

A Hybrid Approach towards Biomedical Relation Extraction Training Corpora: Combining Distant Supervision with Crowdsourcing [★]

Diana Sousa^{✉[0000–0003–0597–9273]} and Andre Lamurias^[0000–0001–7965–6536]
Francisco M. Couto^[0000–0003–0627–1496]

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
`dfsousa@lasige.di.fc.ul.pt`

Abstract. Biomedical Relation Extraction (RE) datasets are vital in the construction of knowledge bases, and to potentiate the discovery of new interactions. There are several ways to create biomedical RE datasets, some more reliable than others, such as resorting to domain expert annotations. However, the emerging use of crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk), can potentially reduce the cost of RE dataset construction, even if the same level of quality cannot be guaranteed. There is a lack of power of the researcher to control who, how, and in what context workers engage in crowdsourcing platforms. Hence, allying distant supervision with crowdsourcing can be a more reliable alternative. The crowdsourcing workers would be asked only to rectify or discard already existing annotations, which would make the process less dependent on their ability to interpret complex biomedical sentences. In this work, we use a previously created distantly supervised dataset of human phenotype-gene relations (PGR dataset) to perform crowdsourcing validation. We divided the original dataset into two annotation tasks: Task 1, 70% of the dataset annotated by one worker, and Task 2, 30% of the dataset annotated by seven workers. Also, for Task 2, we added an extra rater on-site and a domain expert, to further assess the crowdsourcing validation quality. Here, we describe a detailed pipeline for RE crowdsourcing validation, creating a new release of the PGR dataset with partial domain expert revision, and assess the quality of the MTurk platform. We applied the new dataset to two state-of-the-art deep learning systems (BiOnt and BioBERT) and compared its performance with the original PGR dataset, as well as combinations between the two, achieving 0.3494 average increase in F-measure. The code supporting our work and the new release of the PGR dataset will be made publicly available upon acceptance of this manuscript.

Keywords: Crowdsourcing · Distant Supervision · Biomedical Relation Extraction · Knowledge Bases · Training Corpora

[★] This work was supported by FCT through project DeST: Deep Semantic Tagger, ref. PTDC/CCI-BIO/28685/2017, LASIGE Research Unit, ref. UIDP/00408/2020, and PhD Scholarship, ref. SFRH/BD/145221/2019.

1 Introduction

Knowledge bases play a fundamental role in the way we store, organize and retrieve information. More specifically, biological knowledge bases are commonplace for researchers and clinicians to access all types of biomedical data retrieved from biomedical literature [1]. Previous works annotated biomedical literature by resorting to domain expert annotators [12], crowdsourcing platforms [29], or distantly supervised techniques [28]. The main aim of these researchers is to tackle the lack of annotated datasets for biomedical information extraction systems. However, when applying distantly supervised techniques, the annotations are not as reliable as when done by domain experts, and it still needs to be adequately reviewed before the extracted information can be added to any biomedical repository. Hence, the added advantage of automating the extraction of information using distant supervision is slightly impaired by the need to review it, which is not only costly but time and resource-consuming. Moreover, when targeting Relation Extraction (RE) between entities of different domains or document summarization tasks [24], the revision process becomes cumbersome when compared to other information extraction tasks, given its higher complexity that normally requires knowledge of multiple domains.

The alternative way to create reliable gold standard datasets that do not resort to domain expert curation could be allying distant supervision with crowdsourcing [10,21,5]. Before integrating data extracted from distant supervision pipelines into biological knowledge bases or using it as training data for biomedical information extraction systems, the data would go through a confirmation or review phase in the form of crowdsourcing. Crowdsourcing platforms are becoming increasingly popular to address the problem of lack of training corpora for natural language processing (NLP) tasks [3]. Currently, the most popular platform for this purpose is Amazon Mechanical Turk (MTurk) [13,31,14]. Some platforms created a trust layer over MTurk to facilitate task specification and monitoring [30], such as Figure Eight (previously known as CrowdFlower) [20,7], which is widely used by researchers for biomedical NLP related tasks.

One of the problems of using crowdsourcing platforms is the lack of domain expertise. While most platforms allow us to specify some criteria (e.g., degree of education), in exchange for an increased price per task, it is not feasible to specify expertise in particular biomedical domains. Not only that, but there is no guarantee that the quality promised is the quality provided because some malicious workers often take advantage of the difficulty in implementing a verification procedure and submit answers of low quality [3]. Task redundancy can be a solution, but it also increases the costs of using crowdsourcing approaches, partially defeating the purpose of these platforms. The question should be whether the quality of the workers is good enough for the purpose of the task, and if the difference in quality when compared to domain experts is compensated by the decrease in costs. In the case of the MTurk platform, some studies have supported its suitability for a variety of tasks [23]. However, it fails in transparency about its workers' context (e.g., background), if MTurk constitutes their primary form of income or not, what is their motivation for completing the tasks, and

if this introduces bias to the tasks at hand. These and other ethical questions have been discussed in depth by some researchers [8,25].

In this work, we leveraged an existing dataset of biomedical relations, created through distant supervision, and submitted it to the MTurk platform to perform crowdsourcing validation. With the exhaustive review of the performance of the original and new datasets, we assessed the viability of combining distant supervision and crowdsourcing for the field of biomedical RE.

