Data Collection and Annotation for State-of-the-Art NER Using Unmanaged Crowds

Spencer Rothwell, Steele Carter, Ahmad Elshenawy, Vladislavs Dovgalecs, Safiyyah Saleem, Daniela Braga, Bob Kennewick

VoiceBox Technologies, Bellevue, WA, USA

spencerr, steelec, ahmade, vladislavsd, danielab, safiyyahs, bobk [@voicebox.com]

Abstract

This paper presents strategies for generating entity level annotated text utterances using unmanaged crowds. These utterances are then used to build state-of-the-art Named Entity Recognition (NER) models, a required component to build dialogue systems. First, a wide variety of raw utterances are collected through a variant elicitation task. We ensure that these utterances are relevant by feeding them back to the crowd for a domain validation task. We also flag utterances with potential spelling errors and verify these errors with the crowd before discarding them. These strategies, combined with a periodic CAPTCHA to prevent automated responses, allow us to collect high quality text utterances despite the inability to use the traditional gold test question approach for spam filtering. These utterances are then tagged with appropriate NER labels using unmanaged crowds. The crowd annotation was 23% more accurate and 29% more consistent than in-house annotation.

Index Terms: unmanaged crowdsourcing, data collection, spam, NER, Natural Language Understanding (NLU), dialogue systems.

1. Introduction

Building NLU models requires a large amount of text utterances. In order to collect extensive quantities of annotations in a cost-effective manner and with fast turnaround, we leveraged crowdsourcing, specifically using unmanaged crowds. Crowds can generate creative input for open ended questions, which then can be used as bootstrapping data for NLU models. It is difficult, however, to prevent spam when collecting open text. We distinguish spam from low quality data, the former being intentionally produced (e.g. an utterance about sports in a given location). The latter being less obvious and not necessarily produced with potential spelling errors and verify these errors with the crowd before discarding them. These strategies, combined with a periodic CAPTCHA to prevent automated responses, allow us to collect high quality text utterances despite the inability to use the traditional gold test question approach for spam filtering. These utterances are then tagged with appropriate NER labels using unmanaged crowds. The crowd annotation was 23% more accurate and 29% more consistent than in-house annotation.

Building NLU models requires a large amount of text utterances. In order to collect extensive quantities of annotations in a cost-effective manner and with fast turnaround, we leveraged crowdsourcing, specifically using unmanaged crowds. Crowds can generate creative input for open ended questions, which then can be used as bootstrapping data for NLU models. It is difficult, however, to prevent spam when collecting open text. We distinguish spam from low quality data, the former being intentionally produced (e.g. an utterance about sports in a given location). The latter being less obvious and not necessarily produced with potential spelling errors and verify these errors with the crowd before discarding them. These strategies, combined with a periodic CAPTCHA to prevent automated responses, allow us to collect high quality text utterances despite the inability to use the traditional gold test question approach for spam filtering. These utterances are then tagged with appropriate NER labels using unmanaged crowds. The crowd annotation was 23% more accurate and 29% more consistent than in-house annotation.

Sequence labeling tasks, such as Part-of-Speech (POS) tagging or entity labeling for NER, are increasingly common applications of crowdsourcing used in Natural Language Understanding (NLU) tasks. Carmel et al. [1] use crowdsourcing as a means of evaluating the output of various entity recognizers with some success. Others have utilized crowdsourcing for more than just evaluating the performance of an entity recognizer, relying on crowd judgments to identify as well as categorize entity spans in text data. Many papers have been published on crowdsourced NER on Twitter data ([2], [3], [4]), but the approach has also been applied to email data [5].

There are typically two steps in crowdsourcing data for NER model training. First, identify the spans of the entities in a source text. Second, classify any identified spans as belonging to some entity type. Voyer et al. [6] rely on a hybrid method of using expert annotators to identify entity spans and crowdsourcing to classify the entities. Braunschweig et al. [7] use crowdsourcing for each step, with mixed results. They found that using a pure crowdsourcing approach for entity span identification and labeling did not perform as well as expected. This result may be attributable to the lack of any quality control mechanisms whatsoever due to issues with their crowdsourcing platforms. The negative impact of spam on crowdsourced data quality in the absence of proper quality control has been well established [8], [9].

