Achieving Expert-Level Annotation Quality with CrowdTruth

The Case of Medical Relation Extraction

Anca Dumitrache[1,2], Lora Aroyo[1], and Chris Welty[3]

[1] VU University Amsterdam, Netherlands
{anca.dumitrache,lora.aroyo}@vu.nl
[2] IBM CAS, Amsterdam, Netherlands
[3] Google Research, New York, USA
cawelty@gmail.com

Abstract. The lack of annotated datasets for training and benchmarking is one of the main challenges of Clinical Natural Language Processing. In addition, current methods for collecting annotation attempt to minimize disagreement between annotators, and therefore fail to model the ambiguity inherent in language. We propose the CrowdTruth method for collecting medical ground truth through crowdsourcing, based on the observation that disagreement between annotators can be used to capture ambiguity in text. In this work, we report on using this method to build a ground truth for medical relation extraction, and how it performed in training a classification model. Our results show that, with appropriate processing, the crowd performs just as well as medical experts in terms of the quality and efficacy of annotations. Furthermore, we show that the general practice of employing a small number of annotators for collecting ground truth is faulty, and that more annotators per sentence are needed to get the highest quality annotations.

1 Introduction

Clinical Natural Language Processing (NLP) has become an invaluable tool for navigating and processing medical data [19]. Clinical NLP relies on the development of a set of gold standard annotations, or *ground truth*, for the purpose of training, testing and evaluation. Ground truth is usually collected by humans reading text and following a set of guidelines to ensure a uniform understanding of the annotation task. In the medical domain, domain knowledge is usually believed to be required from annotators, making the process for acquiring ground truth more difficult. The lack of annotated datasets for training and benchmarking is considered one of the big challenges of Clinical NLP [8].

Furthermore, the process behind acquiring ground truth often presents flaws [5]. It is assumed that the gold standard represents a universal and reliable model for language. However, previous experiments we performed in medical relation extraction [2] identified two issues with this assumption: (1) disagreement between annotators is usually eliminated through overly prescriptive annotation

guidelines, thus creating artificial data that is neither general nor reflects the ambiguity inherent in natural language, and (2) the process of acquiring ground truth by working exclusively with domain experts is costly and non-scalable, both in terms of time and money.

A possible solution to these issues is using crowdsourcing for collecting the ground truth. Not only is this a much faster and cheaper procedure than expert annotation, it also allows for collecting enough annotations per task in order to represent the diversity inherent in language. Crowd workers, however, generally lack medical expertise, which might impact the quality and reliability of their work in more knowledge-intensive tasks. Previously, we studied medical relation extraction in a relatively small set of 90 sentences [3], comparing the results from the crowd with that of two expert medical annotators. We found that disagreement within the crowd is consistent with expert inter-annotator disagreement. Furthermore, sentences that registered high disagreement tended to be vague or ambiguous when manually evaluated.

Our approach, called *CrowdTruth*, can overcome the limitations of gathering expert ground truth, by using disagreement analysis on crowd annotations to model the ambiguity inherent in medical text. Furthermore, we claim that, even for complex annotation tasks such as relation extraction, lack of medical expertise of the crowd is compensated by collecting a large enough set of annotations. We prove this in two ways, by manually judging the quality of the annotations provided by experts and the crowd, and more importantly by training a model for medical relation extraction with both CrowdTruth data and ground truth from medical experts, and comparing them in a cross-validation experiment.

In this paper, we make the following contributions: (1) a comparison of the quality and efficacy of annotations for medical relation extraction provided by both crowd and medical experts, showing that *crowd annotations are equivalent to those of experts*, with appropriate processing; (2) an openly available *dataset of 900 English sentences for medical relation extraction*, centering primarily on the *cause* relation, that have been processed with disagreement analysis and by experts; (3) an analysis of the *optimal crowd settings for medical relation extraction*, showing that 10 workers per sentence yields the highest quality annotations.

2 Related Work

There exists some research using crowdsourcing to collect semantic data for the medical domain. [18] use crowdsourcing to verify relation hierarchies in biomedical ontologies. On 14 relations from the SNOMED CT CORE Problem List Subset, the authors report the crowd's accuracy at 85% for identifying whether the relations were correct or not. In the field of Biomedical NLP, [7] used crowdsourcing to extract the gene-mutation relations in Medical Literature Analysis and Retrieval System Online (MEDLINE) abstracts. Focusing on a very specific gene-mutation domain, the authors report a weighted accuracy of 82% over a

corpus of 250 MEDLINE abstracts. Both of these approaches present preliminary results from experiments performed with small datasets.