Our work used an open-source dataset, the PGR dataset [28], based on distant supervision, that features both human phenotype and gene annotations and their relations. Since it is a silver standard dataset, it has not been reviewed by domain experts, leading to wrongly labeled relations and other errors. These errors can be from Named-Entity Recognition (NER) (e.g., acronyms of diseases annotated as genes), which was also done automatically, or sentence format errors. To rectify these errors, we used the MTurk platform to validate, alter, or discard the relations within the PGR dataset. We achieved this by dividing the original dataset into two partitions, one of 70% (Task 1), where each relation was rated by one Amazon worker, and another of 30% (Task 2), where each relation was rated by seven distinct workers. We validated our approach through inter-rater agreement using the Fleiss’ kappa [22] and the Krippendorff’s alpha [17] metrics for Task 2. Further, we also provided the 30% partition of the PGR dataset used for Task 2 to an external rater (on-site, without previous curating experience but holding a Biochemistry degree), and to a domain expert (with previous curating experience, holding a PhD in Bioinformatics). These different levels of expertise enlightened the difficulties of curating the dataset and the limitations associated with each level. To evaluate and compare the quality of the crowdsourced Amazon dataset, we applied it to two state-of-the-art deep learning systems and compared its performance with the original PGR dataset, as well as combinations between the two. The deep learning systems used were BiOnt [27] and BioBERT [19], that feature relation extraction between different biomedical entities with high performance, and, in the case of BiOnt, it was already used in conjunction with the PGR dataset.

The performance of the MTurk workers in comparison with our on-site curator and the domain expert was generally good for accessing NER or sentence format errors (approximately 16% of relations). However, the MTurk workers struggle to identify false relations (separate entities with no association in a sentence). The struggle to identify these relations can be due to the complexity of the sentences, or quality issues related to the MTurk platform validation of workers, which we will discuss in more detail in the following sections. Further, the inter-rater agreement for Task 2 showed a slight to a fair agreement (about 0.20-0.21), which is below what we expected and we believe could be related to the problems of sentence complexity and quality reported. Regarding the performance of the crowdsourced Amazon dataset in the application of the BiOnt and BioBERT systems, we had an average increase of 0.3494 in F-measure taking into account all the experiences, concerning the original PGR dataset.

The main takeaways of this work were the need for further validation of the use of crowdsourcing platforms, such as the MTurk platform, and the potential of using distant supervision allied with crowdsourcing to produce gold standard datasets with which we can train viable models and detect relevant biomedical relations. This work resulted in the following contributions:

- Pipeline for RE crowdsourcing, in which we describe in detail all the base concepts and steps taken to produce the new crowdsourced dataset.
- New release of the PGR Dataset, which will be made freely available to the community.
- Assessment of the quality of results obtained with the MTurk platform (through statistical analysis, and direct comparison with on-site rater and domain expert).

2 Materials and Methods

This section presents an overview of the PGR dataset [28], a brief presentation of the Amazon Mechanical Turk (MTurk) platform, and the integration of the dataset into the MTurk platform (including the design, configuration, and evaluation stages). We now describe how we proceeded with each of these stages:

1. Design
 - (a) Set up the tasks (Human Intelligence Task - HIT) to be simple to understand and easy to accomplish by the employees (i.e., workers or turkers).
 - (b) Define the guidelines (instructions) with examples for the workers to better understand the presented HITs.
2. Configuration
 - (a) Configure the MTurk platform specifying different criteria (for workers) and wage (i.e., reward).
 - (b) Submit the HITs within the platform.
3. Evaluation
 - (a) Calculate inter-rater agreement.
 - (b) Compare the PGR dataset before and after MTurk crowdsourcing assessment by employing two different deep learning models (BiOnt [27] and BioBERT [19])

An overview of the pipeline of the work described in this paper can be found in Figure 1.

2.1 PGR Dataset

The PGR dataset is a silver standard corpus of PubMed abstracts featuring human phenotype and gene annotations and their relations [28]. In this dataset, all the annotations were generated in a fully automated fashion (silver standard), taking a distant supervision approach, opposite to a manually annotated dataset where domain experts generate the annotations (gold standard).

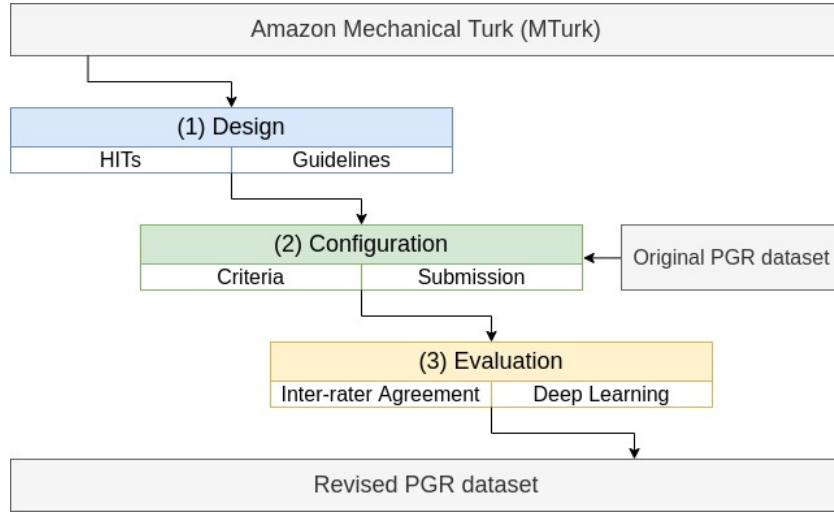


Fig. 1. The pipeline to incorporate the PGR dataset into the Amazon Mechanical Turk (MTurk) platform, including the design, configuration, and evaluation stages.

The first release of the PGR dataset focused mostly on the initial release of the dataset (10/12/2018), where a small subset of relations (6%) was manually reviewed to evaluate the PGR dataset quality and also to use as test corpus for machine learning model evaluation. The second release (11/03/2019) captured a more clear-cut search of the type of abstracts to retrieve, such as abstracts regarding diseases, their associated phenotypes and genes, increasing from about 2.5 relations per abstract to about 3.0 relations per abstract, and the overall number of relations by 2-fold. In this work, we are going to use the second release of the PGR dataset to generate an improved third release.

