# A Fun and Engaging Interface for Crowdsourcing Named Entities

Kara Greenfield[1], Kelsey Chan[2], Joseph P. Campbell[1]
[1]MIT Lincoln Laboratory, 244 Wood St Lexington MA, USA
[2]MIT, 77 Massachusetts Avenue, Cambridge MA, USA
kara.greenfield@ll.mit.edu, kelseyc@mit.edu, jpc@ll.mit.edu

## Abstract

There are many current problems in natural language processing that are best solved by training algorithms on an annotated in-language, in-domain corpus. The more representative the training corpus is of the test data, the better the algorithm will perform, but also the less likely it is that such a corpus has already been annotated. Annotating corpora for natural language processing tasks is typically a time consuming and expensive process. In this paper, we provide a case study in using crowd sourcing to curate an in-domain corpus for named entity recognition, a common problem in natural language processing. In particular, we present our use of fun, engaging user interfaces as a way to entice workers to partake in our crowd sourcing task while avoiding inflating our payments in a way that would attract more mercenary workers than conscientious ones. Additionally, we provide a survey of alternate interfaces for collecting annotations of named entities and compare our approach to those systems.

Keywords: Mechanical Turk, crowd sourcing, named entity recognition, named entity annotation, natural language processing

## 1. Introduction

Annotated linguistic corpora are a key resource in developing natural language processing algorithms. Many of these algorithms require that their annotated training data is in the same domain as the test data in order to achieve maximal system accuracy. Crowdsourcing platforms such as Amazon's Mechanical Turk have been shown to be an effective way to quickly and economically gather annotations on text corpora for a variety of annotation tasks. While annotators who have been trained as professional linguists are able to annotate accurately and consistently from dense annotation guidelines, the amateur annotators who serve as workers on crowdsourcing platforms are not similarly motivated to create the best annotations possible. Financial incentives are the most common motivator used with crowdsourcing workers, but it can be beneficial to include alternative incentives as well, such as making the annotation task enjoyable.

## 2. Named Entity Recognition

Named Entity Recognition (NER) is the subtask of information extraction and consists of automatically extracting named mentions of entities (as opposed to nominal or pronominal mentions) from natural language text.

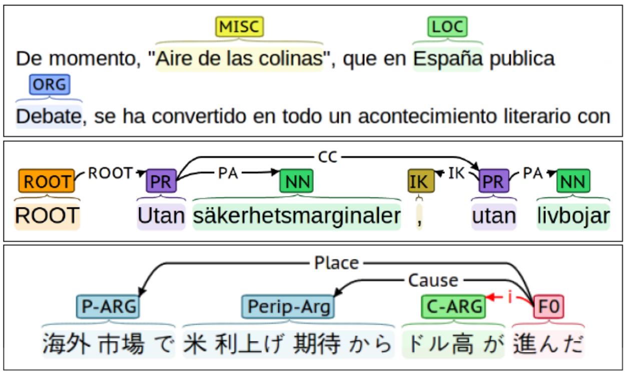
The ontology of which types of named entities are to be extracted varies according to application domain. Common ontology sets include Person, Organization, Location; Person, Organization, Location, Date; and Person, Organization, Geopolitical Entity. There have also been several NER systems developed for more specialized ontologies, such as in the medical domain. There are currently several state-of-the-art named entity extractors; however, due to the limited pool of annotated data available, these models are commonly limited to training on formal domains, such as news articles and scientific texts (Finin, et al., 2010); (Nadeau & Sekine, 2007). It is well known that the domain of the training data, which includes both textual genre (journalistic, scientific, informal, etc.) and topic (politics, arts, medicine, etc.) impacts the performance of the system on test data from other domains. For example, Poibeau and Kosseim (2001) showed that some systems yielding F-scores of more than 0.85 on newspaper articles experienced a drop in performance of up to 50% when tested on more informal texts like manual transcriptions of phone conversations and technical emails. Consequently, there is a need for in-language, in-domain annotated corpora with which to train current state-of-the art NER systems.

## 3. Traditional User Interfaces for NER Annotation

Most traditional user interfaces for collecting NER annotations allow the annotator to read through the passage once, annotating entity mentions of all classes within the ontology as a single task. Two of the most commonly used off-line annotation tools for collecting NER annotation are the BRAT Rapid Annotation Tool shown in Figure 1 **(Stenetorp, et al., 2012)** and Callisto **(MITRE, 2013)**. These tools allow the annotator to select a segment of text and then select the appropriate annotation label for that segment. This allows for the annotator to annotate multiple entity types simultaneously, but consequently requires that they mentally keep track of the definitions for those multiple

*This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

entity types and go through the process of both selecting the mention and then selecting a label for that mention. Combining the subtasks of annotating mentions of each separate entity type typically saves time for an experienced annotator, who has a good understanding of linguistics in general and the specific definition of the entity classes that they are trying to identify. For novice annotators, such as are likely to participate in a crowd sourcing task, combining tasks can prove to be too difficult, lowering the accuracy of the resulting annotations.



