# Non-Expert Correction of Automatically Generated Relation Annotations

**Matthew R. Gormley**$^{\star\dagger}$ and **Adam Gerber**$^{\star\dagger}$ and **Mary Harper**$^{\star\ddagger}$ and **Mark Dredze** $^{\star\dagger}$

$^{\star}$Human Language Technology Center of Excellence
$^{\dagger}$Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21211, USA
$^{\ddagger}$Laboratory for Computational Linguistics and Information Processing
University of Maryland, College Park, MD 20742 USA

mrg@cs.jhu.edu,adam.gerber@jhu.edu,mharper@umd.edu,mdredze@cs.jhu.edu

## Abstract

We explore a new way to collect human annotated relations in text using Amazon Mechanical Turk. Given a knowledge base of relations and a corpus, we identify sentences which mention both an entity and an attribute that have some relation in the knowledge base. Each noisy sentence/relation pair is presented to multiple turkers, who are asked whether the sentence expresses the relation. We describe a design which encourages user efficiency and aids discovery of cheating. We also present results on inter-annotator agreement.

## 1 Introduction

Relation extraction (RE) is the task of determining the existence and type of relation between two textual entity mentions. Slot filling, a general form of relation extraction, includes relations between non-entities, such as a person and an occupation, age, or cause of death (McNamee and Dang, 2009).

RE annotated data, such as ACE (2008), is expensive to produce so systems take different approaches to minimizing data needs. For example, tree kernels can reduce feature sparsity and generalize across many examples (GuoDong et al., 2007; Zhou et al., 2009). Distant supervision automatically generates noisy training examples from a knowledge base (KB) without needing annotations (Bunescu and Mooney, 2007; Mintz et al., 2009). While this method can quickly generate training data, it also generates many false examples. We reduce the noise in such examples by using Amazon Mechanical Turk (MTurk), which has been shown to produce high quality annotations for a variety of natural language processing tasks (Snow et al., 2008).

We use MTurk for annotation of textual relations to establish an inexpensive and rapid method of creating data for slot filling. We present a two step annotation process: (1) automatic creation of noisy examples, and (2) human validation of examples.

## 2 Method

### 2.1 Automatic generation of noisy examples

To create noisy examples we use a similar approach to Mintz et al. (2009). We extract relations from a KB in the form of tuples, $(e, r, v)$, where $e$ is an entity, $v$ is a value, and $r$ is a relation that holds between them; for example (J.R.R. Tolkien, occupation, author). Our KB is Freebase[1], an online database of structured information, and our corpus is from the TAC KBP task (McNamee and Dang, 2009)[2]. For each tuple, we find sentences in a corpus that contain both an exact mention of the entity $e$ and of the value $v$. Of course, such sentences may not attest to the relation $r$, so the process produces many incorrect examples.

### 2.2 Human Intelligence Tasks

A Human Intelligence Task (HIT) is a short paid task on MTurk. In our HITs, we present the turker with ten relation examples as sentence/relation pairs. For each example, the user is asked to select from three annotation options: the sentence (1) expresses the relation, (2) does not express the relation, or (3) the

[1] http://www.freebase.com

[2] http://projects.ldc.upenn.edu/kbp/

*Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 204–207,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

1. The sentence expresses the relation.
   *Sentence*: For the past eleven years, James has lived in Tucson.
   *Relation*: “Tucson” is the residence of “James”
2. The sentence does not express the relation.
   *Sentence*: Samuel first met Divya in 1990, while she was still a student.
   *Relation*: “Divya” is a spouse of “Samuel”
3. The relation does not make sense.
   *Sentence*: Soojin was born in January.
   *Relation*: “January” is the birth place of “Soojin”

Figure 1: The three annotation options with examples.

relation does not make sense (figure 1.)

Of the ten examples that comprise each HIT, seven are automatically generated by the method above. The correct answer is known for the three remaining examples; these are included for quality assurance (control examples.) The three control examples are a positive example (expresses the relation,) a negative example (contradicts the true relation,) and a nonsense example (relation is nonsensical.)