The relations identified in the PGR dataset are either Known if present in the knowledge base of relations provided by the Human Phenotype Ontology (HPO) group [16] or Unknown otherwise. Table 1 presents the numbers for the second release of the PGR dataset.

Table 1. The number of abstracts, phenotype and gene annotations, and of known, unknown and total of relations for the second release (11/03/2019) of the PGR dataset (partial table from [28]).

Abstracts	Annotations		Relations		
	Phenotype	Gene	Known	Unknown	Total
2657	9553	23786	2480	5483	7963

2.2 Amazon Mechanical Turk

The Amazon Mechanical Turk (MTurk) is a crowdsourcing web service (marketplace) that facilitates the use of human intelligence to individuals and businesses that are in demand to complete specific tasks [26]. In this web service, the employees (i.e., workers or turkers) execute tasks (i.e., HITs) submitted by employers (i.e., requesters) to earn a predefined wage (i.e., reward). The type of HITs that MTurk allows requesters to submit ranges from sentiment analysis and document classification in the language domain to image classification in the vision domain. Requesters post HITs to workers who meet their specified criteria (e.g., degree of education), and predefined both a reward and maximum time allotted for the completion of each task. Both requesters and workers remain anonymous throughout the process (workers can be identified through an internal identifier provided by Amazon).

The three main benefits of the MTurk platform are: (1) optimized efficiency by allowing requesters to outsource tasks that need to be handled manually, but do not require the requester or their employees' expertise; (2) increased flexibility for requesters to quickly scale their businesses without needing to scale their in-house workforce, and (3) cost reduction by eliminating the need for requesters to employ a temporary workforce and all the management costs associated with it [13].

Some previous works using MTurk in the biomedical field include named-entity recognition and curation of biomedical entities labels'. Yetisgen-Yildiz et al. [31] used MTurk to extract named-entities such as medical conditions, medication, and laboratory tests, from clinical trial descriptions. Good et al. [9] used it for disease mention annotation in PubMed abstracts. Similarly to our approach, Khare et al. [14] used MTurk to curate indications from drug labels, i.e. to judge whether a drug is used in managing a highlighted disease.

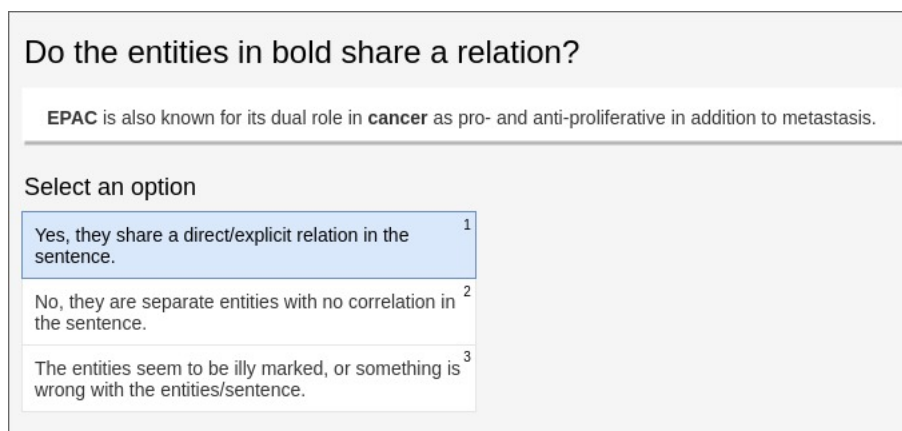
2.3 Integration into Amazon Mechanical Turk Platform

The MTurk platform provides a wide range of customizable templates to start a new project. The template closest to our previously described curation task was the Document Classification template, within the Language field, that we leveraged to set up our PGR HITs. To facilitate the evaluation of the workers performance, we divided the original dataset into partitions of 70% (Task 1) where each relation was rated by one Amazon worker and 30% (Task 2) where each relation was rated seven times, by seven distinct workers. We also had to define guidelines (instructions) with examples for the workers to understand the task at hand thoroughly. Further, each project required defining criteria to select the workers that better suited the goals of the project and determining the reward per HIT for each worker, before submission. Finally, after receiving the results (which took about two weeks), we had to evaluate the performance of our workers. The evaluation was done by calculating the inter-rater agreement, and by comparing the performance of the PGR dataset before and after curation with existing deep learning tools.

We describe the detailed steps that we took and reasoning for each decision made in the following sections.

Design

HITs As stated previously, we adapted the Document Classification template to set up our HITs. Thus, the workers were presented with a sentence with two entities in bold (the human phenotype and the gene entities) and a set of three possible classifications (true relation, false relation, or wrongly labeled relations due to errors in the NER stage or wrong sentence format). Figure 2 represents an example of a HIT as presented to the workers (Task 2).



The screenshot shows a HIT interface with a light gray background. At the top, a question "Do the entities in bold share a relation?" is displayed in bold black text. Below this, a sentence is presented in a white box with a gray border: "EPAC is also known for its dual role in **cancer** as pro- and anti-proliferative in addition to metastasis." The word "cancer" is in bold. Underneath the sentence box, the instruction "Select an option" is shown. Below this instruction are three radio button options, each in a white box with a gray border. The first option, "Yes, they share a direct/explicit relation in the sentence.", is selected and highlighted with a blue background, and has a small "1" to its right. The second option, "No, they are separate entities with no correlation in the sentence.", has a small "2" to its right. The third option, "The entities seem to be illy marked, or something is wrong with the entities/sentence.", has a small "3" to its right.

Fig. 2. An example of a HIT presented to the workers, and the available options.

Guidelines In this work, we considered that rather than defining strict guidelines, it would be more intuitive to the workers to be presented with examples of instances and their gold labels (Supplementary Material Figure S1). Nonetheless, the primary goal of the task presented to the workers was: The goal is to choose among three possible options to classify the relation between a phenotype and a gene in each sentence. The guidelines presented to the workers are illustrated by Supplementary Material Figure S1. We opted out of more exhaustive guidelines to keep the task time manageable and more straightforward to understand.