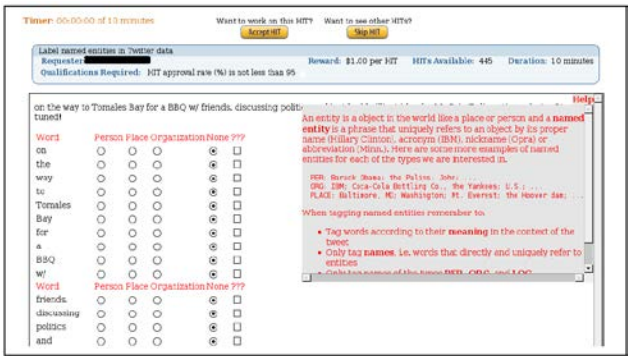
De momento, "Aire de las colinas", que en España publica Debate, se ha convertido en todo un acontecimiento literario con

ROOT Utan säkerhetsmarginaler , utan livbojar

海外 市場 で 米 利上げ 期待 から ドル高 が 進んだ

Figure 1 BRAT Rapid Annotation Tool (Stenetorp, et al., 2012)

In addition to off-line annotation tools, there are also several NER annotation interfaces that have been custom-designed for use by crowd sourcing workers on Amazon Mechanical Turk. Some of these are very similar to typical off-line NER annotation tools, requiring the annotator to simultaneously search for entity mentions of all of the types in the ontology. An example of such a system is the Twitter NER Annotation system shown in Figure 2 (Finin, et al., 2010). An interface such as this is relatively easy to create using the built-in requester tools in Mechanical Turk, but forces the annotator to read the passage with one word on each line, limiting the document length that is reasonable to include in a single human intelligence task (HIT). For named entity mentions that consist of only a single token, this interface allows the annotator to indicate as such with only a single mouse click; however an additional click is required for each additional word in the named entity mention.



Timer: 00:00:00 of 10 minutes

Want to work on this HIT? Want to see other HITs?

Label named entities in Twitter data

Requester: [redacted] Reward: $1.00 per HIT HITs Available: 445 Duration: 10 minutes

Qualifications Required: HIT approval rate (%) is not less than 95

on the way to Tomales Bay for a BBQ w/ friends. discussing politi... tuned!

| Word | Person | Place | Organization | None | ??? |
|---|---|---|---|---|---|
| on | ○ | ○ | ○ | ● | □ |
| the | ○ | ○ | ○ | ● | □ |
| way | ○ | ○ | ○ | ● | □ |
| to | ○ | ○ | ○ | ● | □ |
| Tomales | ○ | ○ | ○ | ● | □ |
| Bay | ○ | ○ | ○ | ● | □ |
| for | ○ | ○ | ○ | ● | □ |
| a | ○ | ○ | ○ | ● | □ |
| BBQ | ○ | ○ | ○ | ● | □ |
| w/ | ○ | ○ | ○ | ● | □ |
| Word | Person | Place | Organization | None | ??? |
| friends. | ○ | ○ | ○ | ● | □ |
| discussing | ○ | ○ | ○ | ● | □ |
| politics | ○ | ○ | ○ | ● | □ |
| and | ○ | ○ | ○ | ● | □ |

An entity is a object in the world like a place or person and a **named entity** is a phrase that uniquely refers to an object by its proper name (Hillary Clinton), acronym (IBM), nickname (Opra) or abbreviation (Minn.). Here are some more examples of named entities for each of the types we are interested in.

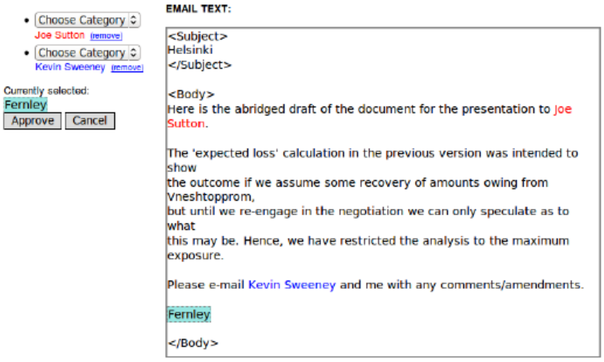
PER: Barack Obama; the Palins; John; ...
ORG: IBM; Coca-Cola Bottling Co.; The Yankees; U.S.; ...
PLACE: Baltimore, MD; Washington; Mt. Everest; the Hoover dam; ...

