task and we show our methodology at work with several examples.

1 Introduction

In the world of the Web 2.0 and user generated content, one important sub-class of peer collaborative production is the phenomenon known as crowdsourcing [6, 9]. In crowdsourcing potentially large jobs are broken into many small tasks that are then outsourced directly to individual workers via public solicitation. One of the best examples is Wikipedia, where each entry or part of an entry could be considered as a task being solicited. However, successful examples of volunteer crowdsourcing are difficult to replicate and hence a financial compensation is used, usually as micro-payments of the order of a few cents per task. This is the model of the Amazon Mechanical Turk (AMT), where many tasks can be done quickly and cheaply.

Later we detail more AMT, but it is quite simple. Workers choose from a list of jobs being offered, where the reward being offered per task and the number of tasks available for that request are indicated. Workers can click on a link to view a brief description or a preview of each task. After seeing the preview, workers can choose to accept the task, where optionally, a qualification exam must be passed to assign officially the task to them. Tasks are very diverse in size and nature, requiring from seconds to minutes to complete. On the other hand, the typical compensation ranges from one cent to less than a dollar per task and is usually correlated to the task complexity. In the case of information retrieval, crowdsourcing is perfect for relevance assessments as tasks are already small, so...
they do not need to be divided in smaller tasks. For this reason crowdsourcing has been used successfully for relevance assessments. For example, Alonso & Mizzaro [1] compared a single topic to TREC and found that workers were as good as the original assessors, and in some cases they were able to detect errors in the golden set. A similar work by Alonso et al. [2] in the context of INEX with a larger data set shows similar results. Kazai et al. [7] propose the collective relevance assessment method for gathering relevance assessments for collections of digital books and videos.

Nevertheless, all the mentioned research have been done in English, as the restrictions of being a registered user in AMT imply that most turkers are American. So a natural question is: can we use crowdsourcing for other languages? According to the last USA census in 2006, Spanish is the second most spoken language. In fact, in 2006 there were 34 million people in the USA for which Spanish was their main language, that is, a 12% of the total population (if we consider Spanish as a second language the number grows to over 40 million). So the natural choice to answer the question above is Spanish to see if there are enough assessors that know a language different from English.

This question is also important considering that most linguistic or retrieval oriented golden sets are in English. Hence, would be very interesting if crowdsourcing could help in generating more resources in other languages. To evaluate the quality of crowdsourcing in Spanish we use one of the few golden sets across languages: CLEF (Cross Language Evaluation Forum), using just the Spanish sub-collection. As a side product, our experiments are also one example of how crowdsourcing should be used.

We found that it is possible to run crowdsourcing experiments in other languages than English on Mechanical Turk. Our largest experiment shows that CLEF experts and workers agree on 70% of the answers, which is very promising. We also found that the quality of the justifications is very high and it can be used as a mechanism to detect workers who are fluent in a particular language. In certain cases, workers were able to identify errors in the golden set and by looking at the justifications we can easily identify the disagreements.

This paper is organized as follows. First, we describe Amazon Mechanical Turk in Section 2. Second we describe our experiments and their results in Section 3. We end with some final remarks in Section 4.

2 Amazon Mechanical Turk

Crowdsourcing has emerged as a feasible alternative for relevance evaluation because it combines the flexibility of the editorial approach at a larger scale. Crowdsourcing is a term used to describe tasks that are outsourced to a large group of people instead of performed by an employee or contractor.

AMT is an example of a crowdsourcing platform\(^3\). AMT is an Internet service that gives developers the ability to include human intelligence as a core

\(^3\) www.mturk.com
component of their applications. Developers use a web services API to submit
tasks, approve completed tasks, and incorporate the answers into their software
applications. To the application, the transaction looks very much like any re-
room procedure call. The application sends the request, and the service returns
the results. People come to the web site looking for tasks and receive payment
for their completed work. In addition to the API, there is also the option to
interact using a dashboard that includes several useful features for prototyping
experiments.

The individual or organization who has work to be performed is known as
the requester. A person who wants to sign up to perform work is described
in the system as a worker. The unit of work to be performed is called a HIT
(Human Intelligence Task). Each HIT has an associated payment and an allotted
completion time; workers can see sample hits, along with the payment and time
information, before choosing whether to work on them. It is possible to control
the quality of the work by using qualification tests. A qualification test is a set
of questions (like a HIT) that the worker must answer to qualify and therefore
work on the assignments.

One of the most important aspects of performing evaluations using crowd-
sourcing is to design the experiment carefully. Asking the right questions in a
simple and effective way involves using guidelines for survey and questionnaire
design.