All control examples derive from a subset of the automatically generated person examples. Positive examples were randomly sampled and hand annotated. Negative examples are familial relations in which we change the relation type so that it would not be expressed in the sentence. For example, the relation “Barack Obama is the parent of Malia Obama” would be changed to “Barack Obama is a sibling of Malia Obama.” To generate nonsense examples we employ the same method for a different mapping of relations, which produces relations like “New Zealand is the gender of John Key.”

### 2.3 HIT Design

MTurk is a marketplace so users have total freedom in choosing which HITs to complete. As such, HIT design should maximize its appeal. We assume that users find appealing those HITs through which they may maximize their own monetary gain, while minimizing moment-to-moment frustrations. We emphasized clarity and ease of use.

The layout consists of three sections (figure 2). The leftmost section is a progress list, which shows the user’s answers and current position; the middle section contains the current relation example and annotation options; the rightmost section (not pictured) contains instructions. All sections and all UI elements remain visible and in the same position for the duration of the HIT, with only the text of the sentence and relation changing according to question number. Because only a single question is displayed at a time, we are able to minimize user actions such as scrolling, clicking small targets, or making large mouse movements. Additionally, we can monitor how much time a user spends on each question.

| | *# HITs* | *Cost* | *Time (hours)* |
|---|---|---|---|
| Trial | 50 | $2.75 | 27 |
| Batch 1 | 500 | $27.50 | 34 |
| Batch 2 | 765 | $42.08 | 25 |
| Batch 3 | 500 | $27.50 | 22 |
| **Total** | 1815 | $99.83 | 108 |

Table 1: Size, cost and time to complete each HITs batch.

At all times the user is able to consult the instructions for the task, which include examples of each annotation option. The user is also reminded of the technical requirements for the HIT and expectations for honesty and accuracy. A comment box provides users with the opportunity to ask questions, make suggestions, or clarify their responses.

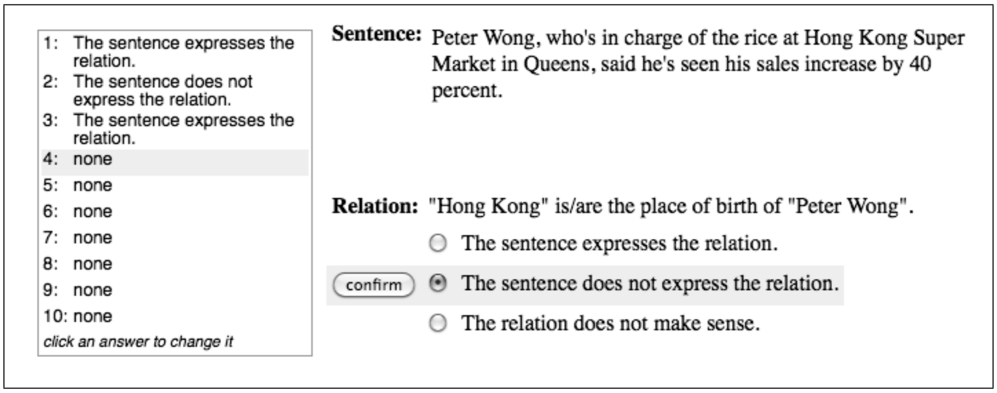
## 3 Results

We submitted a trial run and three full batches of HITs. Table 1 summarizes the costs and completion times for all HITs. The HITs were labeled rapidly and for a low cost ($0.05 per HIT, i.e., .5¢ per annotation). Each HIT was assigned to five unique workers. We found that 50% of the 352 different workers completed 2 or more HITs (figure 3.) Our results exclude a trial run of 50 hits. Across the 17,650 examples the mean time spent was 20.77 seconds, with a standard deviation of 99.96 seconds. The median time per example was 10.0 seconds.