When tagging named entities remember to:

- Tag words according to their **meaning** in the context of the tweet
- Only tag **names**, i.e. words that directly and uniquely refer to entities

Figure 2 Twitter NER Annotation in Mechanical Turk (Finin, et al., 2010)

An alternative user interface for collecting named entity mention annotations through Mechanical Turk was presented by Lawson et. al. (2010). This was an improvement on previous NER annotation systems in that it included several interface features that were specifically designed to ease the annotation burden on novice annotators, such as Mechanical Turk workers. These features included allowing the user to select spans of text instead of individually clicking on each word and having separate tasks for annotating each type of entity in order to decrease the required mental load. Additionally, this interface had workers annotate both named and nominal entity mentions in an attempt to help workers realize that there is a distinction between named and nominal mentions. This interface can be seen in Figure 3. The usability improvements in this interface were obtained at the cost of needing to create a custom interface for the HITs instead of using one of the default HIT templates. The available templates are not optimal for natural language annotations and the developer cost incurred in creating a custom interface is offset by the resulting increase in annotation quality and decrease in annotation time.



- Choose Category
  Joe Sutton (remove)
- Choose Category
  Kevin Sweeney (remove)

Currently selected:
Fernley

EMAIL TEXT:

```
<Subject>
Helsinki
</Subject>

<Body>
Here is the abridged draft of the document for the presentation to Joe
Sutton.

The 'expected loss' calculation in the previous version was intended to
show
the outcome if we assume some recovery of amounts owing from
Vneshtopprom,
but until we re-engage in the negotiation we can only speculate as to
what
this may be. Hence, we have restricted the analysis to the maximum
exposure.

Please e-mail Kevin Sweeney and me with any comments/amendments.

Fernley

</Body>
```

Figure 3 Span-based NER Annotation in Mechanical Turk (Lawson, et al., 2010)

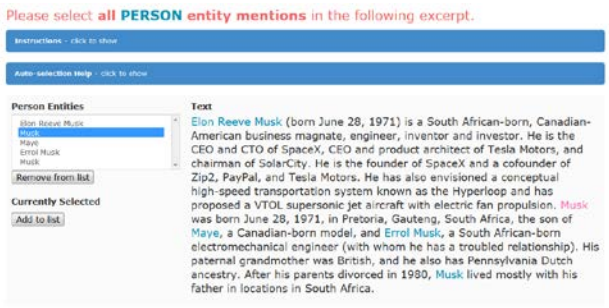
## 4. MITLL NER Crowdsourcing Annotation System

The MIT Lincoln Laboratory named entity crowd sourcing annotation system maximizes annotation accuracy and efficiency through a combination of 1) a clean user interface that minimizes annotator workload, 2) clear annotation guidelines, and 3) and a methodology for assigning HITs to workers which minimizes low annotation recall.

### 4.1 User Interface

Our annotation interface built upon the features developed by Lawson et. al. (2010). Our enhancements were primarily focused on minimizing the effort that a worker had to exert in order to annotate a document. By not having workers annotate nominal entity mentions, they were only required to select a text span and click a single button in order to annotate it as a named entity mention. We used color to allow the user to visually see all of the entity mentions that they already annotated and also the specific text span that they are currently annotating.

We also placed an emphasis on including annotation instructions that were specifically tailored to novice linguistic annotators. Our instructions consisted of a simple definition of a named entity combined with several examples of text spans that were examples of named entities in addition to negative examples. We found that including negative examples in the instructions was particularly beneficial for both increasing annotation accuracy and decreasing the number of workers who emailed us to ask for clarification of the instructions. While detailed instructions are invaluable for assisting the workers, they also require a large amount of screen real estate. We counteracted this by making the bulk of the instructions optionally visible, but always having the simplest form of the instructions (telling the annotator which type of named entity they were supposed to be identifying) visible in large font in a bright color at the top of the screen. Early versions of our experiments didn’t have this and resulted in several annotators who otherwise had very high annotation accuracy accidentally annotating the wrong entity type. The system can be seen in Figure 4.



Please select **all PERSON entity mentions** in the following excerpt.