Configuration

Criteria As we stated before, requesters can predefine specific criteria that the workers have to meet to be able to work on a task. However, specifying that criteria has an added cost per HIT that would make the total value for the

task too expensive, invalidating the use of the crowd (domain expertise would be about the same value). Therefore, the criteria chosen and cost of the crowdsourcing project described in this work are detailed in Table 2. The requirement that workers be "Masters" (high performing workers according to MTurk) adds \$0.001 to the Mechanical Turk Fee, but since the platform rounds it up to the cent, the total value is unaltered.

Table 2. Summary of the crowdsourcing tasks criteria and associated costs.

Setting	Task 1	Task 2
Reward per assignment (USD)	0.02	0.02
Mechanical Turk fee (USD)	0.01	0.01
Number of assignments per task	1	7
Minimum time per assignment	3s	3s
Require that Workers be Masters to do your tasks (high performing workers according to MTurk)	Yes	Yes
Number of tasks	5574	2389
Total cost (USD)	167.22	501.69

Submission We designed a web page template for the tasks and defined the project properties, as required by the MTurk platform. We provided the input instances as a CSV file, where each line corresponded to a HIT. Alternatively, platforms such as Figure Eight [14] simplify task specification and monitoring of MTurk tasks. However, we worked directly with the MTurk platform.

Evaluation

Inter-rater Agreement The original dataset was divided into 70% where each relation was rated by one Amazon worker and 30% where each relation was rated seven times, by seven distinct workers. The goal of rating a subset of relations with overlap (Task 2) was to assess if the raters agreed with each other about the exact rating to be attributed (among the three previously described), by measuring the inter-rater agreement. To determine the previous metric, we used both the Fleiss' kappa [22] and the Krippendorff's alpha [17] metrics, that are appropriate for nominal ratings. The Fleiss's kappa metric is a statistical measure that estimates the reliability of agreement between a fixed number of raters, assuming that our raters were chosen at random from a larger population. Similarly, Krippendorff's alpha is a statistical measure of the agreement, useful when we have multiple raters and multiple possible ratings. We opted by using the two metrics to validate our work. A low deviation between the two metrics will assure an unbiased estimate [32]. Furthermore, we added an additional rater

from our research centre with no previous curating experience, but with a strong background in Biochemistry, to rate the overlapping subset of relations. This additional rater was fundamental to understand the challenges that our workers faced, and to help improve our curation pipeline and guidelines in the future.

To reach a majority consensus between the workers (for Task 2), we used a voting scheme, similar to the approach of Shu Li et al. [20]. Figure 3 illustrates how we chose to classify a relation true, false, or be excluded, according to the voting scheme. We considered that if at least half of the answers voted to exclude the relation from the dataset, the relation should be excluded. Our default label was false because we considered that false relations are more challenging to assess; hence, if a worker is in doubt between true and false, the most likely label would be false. For example, if on one HIT 5 out of 8 raters agreed to exclude, we accepted that rating. However, if 5 agreed true or false, we classified it as false, since considering it a valid sentence (not to exclude), with no agreement, our default label is false.

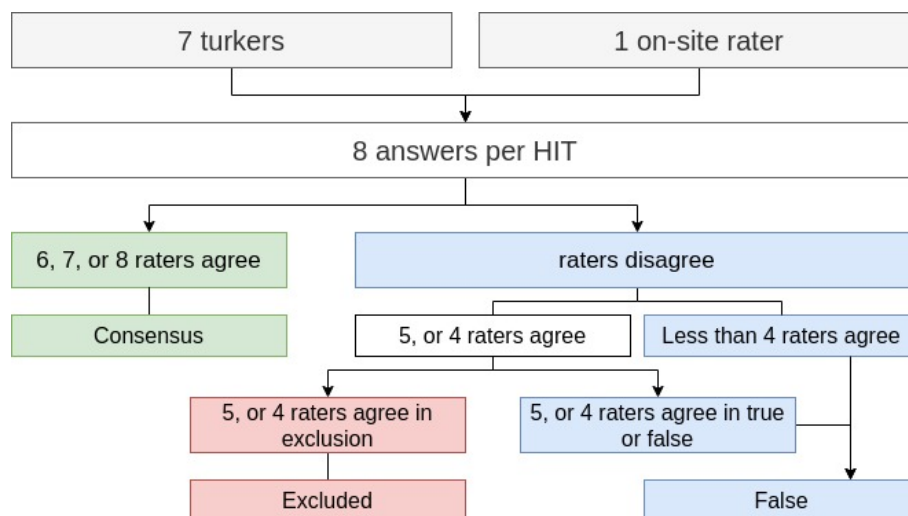


Fig. 3. Flowchart illustrating how to reach majority consensus, according to the answers provided by the workers plus our extra rater on site.

To further assess the quality and challenges of curating the PGR dataset and validate the previous approach, a domain expert with a Bioinformatics background and experience in using and curating corpora also curated the relations in Task 2.

Deep Learning Systems To further assess the quality of the crowdsourced curated dataset, we applied it to two distinct deep learning systems that target the biomedical domain: BiOnt [27] and BioBERT [19]. For comparison, we tested

both the original PGR dataset and the crowdsourced Amazon dataset, as well as combinations between the two (detailed in Table 5).

The BiOnt system is a deep learning system based on the BO-LSTM system [18] that is used to extract and classify relations via long short-term memory networks and biomedical ontologies. This system detects and classifies ten types of biomedical relations, such as human phenotype-gene relations. It takes advantage of domain-specific ontologies, like the Human Phenotype Ontology (HPO) [16] and the Gene Ontology (GO) [2]. The BiOnt system represents each entity as the sequence of its ancestors in their respective ontology.