3 Experiments

We conducted three main experiments to test our approach (see table 1). Each
experiment has specific metadata in Spanish and English so it can be searchable
in both languages. The actual HIT consisted on using the same topic, description,
and narrative from the original CLEF distribution. Note that we discourage
people who are not fluent in Spanish to take the test. We also pay a bonus in
experiment three to get good quality justifications. We followed previous design
guidelines for running experiments as well as dealing with worker quality [3]. All
experiments shared the same metadata (Relevancia, Evaluación, web, Maquinas
de Búsqueda) plus specific keywords (Pena de muerte, Japón, Naciones Unidas,
Los Juegos Olímpicos).

<table>
<thead>
<tr>
<th>ID</th>
<th>Type</th>
<th># of documents</th>
<th>Collection</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>Graded (3 point scale)</td>
<td>22</td>
<td>Wikipedia</td>
<td>145, 150, 152, 153, 157</td>
</tr>
<tr>
<td>E2</td>
<td>Binary</td>
<td>18</td>
<td>CLEF</td>
<td>145, 150, 152, 153, 157</td>
</tr>
<tr>
<td>E3</td>
<td>Binary</td>
<td>135 (124 + 11 honey pots)</td>
<td>CLEF</td>
<td>145, 150, 152, 153, 157</td>
</tr>
</tbody>
</table>

Table 1. Overview of the 3 experiments.
3.1 CLEF

CLEF is an initiative to evaluate cross-language information retrieval that started in 2000 [5]. The CLEF multilingual document collection contains several topics and for each language a different document database. In our case we work with the Spanish subset of CLEF, which is composed by over 450 thousand news documents of 1994 and 1995 of EFE, a Spanish news agency. The number of topics that we used in our experiments was 5 and for relevance assessments we just need the relevant documents for each topic plus a sample of non-relevant documents.

3.2 Enough People Know Spanish?

Most of the experimentation in Mechanical Turk is done in English. The first question is then, do we have enough workers who know Spanish? To answer that question, we designed a simple experiment (E1) with Wikipedia content given a short topic description. We also misclassified a few pages to make sure that we were able to trap potential spammers. We did not rejected any work at this moment. The goal was to see if there was interest in such experiments.

The results for E1 are as follows. There were 5 unique users for the 69 assignments. Of the total, 23/69 (33%) contain a justification. There were three workers per assignment and the total cost for the experiment was $1.72. Because we created the experiment, we assigned the relevance values and then compared to the average worker. There was a 60% agreement. This short experiment showed us that it was possible to run similar tasks but with CLEF content.

3.3 Quality of Assessments

The second experiment (E2) used a subset of the Spanish CLEF and the results are as follows. There were 5 unique users for 59 assignments. We did perform spam detection but did not rejected any work. The total cost for the experiment was $1.99. There was a very high agreement (90%) but only 3/59 (5%) of the assignments contained a justification.

For the rest of CLEF, the third experiment (E3), we used honey pots and spam detection. To increase the number of good quality feedback, we paid a $0.01 bonus if we found that the justification was useful. In this case there were 20 unique workers for 273 assignments. The total cost for the experiment was $14.28. The agreement was strong (70%) and 241/273 (88%) of the assignments contained a justification. We did reject 245 assignments that we considered were not good answers by workers. Figure 1 shows the worker activity for E3.

By honey pot we mean the inclusion of a document and relevance assessment that we know the answer upfront and we expect the worker to know it as well. If a worker misses the answer, then there is a possibility that the person was gaming the system or not paying attention. The idea is to include a number of honey pots in the experiment and use that information to detect potential spammers. In our case, if a worker misses more than 2 easy honey pots, then is a
candidate for exclusion from the results (and also the worker doesn’t get paid). We also examine the speed of a worker for answering the question. If the time to complete an assignment looks similar across all completed work, then looks like a potential spammer.

We also found errors in the golden set. There were five documents where the majority of the workers (at least two out of three) disagree with the experts on the relevance assessment. Table 4 shows the document identifiers and examples of the justifications by the workers.

3.4 Justification Analysis

One way to assess the knowledge of Spanish is by analyzing the justifications. As we can see in Table 2 the explanations are reasonable well written for the first experiment. Notice that some people used uppercase letters, perhaps to avoid the need to use accents, although even in this case some people included the accents. On the other hand accents could be an issue if the assessor uses an English keyboard. It is worth mentioning that in some cases workers did assess a document and provided the justification in English as we can see on the last entry of Table 4. It is unclear if this particular example was done using a translation tool, but at least the worker was following instructions.

Interestingly enough, the length of the justifications increased in the second experiment (the average length went from 55.73 to 201 characters). However, the overall number of justifications in E2 decreased. After introducing the bonus
payment, the number of justifications went up (compared to E1 and E2). In terms of the length of the explanations, the third experiment was 139.34 characters on average. This is remarkable considering the small value of the bonus ($0.01).