To our knowledge, the most extensive study of medical crowdsourcing was performed by [23], who describe a method for crowdsourcing a ground truth for medical named entity recognition and entity linking. In a dataset of over 1,000 clinical trials, the authors show no statistically significant difference between the crowd and expert-generated gold standard for the task of extracting medications and their attributes. We extend these results by applying crowdsourcing to the more complex task of medical relation extraction, that *prima facie* seems to require more domain expertise than named entity recognition. Furthermore, we test the viability of crowdsourced ground truth for relation extraction.

Crowdsourcing ground truth has shown promising results in a variety of other domains. [13] compared the crowd versus experts for the task of part-of-speech tagging. The authors also show that models trained based on crowdsourced annotation can perform just as well as expert-trained models. [15] studied crowdsourcing for relation extraction in the general domain, comparing its efficiency to that of fully automated information extraction approaches. Their results showed the crowd was especially suited to identifying subtle formulations of relations that do not appear frequently enough to be picked up by statistical methods.

Other research for crowdsourcing ground truth includes: entity clustering and disambiguation [16], Twitter entity extraction [12], multilingual entity extraction and paraphrasing [9], and taxonomy creation [10]. However, all of these approaches rely on the assumption that one black-and-white gold standard must exist for every task. Disagreement between annotators is discarded by picking one answer that reflects some consensus, usually through using majority vote. The number of annotators per task is also kept low, between two and five workers, also in the interest of eliminating disagreement. The novelty in our approach is to consider language ambiguity, and consequently inter-annotator disagreement, as an inherent feature of the language. The metrics we employ for determining the quality of crowd answers are specifically tailored to quantify disagreement between annotators, rather than eliminate it.

3 Experimental Setup

In order to perform the comparison between expert and crowdsourced gold standards, we set up an experiment to train and evaluate a relation extraction model for a sentence-level relation classifier. The classifier takes, as input, sentences and two terms from the sentence, and returns a score reflecting the likelihood that a specific relation, in our case the *cause* relation between symptoms and disorders, is expressed in the sentence between the terms. Starting from a set of 902 sentences that are likely to contain medical relations, we constructed a workflow for collecting annotations through crowdsourcing. This output was analyzed with CrowdTruth metrics for capturing disagreement, and then used to train a model for relation extraction. In parallel, we also constructed a model based using a

traditional gold standard acquired from domain experts, that we then compare to the crowd model.

3.1 Data

The dataset used in our experiments contains 902 medical sentences extracted from PubMed article abstracts. The MetaMap parser [1] ran over the corpus to identify medical terms from the UMLS vocabulary [6]. Distant supervision [17] was used to select sentences with pairs of terms that are linked *in UMLS* by one of our chosen seed medical relations. The intuition of distant supervision is that since we know the terms are related, and they are in the same sentence, it is more likely that the sentence expresses a relation between them (than just any random sentence). The seed relations were restricted to a set of eleven UMLS relations important for clinical decision making [22] (listed in Tab.1). Given a relation, each sentence in the dataset can be either *positive* (i.e. the relation is expressed between the two terms in the sentence), or *negative* (i.e. the relation is not expressed). All of the data that we have used is available online at: `http://data.crowdtruth.org/medical-relex`.

For collecting annotations from medical experts, we employed medical students, in their third year at American universities, that had just taken United States Medical Licensing Examination (USMLE) and were waiting for their results. Each sentence was annotated by exactly one person. The annotation task consisted of deciding whether or not the UMLS seed relation discovered by distant supervision is present in the sentence for the two selected terms.

3.2 Crowdsourcing setup

The crowdsourced annotation setup is based on our previous medical relation extraction work [4], adapted into a workflow of three tasks (Fig.1). First, the sentences were pre-processed using a named-entity recognition tool combining the UMLS vocabulary with lexical parsing, to determine whether the terms found with distant supervision are complete or not. The incomplete terms were parsed through a crowdsourcing task (*FactSpan*) in order to get the full word span of the medical terms. Next, the sentences with the corrected term spans were sent to a relation extraction task (*RelEx*), where the crowd was asked to decide which relation holds between the two extracted terms. To simplify the task for the crowd, we combined the UMLS relations from distant supervision, merging relations with similar meanings (e.g. *disease has primary anatomic site* and *has finding site*). We also added four new relations (e.g. *associated with*), to account for weaker, more general links between the terms. The full set of the relations presented to the crowd is available in Tab.1. The workers were also able to read the definition of each relation. The task was multiple choice, workers being able to choose more than one relation at the same time. There were also options available for cases when the medical relation was other than the ones we provided (*other*), and for when there was no relation between the terms (*none*). Finally, the results from *RelEx* were passed to another crowdsourcing

Table 1: Set of medical relations.