### 3.1 Analysis

To evaluate the annotations, two of the authors annotated a random sample of 247 (10%) of the 2471 noisy examples. In addition, we analyzed the workers agreement with the control examples.

We used two metrics to assess agreement. The first metric is pairwise percent agreement (*Pairwise*): the average of the example agreement scores, where the example agreement score is the percent of



1: The sentence expresses the relation.
2: The sentence does not express the relation.
3: The sentence expresses the relation.
4: none
5: none
6: none
7: none
8: none
9: none
10: none
click an answer to change it

**Sentence:** Peter Wong, who's in charge of the rice at Hong Kong Super Market in Queens, said he's seen his sales increase by 40 percent.

**Relation:** "Hong Kong" is/are the place of birth of "Peter Wong".

○ The sentence expresses the relation.

● The sentence does not express the relation.

○ The relation does not make sense.

Figure 2: An example HIT with instructions excluded.

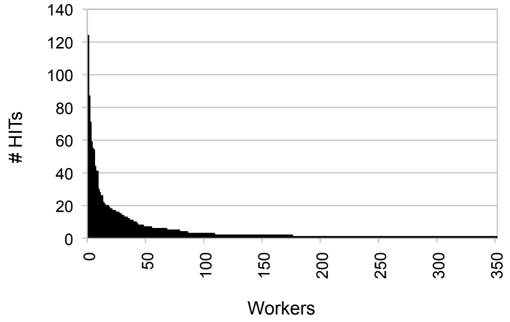


Figure 3: The number of HITs per worker, with columns sorted left to right.

pairs of annotators that agreed for a particular example. The second metric is the exact kappa coefficient (*Exact-$\kappa$*) (Conger, 1980), which takes into account that agreement can occur by chance. The number of annotators ($R$) varies with the test scenario.

Table 2 presents the inter-annotator agreement scores for various subsets of the examples and combinations of annotators. On a sample of examples, we evaluated agreement between the first and second expert annotators (*E1/E2*) and also the agreement between each expert and the majority vote of the workers (*E1/M* and *E2/M*). The agreement between the two experts is substantially higher than their individual agreements with the majority. Yet, we achieve our goal of reducing noise.

We also analyzed the agreement between the known control answer and the majority vote of the workers (*C/M*). This high level of agreement supports our belief that the automatically generated negative and nonsense examples were easier to identify than noisy negative and nonsense examples. Finally, we evaluated the agreement between the five workers for different subsets of the data: the sample of noisy examples (*T(sample)*), the control examples only (*T(control)*), and all examples (*T(all)*). Table 3 lists the number of examples collected and the agreement scores for all workers for each relation type.

|  | *# Ex.* | *R* | *Exact-$\kappa$* | *Pairwise* |
|---|---|---|---|---|
| *E1/E2* | 247 | 2 | 0.64 | 0.81 |
| *E1/M* | 247 | 2 | 0.29 | 0.60 |
| *E2/M* | 247 | 2 | 0.39 | 0.70 |
| *C/M* | 1059 | 2 | 0.90 | 0.93 |
| *T(sample)* | 247 | 5 | 0.31 | 0.69 |
| *T(control)* | 1059 | 5 | 0.52 | 0.68 |
| *T(all)* | 3530 | 5 | 0.45 | 0.68 |

Table 2: Inter-annotator agreement

Table 4 shows the divergence of the workers' annotations from those of an expert. The high level of confusability for those examples which the expert annotated as *Not Expressed* suggests their inherent difficulty. The workers labeled more examples as *Expressed* than the expert, but both labeled few examples as *Nonsense*.