The BioBERT system is a pre-trained biomedical language representation model for biomedical text mining based on the BERT [6] architecture. This system can perform diverse biomedical text mining tasks, namely NER, RE, and Question Answering (QA), when trained on large-scale biomedical corpora. The novelty of the architecture is that their authors designed these systems (BioBERT and BERT) to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. This feature allows easy adaption to several tasks without loss in performance.

3 Results and Discussion

3.1 Ratings Statistics

To assess the performance of the workers, we conducted some statistical analyses, including the time spent on average rating each sentence. Figure 4 and 5 reflect the average time spent by the workers with each sentence, with a cutoff of 50 seconds (using box plot and standard deviation analysis). We decided to set the cutoff for work time to 50 seconds because we considered that was enough time for a worker to make an assessment, and anything longer than that was probably the worker having a mid-task break (the longest time for a HIT completion was 40322 seconds, about 11 hours).

Our domain expert did a similar time self-evaluation, which resulted in an average of about 20 seconds per sentence (for Task 2). The domain expert consulted some abstracts to clarify whether an abbreviation was referring to a gene or other type of entity for a specific sentence. Through Figures 4 and 5 it is possible to assess that workers took an average of 13 seconds per HIT (sentence). By comparing this time to the average time done by our domain expert (20 seconds), it is possible to question the level of attention with which our workers performed their ratings, questioning the trust that we can deposit on MTurk crowdsourcing. However, taking into consideration that our domain expert took some time checking some abstracts to which workers did not have access, it can justify the differences in average time.

To further characterize the workers that performed our tasks, we checked their WorkerId tab in the results file provided by MTurk. There, we realized that six sentences were rated but did not have an associated WorkerId and that both tasks (7983 relations, 22255 HITs) were performed by only 64 different

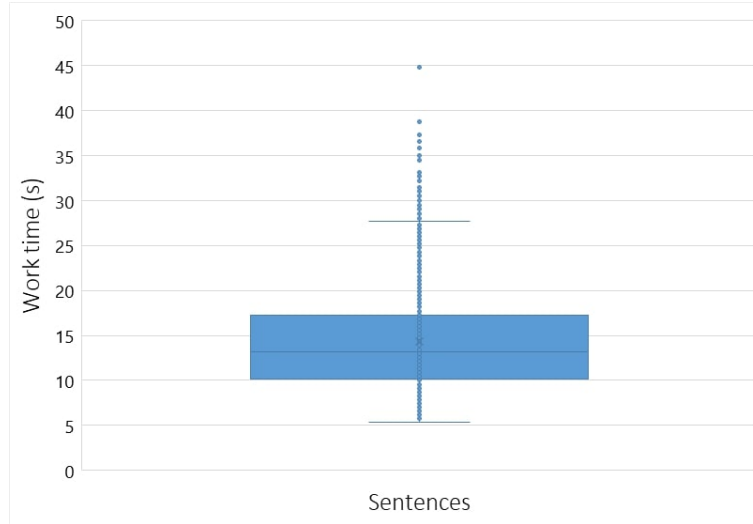


Fig. 4. Box plot expressing the average worker work time distribution (in seconds) per sentence (with a cutoff of 50 seconds).

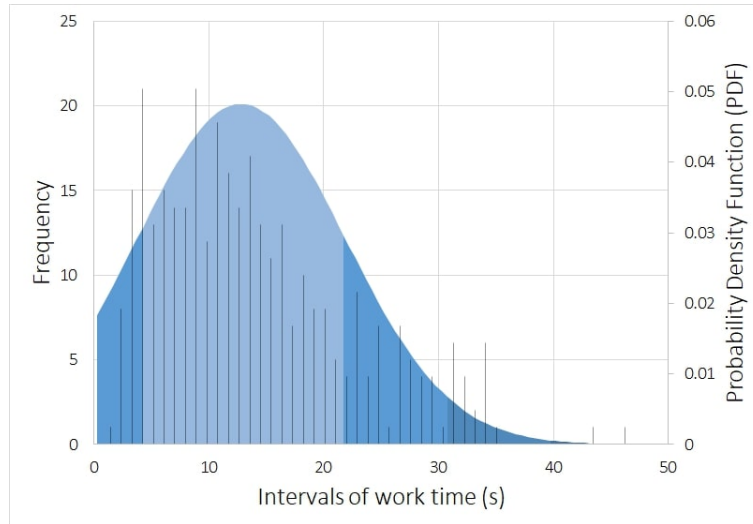


Fig. 5. Standard deviation expressing the average worker work time distribution (in seconds), and the histogram of the occurrence events (with a cutoff of 50 seconds).

workers, making in on average 348 HITs per worker. Therefore, if we had a malicious worker that classified their respective HITs at random or close to it, it would damage the whole dataset. The MTurk platform should guarantee a more diverse group of workers working on the same task since that is what employees are expecting, even to avoid some bias ratings, or a more strict selection process.

3.2 Inter-rater Agreement

Table 3 presents the inter-rater agreement score, using both Fleiss’ kappa [22] and Krippendorff’s alpha [17] metrics, for the dataset corresponding to Task 2, considering only the Amazon workers, the Amazon workers plus the extra rater (on-site), and the extra rater (on-site) plus the domain expert.

Table 3. The inter-rater agreement score, using both Fleiss’ kappa and Krippendorff’s alpha metrics, considering only the Amazon workers, the Amazon workers plus the extra rater (on-site), and the extra rater (on-site) plus the domain expert (Task 2).