Relation	**Corresponding UMLS relation(s)**	**Definition**	**Example**
treat	may treat	therapeutic use of a drug	penicillin treats infection
prevent	may prevent	preventative use of a drug	vitamin C prevents influenza
diagnosis	may diagnose	diagnostic use of an ingredient, test or a drug	RINNE test is used to diagnose hearing loss
cause	cause of; has causative agent	the underlying reason for a symptom or a disease	fever induces dizziness
location	disease has primary anatomic site; has finding site	body part in which disease or disorder is observed	leukemia is found in the circulatory system
symptom	disease has finding; disease may have finding	deviation from normal function indicating the presence of disease or abnormality	pain is a symptom of a broken arm
manifestation	has manifestation	links disorders to the observations that are closely associated with them	abdominal distention is a manifestation of liver failure
contraindicate	contraindicated drug	a condition for which a drug or treatment should not be used	patients with obesity should avoid using danazol
associated with		signs, symptoms or findings that often appear together	patients who smoke often have yellow teeth
side effect		a secondary condition or symptom that results from a drug	use of antidepressants causes dryness in the eyes
is a		a relation that indicates that one of the terms is more specific variation of the other	migraine is a kind of headache
part of		an anatomical or structural sub-component	the left ventricle is part of the heart

task *RelDir* to determine the direction of the relation with regards to the two extracted terms.

All three crowdsourcing tasks were run on the CrowdFlower platform [4] with 10-15 workers per sentence, to allow for a distribution of perspectives; the precise settings for each task are available in Tab.2. Even with three tasks and 10-15 workers per sentence, compared to a single expert judgment per sentence, the total cost of the crowd amounted to 2/3 of the sum paid for the experts. In our case, cost was not the limiting factor for the experts, but their time and availability.

	FactSpan	**RelEx**	**RelDir**
judgments (i.e. workers per sentence)	10	15	10
pay per sentence annotation (in $)	0.04	0.05	0.01

Table 2: CrowdFlower Settings for the Tasks in CrowdTruth Workflow.

3.3 CrowdTruth metrics

For each crowdsourcing task in the crowd annotation workflow, the crowd output was processed with the use of CrowdTruth metrics – a set of general-purpose

[4] `http://CrowdFlower.com`

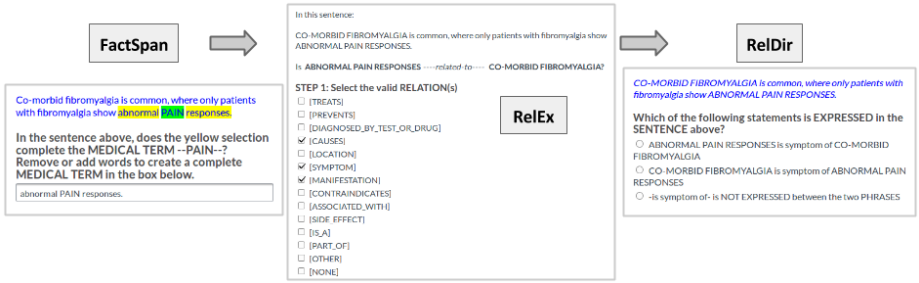

Fig. 1: CrowdTruth Workflow for Medical Relation Extraction on CrowdFlower [11].

crowdsourcing metrics [14], that have been successfully used to model relation extraction [4]. These metrics attempt to model the crowdsourcing process based on the triangle of reference [20], with the vertices being the input sentence, the worker, and the seed relation. Ambiguity and disagreement at any of the vertices (e.g. a sentence with unclear meaning, a poor quality worker, or an unclear relation) will propagate in the system, influencing the other components. For example, a worker who annotates an unclear sentence is more likely to disagree with the other workers, and this can impact that worker's quality. Therefore, the CrowdTruth metrics model quality at each vertex in relation to all the others, so that a high quality worker who annotates many low clarity sentences will be recognized as high quality. In our workflow, these metrics are used both to eliminate spammers [21], and to determine the clarity of the sentences and relations. The main concepts are:

- *annotation vector:* This construct is used to model the annotation of one worker for one sentence. For each worker i submitting their solution to a task on a sentence s, the vector $W_{s,i}$ records their answers. If the worker selects an answer, its corresponding component would be marked with '1', and '0' otherwise. For instance, in the case of $RelEx$, the vector will have fourteen components, one for each relation, as well as *none* and *other*.
- *sentence vector:* This is the main component for modeling disagreement in the crowdsourcing system. There is one such vector for every input sentence. For every sentence s, it is computed by adding the annotation vectors for all workers on the given task: $V_s = \sum_i W_{s,i}$.
- *sentence-annotation score:* A core CrowdTruth concept, this metric computes annotation ambiguity in a sentence with the use of cosine similarity. In the case of $RelEx$, it becomes the *sentence-relation score*, and is computed as the cosine similarity between the sentence vector and the unit vector for the relation: $srs(s, r) = cos(V_s, \hat{r})$. The higher the value of this metric, the more clearly the relation is expressed in the sentence.