## 4 Quality Control

We identify spurious responses and unreliable users in two ways. First, worker responses are compared to control examples; greater agreement with controls should indicate greater confidence in the user. We filtered any worker whose agreement with the controls was less than 0.85 (*Control Filtered*). The second approach uses behavioral data. Because only a single example is visible at any time, we can mea-

| *Relation* | *# Ex.* | *Exact-$\kappa$* | *Pairwise* |
|---|---:|---:|---:|
| siblings | 13 | 0.67 | 0.82 |
| children | 12 | 0.57 | 0.83 |
| gender | 80 | 0.46 | 0.70 |
| place_of_death | 40 | 0.43 | 0.68 |
| parent | 12 | 0.40 | 0.64 |
| spouse | 54 | 0.37 | 0.65 |
| title | 71 | 0.30 | 0.78 |
| residences | 228 | 0.29 | 0.60 |
| ethnicity | 38 | 0.28 | 0.54 |
| occupation | 551 | 0.26 | 0.77 |
| activism | 4 | 0.26 | 0.55 |
| religion | 22 | 0.23 | 0.55 |
| place_of_birth | 160 | 0.20 | 0.64 |
| nationality | 1044 | 0.19 | 0.67 |
| schools_attended | 8 | 0.16 | 0.55 |
| employee_of | 132 | 0.16 | 0.70 |
| charges | 2 | 0.14 | 0.70 |
| **Total** | 2471 | 0.35 | 0.69 |

Table 3: Inter-annotator agreement across relation type. *# Ex.* is the number of noisy examples. *Exact-$\kappa$* and *Pairwise* agreement are among the five workers.

| | | Worker | | | |
|---|---|---|---|---|---|
| | | E | NE | Nn | Total |
| Expert-1 | E | 561 | 89 | 20 | 670 |
| | NE | 284 | 248 | 28 | 560 |
| | Nn | 1 | 1 | 3 | 5 |
| | Total | 846 | 338 | 51 | 1235 |

Table 4: Confusion matrix of expert-1 and user's annotations on the sample of noisy examples, for the choices Expressed (E), Not Expressed (NE), and Nonsense (Nn)

sure how much time a user spends on each example. The UI is designed to allow for the extremely rapid completion of examples and of the HIT in general. Thus, a user could complete the HIT in only a few seconds without even reading any of the examples. Still other users spend only a moment on all-but-one question, and then several minutes on the remaining question. Here, we filter a user answering three or more questions each in under three seconds (*Time Filtered*). We combine these two approaches (*Control and Time)*, which yields the highest expert-agreement levels (table 5.)

| | *E1/M* | *E2/M* |
|---|---:|---:|
| *Unfiltered* | 0.28 | 0.38 |
| *Time Filtered* | 0.32 | 0.43 |
| *Control Filtered* | 0.34 | 0.47 |
| *Control and Time* | 0.37 | 0.48 |

Table 5: Exact-$\kappa$ scores for three levels of quality control and a baseline, between each expert and the majority vote on 231 sampled examples. For a fair comparison, we reduced the sample size to include only examples for which each level of quality control had at least one worker annotation remaining.

## 5 Conclusion

Using non-expert annotators from Amazon Mechanical Turk for the correction of noisy, automatically generated examples is inexpensive and fast. We achieve good inter-annotator agreement using quality assurance measures to detect cheating. The result is thousands of new annotated slot filling example sentences for 17 person relations.

## Acknowledgments

We would like to thank the 352 turkers who made this work possible.

## References

ACE. 2008. Automatic content extraction. http://projects.ldc.upenn.edu/ace/.

R. Bunescu and R. Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *Association for Computational Linguistics (ACL)*.

A.J. Conger. 1980. Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88(2):322–328.

Z. GuoDong, M. Zhang, D. H Ji, and Z. H. U. QiaoMing. 2007. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Paul McNamee and Hoa Dang. 2009. Overview of the TAC 2009 knowledge base population track. In *Text Analysis Conference (TAC)*.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Association for Computational Linguistics (ACL)*.

R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Empirical Methods in Natural Language Processing (EMNLP)*.

G. Zhou, L. Qian, and J. Fan. 2009. Tree kernel-based semantic relation extraction with rich syntactic and semantic information. *Information Sciences*.