Inter-rater agreement metric	Inter-rater agreement		
	Amazon workers	Amazon workers + extra rater (on-site)	Extra rater (on-site) + expert
Fleiss’ kappa	0.2028	0.2050	0.6549
Krippendorff’s alpha	0.2029	0.2051	0.6550

Given the small number of workers working on Task 2 (33), and the high number of sentences to rate (2389) it is challenging to find an inter-rater agreement metric that can return an accurate value of agreement between the workers. The Fleiss’ kappa metric assumes that the raters are deliberately chosen and fixed, while the Krippendorff’s alpha metric is indicated for when we have multiple raters and multiple possible ratings. Since none of the two cases is precisely right, we do not have a metric that fully expresses the results of our experiment with Task 2. We can say that probably the agreement between raters was only moderate (on a qualitative scale). Some of the reasons for moderate agreement could be due to difficulties in understanding the task, complex biomedical sentences that are beyond the scope of the average worker, or random answers provided by malicious workers.

It was particularly interesting to have an extra rater (on-site) that could express doubts while performing the task. Some of these doubts could be the ones that the workers had, while others we considered to be beyond their expertise. For our on-site rater, one of the most prominent problems was if the gene entities tagged were, in fact, gene entities or their proteins products, that frequently share the same names. One could argue that a relation between a gene product and a human phenotype implies a relation between a gene and a human phenotype.

Nonetheless, the extra rater considered that these relations hold even if the mention was of a protein and not the gene if this distinction was not clear by the sentence (only when reading the abstract or full-text article) or if the gene name was not capitalized. This particular problem was not one that a person not familiarized with Biochemistry related domains would have. However, assessing if an abbreviation that is used both as a gene name and in other biomedical topics (e.g., disease abbreviation) is a gene, is a transversal problem, to both the workers and our extra rater on-site.

The difficulties that our extra-rater experienced are evident by the inter-rater agreement between this rater and our domain expert. One example sentence where they disagreed was:

*While examining pedigrees of JEB patients with **LAMA3** mutations, we observed that heterozygous carriers of functional null mutations displayed subtle enamel pitting in the absence of **skin fragility** or other JEB symptoms.* (PMID:27827380)

where the domain expert considered a true relation, and the extra-rater a false relation, this happens because the relation is one of negation (absence) which often confuses non-experts with being false. However, an implication of relation of any sort is a true relation which can be then classified as positive or negative. This confusion is also noticeable by the diversity in the workers' answers for this sentence (four classified as true, two as false, and one as an error).

3.3 Corpus Statistics

Table 4 presents the final numbers both in total count and percentage for each task. For Task 2, we considered the majority consensus described previously, and the domain expert numbers separately.

From analyzing Table 4, what becomes immediately evident is the inversion between the number of true and false relations from the original datasets to the Amazon crowdsourced datasets. These final numbers demonstrate quite clearly that most relations described in the original PGR dataset as false were, in fact, true. This inversion can be due to the way that the PGR dataset was built, using a gold standard knowledge base of human phenotype-gene relations. This knowledge base, at the time of the dataset creation, was quite incomplete, since, for instance, if a child ontological term had a relation with a gene, its immediate parent would not necessarily share the same relation, which should be explicit. Thus, these parent concepts in PGR relations would always hold false.

Also, to understand the difference between an annotation error and a false relation requires more expertise than the one that MTurk provides, and that inexperienced raters have (even if in the field). Thus, to differentiate between false relations and an annotation error, we need expert knowledge, such as in the following annotation error example:

*We show that the **miR-106b-25** cluster upregulates **NOTCH1** in multiple breast **cancer** cell lines, representing both estrogen receptor (ER+) and triple*

Table 4. The inter-rater agreement score, using both Fleiss’ kappa and Krippendorff’s alpha metrics, considering only the Amazon workers, the Amazon workers plus the extra rater (on-site), and the extra rater (on-site) plus the domain expert (Task 2).

Dataset		Relations			
		True	False	Excluded	Total
Task 1 (70%)	Original	1751 (31.41%)	3823 (68.59%)	-	5574 (100%)
	Amazon workers	4220 (75.71%)	283 (5.08%)	1071 (19.21%)	4503 (80.79%)
Task 2 (30%)	Original	729 (30.51%)	1660 (69.49%)	-	2389 (100%)
	Amazon workers + extra rater (on-site) (after reaching consensus)	41179 (49.35%)	613 (25.66%)	240 (10.05%)	1792 (75.01%)
	Expert	1281 (53.62%)	343 (14.36%)	765 (32.02%)	1624 (67.98%)

negative breast cancer (TNBC) through direct repression of the E3 ubiquitin ligase, NEDD4L. (PMID:29662198)

where the workers had difficulties accessing that miR just by itself is not a gene entity, but stands for microRNA genes (a large group of genes).

3.4 Deep Learning Impact

Table 5 presents the performance of both the original PGR dataset and the crowdsourced Amazon dataset, and combinations between the two, on the BiOnt [27] and BioBERT (version 1.1) [19] systems, in terms of precision, recall, F-measure, and accuracy. To assess the performance of the dataset (before and after crowdsourcing) when applied to deep learning systems, we used the suggested parameters by the authors of each system. The only exception to the default parameters, since we had a class imbalance, was to add a class weight of 5 to the label false to both systems (the full multiplier to balance was approximately 14.9 for the Task 1 dataset). The full multiplier is a result of dividing the percentage of true relations by the percentage of false relations for the training dataset. For the class weight, we chose a number between 1 and the full multiplier which is usually the standard practice [4], to maintain a more accurate representation of the natural unbalance between labels when applying the models to real-world data. Using this class weight translates to treating every training instance with the label false as five instances of the label true, meaning that in the loss function, we assign a higher value to these instances. Hence, the loss becomes a weighted

average, where the weight of each sample is specified by the class weight and its corresponding class, providing a weight or bias for each output class. To achieve this, we had to alter the loss function of the BioBERT system to allow class weights.

Table 5. Precision, recall, F-measure, and accuracy of the application of the PGR dataset (original, new, and combinations between the two) to the BiOnt and BioBERT systems. The highest scores for each metric are presented in bold.