Elon Reeve Musk (born June 28, 1971) is a South African-born, Canadian-American business magnate, engineer, inventor and investor. He is the CEO and CTO of SpaceX, CEO and product architect of Tesla Motors, and chairman of SolarCity. He is the founder of SpaceX and a cofounder of Zip2, PayPal, and Tesla Motors. He has also envisioned a conceptual high-speed transportation system known as the Hyperloop and has proposed a VTOL supersonic jet aircraft with electric fan propulsion. Musk was born June 28, 1971, in Pretoria, Gauteng, South Africa, the son of Maye, a Canadian-born model, and Errol Musk, a South African-born electromechanical engineer (with whom he has a troubled relationship). His paternal grandmother was British, and he also has Pennsylvania Dutch ancestry. After his parents divorced in 1980, Musk lived mostly with his father in locations in South Africa.

Figure 4 MITLL NER Annotation Interface

## 4.2 Data Selection and Incentives

Annotator fatigue is a common problem in many annotation scenarios, including crowd sourcing. Failing to counteract this leads to generating annotated corpora that are missing many annotations and consequently can’t be utilized as gold standards. This problem occurs even when the annotators are trained linguists, but is compounded in crowd sourced annotation due to the fact that many of the workers are not motivated to care about the quality of the final corpus. Lawson et. al. (2010) addressed this problem by monetarily incentivizing workers based on the number of entities that they annotated. While this methodology did encourage workers to annotate more than just the first few entity mentions in each HIT, it can have the unintended negative consequence of motivating workers to annotate text spans that are not actually entity mentions. The same financial motivations that would lead to a worker not annotating all of the entity mentions in order to annotate more documents when the financial reimbursement is proportional to the number of documents would lead to those workers annotating an abundance of false positives when the financial reimbursement is proportional to the number of annotations. An additional shortcoming of incentivizing workers based on the number of annotations they return is that the cost of creating the corpus increases by an unpredictable amount.

We primarily chose to address the problem of annotator fatigue by identifying and correcting for it rather than disincentivizing it as Lawson et. al. did (2010). The first way in which we did this (as shown in Figure 5) was to avoid having the same portion of the text occur at the end of the HIT for all of the workers who annotated that HIT. Each document was split into chunks of no more than 500 characters. All excerpts began and ended at sentence breaks so that workers would understand the context of the excerpt. Every HIT contained two excerpts. If an excerpt appeared first out of two in one HIT, it would appear again as the second of two in another HIT. Additionally, we ran all of the documents through our automatic NER system, MITIE (King, n.d.), (Geyer, et al., 2016). We took all of the documents in which MITIE identified entity mentions that weren’t annotated by either of the original two workers for that document and presented those sections of text again to a new worker in order to either verify that there was no entity mention or to recover from the low recall of the other workers. Adjudicating automated system output allowed us to benefit from having additional annotations only where they were needed without having to pay to have them on the entire corpus.

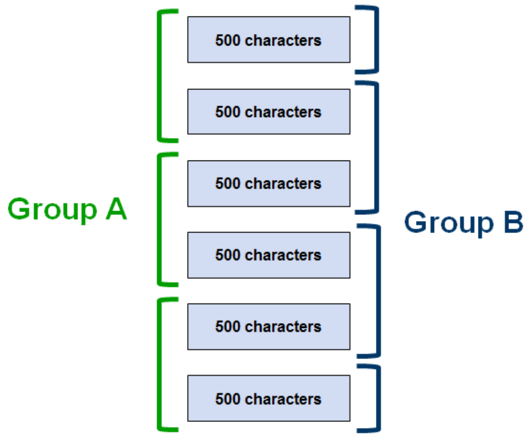


Figure 5 Document partitioning

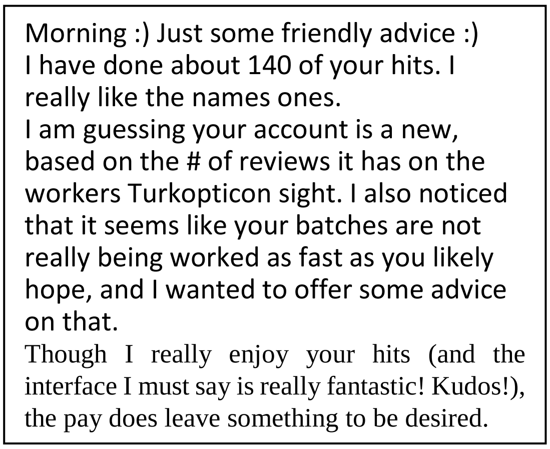
We did also appeal to workers’ morals and sense of human connection to discourage them from submitting HITs without reading or annotating the text. We accomplished this via a text prompt whenever a user submitted a HIT without any annotations, asking them if they were sure that there weren’t any named entity mentions in that HIT.