Finin et al. [2] and Bontcheva et al. [3] combine the boundary detection and entity labeling steps together, presenting workers with an entity-containing phrase and allowing them to selectively identify and classify any possible entities in context, word-by-word. Finin et al. [2] present each word in a phrase with a radial selection menu for each entity type being considered. Bontcheva et al. [3] present the workers with a phrase and ask them to highlight which words belong to a given entity type (e.g. Location). The results of [2] suggest...
that the radial selection menu format makes it difficult for workers to reliably understand entity span boundaries.

Numerous studies ([2], [3], [6], [7]) have leveraged crowdsourcing for NER with some measure of success, finding that crowdsourced annotations perform as well as or better than expert annotations if implemented correctly. Our work outlined in this paper aims to improve upon existing methods for crowdsourcing NER annotations by more reliably identifying entity span boundaries through a completely unsupervised crowdsourcing workflow. We demonstrate the quality of the annotations generated by improving language models trained on the resulting data.

This paper is organized as follows: section 2 outlines the process of collecting and cleaning text variants; section 3 describes how named entity annotations are generated for these text variants, in section 4, we evaluate the resulting data set and section 5 presents the discussion and conclusions.

2. Text variant elicitation

The process of collecting text variants begins with defining the scenarios that need to be represented within the collection. Scenarios are based entirely upon what a given product is designed to support and are intended to provoke responses from workers in the crowd that simulate what a real user would say when using a particular voice recognition product.

Once the scenarios are defined, we begin the first step of our crowdsourcing utterance collection process. We refer to these collections as variant elicitation. During the variant elicitation process, scenarios are presented to workers through a crowdsourcing platform. The scenarios give the workers a brief context pertaining to the kind of product they are using and the worker is asked to write a text command (variant) as if they were asking a voice recognition enabled personal assistant to perform a specified function. For example, if we wanted to collect variations of commands that might be used to turn off a device, we might ask workers the following: "Imagine you have just finished using our device. Think of two ways you might speak to the device to turn it off."

Variant elicitation is, by definition, a creative task and is therefore difficult to control when done through crowdsourcing. Standard crowdsourcing means of controlling quality such as using test questions do not apply to this task. As a means of improving quality and preventing automated spam from bots in such tasks, we have introduced a simple mathematical CAPTCHA question that is presented with every five variant elicitation scenarios. The inclusion of a periodic CAPTCHA has proven to be an effective means of quality control (see Table 1, where we display encountered spam rates for Calendar, Point of Interest (POI) and Short Message Dictation (SMD) variant elicitation tasks, across 240 units per domain).

<table>
<thead>
<tr>
<th>Domain</th>
<th>Spam % No CAPTCHA</th>
<th>Relative Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calendar</td>
<td>16.7</td>
<td>7.9</td>
</tr>
<tr>
<td>POI</td>
<td>27.1</td>
<td>18.8</td>
</tr>
<tr>
<td>SMD</td>
<td>21.7</td>
<td>15.8</td>
</tr>
</tbody>
</table>

Table 1. Spam rejection rates of quality control measures.

Once the variants have been collected, they undergo two spellchecking stages. The first stage involves running all the collected variants through an automated spellchecker. Variants are flagged during this process but not thrown out. In stage two, the variants that were flagged by the automatic spellchecker are deployed in another crowdsourcing task wherein workers are asked to identify if an utterance really contains a spelling error. If a variant is found to contain an error, it is discarded. Test questions and inter-annotator agreement are used to control quality in this stage. In prior collections, roughly 10% has been flagged for spelling errors.

After variants have been processed through the spellchecking stage, those that pass are sent to a domain-validation stage. In this stage, the variants in question are displayed to workers alongside several golden variants which we found to be a representative sample of relevant responses to the original scenarios that were presented to workers during the elicitation phase. Workers are asked to decide if the variant being evaluated belongs with the group of golden variants. This strategy allows untrained workers to accurately match responses to domains without having to explicitly describe the constraints of that domain to the worker. It is easier for an untrained worker to decide if a variant is similar to a collection of domain appropriate variants than it is for a worker to explicitly learn and apply the constraints of that domain. Approximately 20% of the data set is typically determined to be irrelevant in this step and is discarded.

The variants that pass this stage are considered clean and we now refer to them as utterances. An internal review of 1000 clean utterances by a product expert found that 99% fit within the scope of the original intent of the collection.

3. Annotation Process

After utterances have been collected and processed through the entire quality control pipeline, they are ready to start the annotation process using unmanaged crowds.