<table>
<thead>
<tr>
<th>Query</th>
<th>Answer</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shoemaker-Levy y Júpiter</td>
<td>Yes</td>
<td>Se refiere al cometa precitado y a su impacto en el planeta Júpiter, y sus consecuencias.</td>
</tr>
<tr>
<td>java</td>
<td>Somewhat</td>
<td>Habla sobre la versión de código abierto de dicho programa.</td>
</tr>
<tr>
<td>Ib barcelona</td>
<td>No</td>
<td>Trata acerca de la ciudad, mas no del club de fútbol.</td>
</tr>
</tbody>
</table>

Table 2. Sample of justification answers by workers on the first experiment (E1).

<table>
<thead>
<tr>
<th>Query</th>
<th>Answer</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>los derechos de la infancia</td>
<td>No</td>
<td>No es relevante porque mas bien es un articulo informativo sobre una reunion y la agenda que se va a revisar en dicha reunion y no hay datos referidos a la Convención de las Naciones Unidas sobre los derechos de la infancia.</td>
</tr>
<tr>
<td>AI contra la pena de muerte</td>
<td>Yes</td>
<td>Se lee claramente que AI esta en contra de la pena de muerte y que ha tomado acciones para que el gobierno jordano tome consciencia en cuanto a la sentencia y ejecución de la misma.</td>
</tr>
<tr>
<td>Importaciones de arroz en Japon</td>
<td>No</td>
<td>Habla de los recursos que el gobierno invertira para sacar al pais de la recesion y solo hace mención breve sobre los agricultores de arroz que han sido afectados por la importación de arroz.</td>
</tr>
</tbody>
</table>

Table 3. Sample of justification answers by workers on the second experiment (E2).

### 4 Summary and Conclusions

The results of our research are promising. We cannot tell for sure if all workers are Spanish-speaking or if they were using a translation tool and based their decisions on that. That said, the justification data is a very useful tool to identify workers who are fluent in that language. In addition, the overall quality of the justification/feedback is extremely high, and most probably they are fluent in Spanish, as justifications would have had lesser quality if a translation software was being used.

The initial results show that it is possible to assess multilingual content using a crowdsourcing platform, at least in the English-Spanish case. This is consistent with previous research and we are showing that this can be extended to non-English relevance tasks. Not only the results show an agreement with the CLEF
<table>
<thead>
<tr>
<th>Query</th>
<th>Answer</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Los juegos olímpicos y la paz</td>
<td>Yes</td>
<td>Yo creo que es desde ya muy relevante, ya que explica cómo cambiando una fecha de un partido se pudo mantener la paz y evitar conflictos. La fecha se cambió para que Alemania no jugara el día del nacimiento de Hitler.</td>
</tr>
<tr>
<td>Importaciones de arroz en Japón</td>
<td>No</td>
<td>El documento no contiene nada referido al tema que es importaciones de arroz en Japón, todo el documento está referido a la Copa del Rey de España, es decir trata sobre resultados futbolísticos.</td>
</tr>
<tr>
<td>Ganadoras de Wimbledon</td>
<td>No</td>
<td>El documento no es relevante al tema que son las ganadoras del torneo de tenis en Wimbledon, por el contrario el documento trata sobre la final que jugaron en el Roland Garros la alemana Steffi Graf y a la española Arantxa Sánchez Vicario.</td>
</tr>
</tbody>
</table>

Table 4. Sample of justification answers by workers on the third experiment (E3).

<table>
<thead>
<tr>
<th>Topic</th>
<th>DocId</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>145</td>
<td>EFE19940222-12530</td>
<td>No es relevante ya que se rehíere a un libro sobre economía y la economía entre EEUU y Japón y no se trata explícitamente el tema sobre las importaciones de arroz en Japón.</td>
</tr>
<tr>
<td>145</td>
<td>EFE19940315-08328</td>
<td>No, sector del petróleo en este caso.</td>
</tr>
<tr>
<td>145</td>
<td>EFE19940322-14553</td>
<td>Este documento no es relevante debido a que no menciona información sobre las importaciones de arroz en Japón en ninguno de sus párrafos. Es un documento que habla sobre el acero.</td>
</tr>
<tr>
<td>152</td>
<td>EFE19940202-01089</td>
<td>Documento no relevante por cuanto no proporciona información sobre la convención de la ONU acerca de los derechos de la infancia. El documento trata sobre los nuevos derechos de los niños en Francia.</td>
</tr>
</tbody>
</table>

Table 5. Examples of disagreements between experts (CLEF) and workers.
experts, but also show examples when there is a disagreement. This finding was only possible because of the high quality of the feedback.

Compared to previous experience running English relevance experiments, these tests were slower. For a similar experiment in English that usually takes 24hrs to complete, the Spanish version finalizes in 5 days. One potential reason is that there are less Spanish workers than in English. Another reason is the current limitation on Euro payments from Amazon that may impact potential workers. Based on this, the percentage of Spanish-speaking turkers is less than 20% and more careful experiments may show that the actual percentage is consistent with the real percentage (today could be larger than 15%).

5 Acknowledgments

We thank Cristina González for helping us with the CLEF data set.

References