Method		Precision	Recall	F-measure	Accuracy
BiOnt	PGR	0.8140	0.3070	0.4459	0.4821
	Amazon (train) + PGR (test)	0.7000	0.9825	0.8175	0.7024
	Amazon (train) + Amazon workers consensus (test)	0.6810	0.9670	0.7992	0.6726
	Amazon (train) + Expert (test)	0.8142	0.9721	0.8861	0.7989
	Amazon workers consensus (train) + PGR (test)	0.6880	0.8509	0.7608	0.6369
	Expert (train) + PGR (test)	0.6894	0.9737	0.8072	0.6845
BioBERT	PGR	0.8542	0.3445	0.4910	0.5143
	Amazon (train) + PGR (test)	0.6744	0.9856	0.8000	0.6775
	Amazon (train) + Amazon workers consensus (test)	0.6700	0.9763	0.7946	0.6680
	Amazon (train) + Expert (test)	0.8103	0.9906	0.8915	0.8096
	Amazon workers consensus (train) + PGR (test)	0.7315	0.9160	0.8134	0.7143
	Expert (train) + PGR (test)	0.7857	0.8319	0.8082	0.7314

The deep learning systems performance is quite similar, with BioBERT achieving slightly better results. In both systems, the performance of the new PGR dataset (through MTurk crowdsourcing) was superior to the one of the original PGR dataset, with a slight decrease in precision but a considerable gain in recall. We chose to include the accuracy metric to consider the ability to recognize true negatives as well (due to the class imbalance). The best performance overall was

the Amazon MTurk (Task 1) as training corpus and the expert (Task 2) as test corpus. This performance can be due to the amount of available training data in Task 1, and the more reliable test set from the domain expert. The PGR original test set underperformed probably due to its small size which was not being representative of the data (260 relations). Also, other experiences with using the majority consensus (Task 2) and the expert (Task 2) as training sets showed that these smaller corpora also hold the ability to train a model.

4 Conclusion and Future Directions

This work describes our proposal of a complete pipeline for RE crowdsourcing. The pipeline generated an openly available new release of the PGR dataset, as well as a domain expert revision into 30% of the original dataset. Additionally, we assessed MTurk workers performance, by comparing them to an extra rater on-site and to a domain expert. Moreover, we applied the new dataset as training data in two state-of-the-art deep learning systems (BiOnt [27] and BioBERT [19]) to measure the usefulness of the annotations. This study showed that it is possible to use the wisdom of the crowd to at least improve existing silver standard datasets, since in our case it was able to exclude previous annotation errors (16.46%) and modify wrongly labelled relations. This improvement had a significant impact on model training, since we had 0.3494 average increase in F-measure, taking into account all the experiences when comparing it with the original PGR dataset.

Regarding future work, it will be interesting to improve on the existing pipeline by providing different guidelines and assess if that would make a difference in performance. Also, we can differentiate between what constitutes a false, and a negative relation. To solve the lack of domain expertise of MTurk workers, we could create a specialized crowdsourcing platform for the RE biomedical field, similar to the one developed by the company Unbabel that focus on translation [11], as well as other biomedical crowdsourcing projects [29,15]. Finally, we could apply the same methods to datasets from other biomedical domains and assess the differences in performance.

5 Acknowledgements

The authors express their gratitude to Priberam for facilitating the use of the platform Amazon Mechanical Turk. Also, we acknowledge the help of André Nascimento, as our extra on-site rater.

6 Supplementary Material

Relation Extraction Instructions

Option 1 Examples (Yes, they share a direct/explicit relation in the sentence.):

"In fact, **WRAP53** has been considered as a candidate **cancer** susceptibility gene."

"We further review the variety of melanocytic **tumors** associated with such **BRAF** fusions."

"**FLNB**-related disorders are classified as spondylocarpotarsal synostosis (SCT), Larsen syndrome (LS), atelosteogenesis (AO), boomerang dysplasia (BD), and isolated congenital talipes equinovarus, presenting with scoliosis, short-limbed dwarfism, clubfoot, joint dislocation and other unique **skeletal abnormalities**."

Option 2 Examples (No, they are separate entities with no correlation in the sentence.):

"Here, we report that these defects can occur independently of **albinism** in people with recessive mutations in the putative glutamine transporter gene **SLC38A8**."

"In particular, MYH9 mutations result in congenital macrothrombocytopenia and predispose to kidney failure, **hearing loss**, and cataracts, MPL and MECOM mutations cause congenital thrombocytopenia evolving into bone marrow failure, whereas thrombocytopenias caused by RUNX1, ANKRD26, and **ETV6** mutations are characterized by predisposition to hematological malignancies."

"The created databases include ACAD8 (isobutyryl-CoA dehydrogenase deficiency (IBD)), ACADSB (short-chain acyl-CoA dehydrogenase (SCAD) deficiency), AUH (3-methylglutaconic aciduria (3-MGCA)), DHCR7 (Smith-Lemli-Opitz syndrome), **HMGCS2** (3-hydroxy-3-methylglutaryl-CoA synthase 2 deficiency), HSD17B10 (17-beta-hydroxysteroid dehydrogenase X deficiency), FKBP14 (Ehlers-Danlos syndrome with progressive **kyphoscoliosis**, myopathy, and hearing loss; EDSKMH) and ROGDI (Kohlschutter-Tonz syndrome)."

Option 3 Examples (The entities seem to be illy marked, or something is wrong with the entities/sentence.):

"Here, we identify new patients with **dextrocardia** who have mutations in CFAP53, a coiled-coil domain containing protein." - **WRONG ANNOTATION:** *coil* is not a gene.

"The human Shwachman-Diamond syndrome (**SDS**) is an **autosomal recessive** disease caused by mutations in a highly conserved ribosome assembly factor SBDS." - **WRONG ANNOTATION:** *SDS* is not a gene.