# 5. Worker Feedback

We found emails from workers to be an extremely valuable source of feedback on both our interface design and annotation instructions. While we never explicitly prompted or asked users for feedback, many voluntarily provided it.

One of the greatest benefits that we gained from the pilot runs of our experiments was user feedback on examples in the data where they were unsure of whether or not they should annotate a particular span of text as a named entity mention. In addition to responding to that worker, we used many of those cases as examples in our instructions in the final run of the experiment and correspondingly saw a decrease in such clarification requests which lessened our workload.

Of particular interest was that many of the workers were particularly motivated to maintain their approval rating on Amazon Mechanical Turk. While we didn't say we would reject any HITs or actually reject any HITs, or have any history of ever rejecting HITs on this requestor account, the vast majority of email requests for clarification on the guidelines also informed us that they were diligently trying to complete the HITs accurately and requested that we not reject their HITs if they made a mistake because they were afraid of that negatively affecting their approval rating. There is very likely a positive correlation between a worker being motivated enough to ask for clarification on the guidelines rather than taking their best guess and that worker caring about requestors' opinions of them, so this motivation may not be present in all workers, but it is very strong in those who do possess it.



Morning :) Just some friendly advice :)
I have done about 140 of your hits. I really like the names ones.
I am guessing your account is a new, based on the # of reviews it has on the workers Turkopticon sight. I also noticed that it seems like your batches are not really being worked as fast as you likely hope, and I wanted to offer some advice on that.
Though I really enjoy your hits (and the interface I must say is really fantastic! Kudos!), the pay does leave something to be desired.

Figure 6 Worker Feedback

We also found that many workers were motivated by the ease of use of the interface, even when they thought that the task warranted a higher financial incentive. Figure 6 shows feedback from one of the workers who completed our HITs. Due to their enjoyment in completing these HITs and the clean interface design, this worker accurately annotated many of our HITs, despite believing that they could obtain a higher hourly rate by completing other HITS. As this worker illuminated, increased financial incentives can serve to decrease the time required to complete a batch of HITs, but with a good interface design, a slightly lower rate can also yield accurate annotations, just in a slightly longer time frame. While this was the only worker who provided us with feedback on pricing, we received many comments from other workers stating that they found the task enjoyable and especially liked the interface.

## 6. Conclusions

In this work, we presented a system for gathering named entity recognition annotations via crowd sourcing that builds upon prior work in developing natural language annotation interfaces. We provided a methodology for overcoming the low recall rates that are common among novice annotators. Additionally, we analysed worker feedback to show that having an annotation interface that is easy to use can be a strong incentive for crowd sourcing workers. The primary motivators that we identified other than HIT pricing were maintaining a positive worker rating (which is indirectly a financial incentive) and ease of interface use. In future work, we would like to expand this system to allow for more complicated linguistic annotations, especially those that require annotating multiple disjoint spans of text for a single annotation.

## 7. Acknowledgements

The authors would like to thank Scott Briere and Nicholas Malyska for their assistance with setting up our crowdsourcing experiments. We would also like to thank all of the Amazon Mechanical Turk workers for completing our HITs and providing feedback to us.

## 8. References

Finin, T. et al., 2010. *Annotating Named Entities in Twitter Data with Crowdsourcing.* s.l., s.n., pp. 80-88.

Geyer, K., Greenfield, K., Mensch, A. & Simek, O., 2016. *Named Entity Recognition in 140 Characters or Less.* s.l., s.n.

King, D., n.d. *MITLL/MITIE.* [Online] Available at: https://github.com/mit-nlp/MITIE

Lawson, N., Eustice, K., Perkowitz, M. & Yetisgen-Yildiz, M., 2010. *Annotating Large Email Datasets for Named Entity Recognition with Mechanical Turk.* s.l., s.n., pp. 71-79.

MITRE, 2013. *Callisto.* [Online] Available at: http://mitre.github.io/callisto/index.html

Nadeau, D. & Sekine, S., 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes,* 30(1), pp. 3-26.

Poibeau, T. & Kosseim, L., 2001. *Proper Name Extraction from Non-Jornalistic Texts.* s.l., s.n.

Stenetorp, P. et al., 2012. *BRAT: a Web-based Tool for NLP-Assisted Text Annotation.* s.l., s.n., pp. 102-107.