3.1 Background

Our own research has indicated that a majority of workers on the particular platform that we use for crowdsourcing do not read the instructions for the tasks they complete. We created an experimental task in which the correct answer to every question was hidden within the instructions and could not be deduced from the question alone. Only 209 of 750 (28%) workers read the instructions carefully and entered the correct response. In another iteration of the experiment, we explicitly warned workers in the title that they needed to read the instructions carefully and just 217 of 384 (56%) workers passed. From these experiments we concluded that an effective crowdsourcing task must assume that workers will not read the instructions, meaning the actual task must be intuitive enough to be understood without external explication. For this reason, as will be shown, we break our annotation process into smaller, more manageable pieces. Complex annotations can then be synthesized from simplified constituent tasks that minimize the amount of learning each worker is required to do. Here we will detail this piecewise annotation process for NER annotation in particular, but it should be noted that this process can be used just as effectively for more complex types of annotation, such as intent and action annotation for NLU modeling.
3.2 Domain classification

If utterances have been collected in a manner similar to the process described in section 2 of this paper, then domain classification is unnecessary as one should already possess the information necessary to classify the utterances. If one does not already possess domain information, a crowdsourcing task must be run in order to classify the utterances so that they can be subdivided for efficiency.

A simple classification task is deployed where workers are presented with individual utterances along with a list of possible domains and asked which domain is appropriate. Test questions and inter-annotator agreement rules work as effective quality control measures in this stage.

With the utterances classified by domain, they then enter the entity annotation phase with specialized annotation tasks for each domain. The domains can be further divided into batches depending on the complexity of the annotation.

3.3 Entity annotation

In this paper, we report an experiment within the media/entertainment field; utterances were divided into four domains: TV SHOW (e.g. abc comedy shows), MOVIE (e.g. bruce willis films), SPORTS (e.g. yankees versus mariners), and CONTROL (e.g. change to channel three hundred). After utterances are classified under their appropriate domains, they are deployed in multiple specialized entity annotation tasks. In these jobs, the utterances are presented to workers alongside empty tagging fields where they can insert words that fit under the scope of each specific tag. For simplicity, each tag can only be used once for a contiguous span of text. To prevent workers from rushing through the task, they are required to enter a token into the empty fields if there are no words that belong there. The merits of allowing workers to indicate any uncertainty in their judgments are explained by [2]; we also utilized this approach in our implementation.

![Screenshot of our crowdsourcing task.](image)

For the purpose of making the tasks simpler and reducing the learning curve, a maximum of four possible tags are presented to workers within each job; our experiments found that quality suffered when more than four tags were presented to workers within a single task. These findings were consistent with [10], which shows how reduced cognitive load improves worker quality. If the utterances within a domain require tagging of more than four entities, the task is split into multiple batches. This allows us to accommodate any number of entities and the unmanaged crowd can still tag accurately without being overloaded with information. Table 2 shows the way in which we divided the domains into batches. Partially annotated utterances from multiple batches are then combined using an automated script. One potential disadvantage of this approach is that the entities from different batches might have conflicts (e.g. overlapping entity spans); this can be resolved in the aggregation stage using rules that specify cases in which some tags take precedent over others.

The results in section 4 show that the inter-annotator agreement rule, whereby two annotators must come up with the exact same answer in order for it to be considered valid, proved to be a sufficient mechanism for spam prevention when paired with test questions.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Batch</th>
<th>Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV Shows</td>
<td>1</td>
<td>CHANNEL, GENRE, PROGRAM, PERSON</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>SEASON, EPISODE, PERIOD</td>
</tr>
<tr>
<td>Movies+</td>
<td>1</td>
<td>CHANNEL, GENRE, PROGRAM, PERSON</td>
</tr>
<tr>
<td>Control</td>
<td>1</td>
<td>CHANNEL, GENRE, PROGRAM, PERSON</td>
</tr>
<tr>
<td>Sports</td>
<td>2</td>
<td>TEAM1, TEAM2, LEAGUE, SPORTS</td>
</tr>
</tbody>
</table>

Table 2. List of domains and schema for dividing their entities into batches.

The inter-annotator agreement rule is most effective at preventing spam annotations for longer utterances. The utterance “find seattle mariners baseball game” when deployed in its first batch for sports annotation has 1369 unique possible unique annotations. Therefore the probability of chance agreement between the first two annotators is 1 in 1369. We become increasingly confident in the accuracy of annotations as the length of the utterance increases. Table three shows the probabilities of chance agreement for different utterance lengths. As the length increases, probability of chance agreement decreases exponentially.