"Barter syndrome (BS) is a hereditary condition transmitted as an **autosomal recessive** (Barter **type** 1 to 4) or dominant trait (Barter type 5)." - **WRONG ANNOTATION:** *type* is not a gene.

Fig. S1. The guidelines, in the form of examples of answers to different annotations, presented to the workers.

References

1. Arnaboldi, V., Raciti, D., Van Auken, K., Chan, J.N., Müller, H.M., Sternberg, P.W.: Text mining meets community curation: a newly designed curation platform to improve author experience and participation at wormbase. Database **2020** (2020)
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25–29 (2000)

3. Callison-Burch, C., Dredze, M.: Creating speech and language data with amazon's mechanical turk. In: *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*. pp. 1–12 (2010)
4. Chen, P.R., Lo, S.Y., Hang, H.M., Chan, S.W., Lin, J.J.: Efficient road lane marking detection with deep learning. In: *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*. pp. 1–5. IEEE (2018)
5. Collovini, S., Pereira, B., dos Santos, H.D., Vieira, R.: Annotating relations between named entities with crowdsourcing. In: *International Conference on Applications of Natural Language to Information Systems*. pp. 290–297. Springer (2018)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. pp. 4171–4186 (2019)
7. Feyisetan, O., Luczak-Roesch, M., Simperl, E., Tinati, R., Shadbolt, N.: Towards hybrid ner: A study of content and crowdsourcing-related performance factors. In: *European Semantic Web Conference*. pp. 525–540. Springer (2015)
8. Fort, K., Adda, G., Cohen, K.B.: Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics* **37**(2), 413–420 (2011)
9. Good, B.M., Nanis, M., Wu, C., Su, A.I.: Microtask crowdsourcing for disease mention annotation in pubmed abstracts. In: *Pacific Symposium on Biocomputing Co-Chairs*. pp. 282–293. World Scientific (2014)
10. Gormley, M.R., Gerber, A., Harper, M., Dredze, M.: Non-expert correction of automatically generated relation annotations. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. pp. 204–207 (2010)
11. Graça, J.: Unbabel: How to combine ai with the crowd to scale professional-quality translation. In: *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*. pp. 41–85 (2018)
12. Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., Declerck, T.: The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics* **46**(5), 914–920 (2013)
13. Ipeirotis, P.G., Provost, F., Wang, J.: Quality management on amazon mechanical turk. In: *Proceedings of the ACM SIGKDD workshop on human computation*. pp. 64–67 (2010)
14. Khare, R., Burger, J.D., Aberdeen, J.S., Tresner-Kirsch, D.W., Corrales, T.J., Hirschman, L., Lu, Z.: Scaling drug indication curation through crowdsourcing. *Database* **2015** (2015)
15. Kleffner, R., Flatten, J., Leaver-Fay, A., Baker, D., Siegel, J.B., Khatib, F., Cooper, S.: Foldit standalone: a video game-derived protein structure manipulation interface using rosetta. *Bioinformatics* **33**(17), 2765–2767 (2017)
16. Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurphy, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., et al.: The human phenotype ontology in 2017. *Nucleic acids research* **45**(D1), D865–D876 (2017)
17. Krippendorff, K.: *Computing krippendorff's alpha-reliability* (2011)
18. Lamurias, A., Sousa, D., Clarke, L.A., Couto, F.M.: BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies. *BMC bioinformatics* **20**(1), 1–12 (2019)
19. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**(4), 1234–1240 (2020)

20. Li, T.S., Bravo, À., Furlong, L.I., Good, B.M., Su, A.I.: A crowdsourcing workflow for extracting chemical-induced disease relations from free text. *Database* **2016** (2016)
21. Liu, A., Soderland, S., Bragg, J., Lin, C.H., Ling, X., Weld, D.S.: Effective crowd annotation for relation extraction. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 897–906 (2016)
22. McHugh, M.L.: Interrater reliability: the kappa statistic. *Biochemia medica* **22**(3), 276–282 (2012)
23. Mortensen, K., Hughes, T.L.: Comparing amazon’s mechanical turk platform to conventional data collection methods in the health and medical research literature. *Journal of General Internal Medicine* **33**(4), 533–538 (2018)
24. Narayan, S., Cohen, S.B., Lapata, M.: Ranking sentences for extractive summarization with reinforcement learning. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. pp. 1747–1759 (2018)
25. Paolacci, G., Chandler, J.: Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science* **23**(3), 184–188 (2014)
26. Paolacci, G., Chandler, J., Ipeirotis, P.G.: Running experiments on amazon mechanical turk. *Judgment and Decision making* **5**(5), 411–419 (2010)
27. Sousa, D., Couto, F.M.: BiOnt: deep learning using multiple biomedical ontologies for relation extraction. In: *European Conference on Information Retrieval*. pp. 367–374. Springer (2020)
28. Sousa, D., Lamurias, A., Couto, F.M.: A silver standard corpus of human phenotype-gene relations. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. pp. 1487–1492 (2019)
29. Tsueng, G., Nanis, M., Fouquier, J.T., Mayers, M., Good, B.M., Su, A.I.: Applying citizen science to gene, drug and disease relationship extraction from biomedical abstracts. *Bioinformatics* **36**(4), 1226–1233 (2020)
30. Wang, A., Hoang, C.D.V., Kan, M.Y.: Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation* **47**(1), 9–31 (2013)
31. Yetisgen-Yildiz, M., Solti, I., Xia, F., Halgrim, S.: Preliminary experiments with amazon’s mechanical turk for annotating medical named entities. In: *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon’s Mechanical Turk*. pp. 180–183 (2010)
32. Zapf, A., Castell, S., Morawietz, L., Karch, A.: Measuring inter-rater reliability for nominal data—which coefficients and confidence intervals are appropriate? *BMC medical research methodology* **16**(1), 93 (2016)