<table>
<thead>
<tr>
<th>Length (words)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>1/5</td>
<td>1/25</td>
<td>1/109</td>
<td>1/413</td>
<td>1/1369</td>
<td>1/4021</td>
</tr>
</tbody>
</table>

Table 3. Probability of chance agreement between two annotators in a task with four tags.

The media/entertainment utterances used in our experiments had an average utterance length of 3.9 words. This suggests an average chance agreement of about 1 in 400 for our data set. We found that this probability was sufficiently low to ensure high quality annotations. Data sets with shorter utterances may require additional spam prevention measures to counter the higher probability of chance agreement.

4. Experimental design and results

Several experiments were designed to demonstrate the usefulness of the proposed process leveraging unmanaged crowdsourced annotation. In the first experiment we compare in-house annotations with crowd annotations and measure the agreement in terms of NER performance. The in-house annotation was used as our baseline and was done by an experienced annotator who is familiar with the domain scope and annotation ontology and requirements. A set of 2000 utterances was randomly selected from a large unlabeled corpus of media and entertainment related utterances. Utterances can contain one or more entities, as well
as no entities. As shown in Table 2, a total of 11 unique entities are possible. The same dataset was manually annotated by the in-house annotator and the crowd. We refer to these datasets as in-house and crowd in the following discussion.

To measure the extent of agreement and hence the annotation consistency, a test protocol was created. The annotated dataset is shuffled and a random 80% is used for NER (standard Conditional Random Fields) model training, whereas the remaining 20% is used for evaluation. The process is repeated five times. In this experiment, average and standard deviation over the five runs are reported for Precision, Recall and F1-score measures. The average precision is expected to be higher and standard deviation lower for higher consistency annotations. The former indicates the overall annotation stability within a dataset (e.g. following the same convention of annotation for all utterances), while the latter measures the amount of noise (e.g. different annotation of the same structure utterances).

The summary of results reporting average Precision (Prec), Recall (Rec), and F1-score over 5 runs for both annotated datasets is shown in Table 4.

In a second part of this experiment, we compared in-house annotation and crowd annotations with respect to a golden set. In this experiment no NER model was trained. As shown in Table 5, crowd annotations match the golden set more often than in-house annotations. The F1-measure is 0.579 and 0.844 for one person and two person annotation respectively.

**Table 5. Comparison of in-house annotations against the golden set and crowd annotations against the golden set.**

![Table 5](image)

True Positive (TP), False Positive (FP) and False Negative (FN) counts are frequently used to assess performance of an information retrieval system. In this analysis, the counts are performed at utterance level evaluating the tagging performance within an utterance. True Positive counts reflect the number of correctly tagged entities: both the label and the boundary must match with respect to some reference. The False Positives include cases when the label is incorrect or the boundary does not align perfectly with the reference. False Negatives include cases when an entity was not tagged.

**5. Discussion and conclusion**

Our results indicate that the crowd produced more accurate annotations than our in-house annotator. The relatively low performance of our in-house annotator might suggest that the annotator chosen is unskilled; however our annotator is a professional within the speech field with years of experience. The lower performance may be due to unfamiliarity with American television. We believe that this lack of expertise with a particular subject matter is typical of available in-house annotators within the industry. Given that the crowd was unmanaged, their superior performance indicates that combining the work of individuals who are not experts can allow us to synthesize high-quality annotations. A single worker may not have broad expertise, but it is likely that some workers within the crowd are familiar with given entities and their cumulative knowledge can therefore fill in the gaps.

Creating a golden NER annotation set in the media/television domain is difficult due to ambiguities which allow for multiple correct annotations. It can be argued that the utterance *The Tonight Show with Jay Leno* should be tagged entirely as a program, or that *Tonight Show* should be tagged program and *Jay Leno* should be tagged as person. By removing utterances that caused disagreement between annotators, we tried to eliminate some of this ambiguity; however, we were able to find crowd annotations which differed from our gold annotations but could still be considered correct. An internal review of 1000 crowdsourced annotations by two in-house annotators judged 99% of annotations were correct. These did not necessarily match golden annotations but they still fell within the acceptable parameters set by our annotation guidelines.

In this paper, we described our process for generating high quality NER annotated text utterances from unmanaged crowds. Through the use of a battery of quality control measures and piecewise division of labor to reduce task complexity, we were able to produce annotations which were more reliable and accurate than those of a single annotator.
6. References