3.4 Training the model

The sentences together with the relation annotations were then used to train a manifold model for relation extraction [22]. This model was developed for the medical domain, and tested for the relation set that we employ. It is trained per individual relation, by feeding it both *positive* and *negative* data. It offers support for both discrete labels, and real values for weighting the confidence of the training data entries, with positive values in $(0, 1]$, and negative values in $[-1, 0)$. Using this system, we train several models using five-fold cross validation, in order to compare performances of the crowd and expert dataset. In total, we use four datasets:

1. *baseline:* Discrete (positive or negative) labels are given for each sentence by the distant supervision method – for any relation, a positive example is a sentence containing two terms related by *cause* in UMLS. This dataset constitutes the baseline against which all other datasets are tested. Distant supervision does not extract negative examples, so in order to generate a negative set for one relation, we use positive examples for the other (non-overlapping) relations shown in Tab. 1.
2. *expert:* Discrete labels based on an expert's judgment as to whether the *baseline* label is correct. The experts do not generate judgments for all combinations of sentences and relations – for each sentence, the annotator decides on the seed relation extracted with distant supervision. Similarly to the baseline data, we reuse positive examples from the other relations to extend the number of negative examples.
3. *single:* Discrete labels for every sentence are taken from one randomly selected crowd worker who annotated the sentence. This data simulates the traditional single annotator setting.
4. *crowd:* Weighted labels for every sentence are based on the CrowdTruth *sentence-relation score*. The classifier expects positive scores for positive examples, and negative scores for negative, so the sentence-relation scores must be re-scaled. An important variable in the re-scaling is a threshold to select positive and negative examples. The Results section compares the performance of the crowd at different threshold values. Given a threshold, the *sentence-relation score* is then linearly re-scaled into the $[0.85, 1]$ interval for the positive label weight, and the $[-1, -0.85]$ interval for negative. An example of how the scores were processed is given in Tab.3.

3.5 Evaluation setup

In order for a meaningful comparison between the crowd and expert models, the evaluation set needs to be carefully selected. The sentences in the test folds were picked through the cross validation mechanism, but the scores were selected from our *test partition*, which we verified to ensure correctness. To build the *test partition* , we first selected the positive/negative threshold for *sentence-relation score* such that the crowd agrees the most with the experts. We assume that, if both the expert and the crowd agree that a sentence is either a positive or

Sent.1: **Renal osteodystrophy** is a general complication of chronic renal failure and **end stage renal disease**.
Sent.2: If **TB** is a concern, a **PPD** is performed.

	treat	*prevent*	*diagnosis*	*cause*	*location*	*symptom*	*manifestation*	*contraindicate*	*associated with*	*side effect*	*is a*	*part of*	*other*	*none*	Sent.
sentence	0	0	1	10	1	2	0	0	1	0	0	0	0	0	**Sent.1**
vector	3	1	7	0	0	0	0	0	3	0	0	0	1	0	**Sent.2**
sentence –	0	0	0.09	0.96	0.09	0.19	0	0	0.09	0	0	0	0	0	**Sent.1**
relation score	0.36	0.12	0.84	0	0	0	0	0	0.36	0	0	0	0.12	0	**Sent.2**
crowd model	-1	-1	-0.97	0.99	-0.97	-0.94	-1	-1	-0.97	-1	-1	-1	-1	-1	**Sent.1**
training score	-0.89	-0.96	0.95	-1	-1	-1	-1	-1	-0.89	-1	-1	-1	-0.96	-1	**Sent.2**

Table 3: Example sentence with scores from the crowd dataset; training score calculated for negative/positive sentence-relation threshold equal to 0.5*, and linear rescaling in the* $[-1, -0.85]$ *interval for negative,* $[0.85, 1]$ *for positive.*

negative example, it can automatically be used as part of the test set. Such a sentence was labeled with the crowd score. In the cases where the crowd and experts disagree, we manually verified and assigned either a positive, negative, or ambiguous value. The ambiguous cases were subsequently removed from the test folds. In this way we created reliable, unbiased test scores, to be used in the evaluation of the models.

4 Results

We compared each of the four datasets using the test partition as a gold standard, to determine the quality of the *cause* relation annotations, as shown in Fig.2. As expected, the baseline data performed the lowest, followed closely by the single crowd worker. The expert annotations achieved an F1 score of 0.844. Since the baseline, expert, and single sets are binary decisions, they appear as horizontal lines. For the crowd annotations, we plotted the F1 against different sentence-

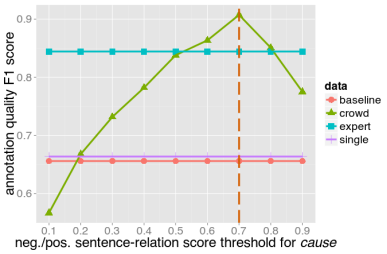

Fig. 2: Annotation quality F1 per neg./pos. threshold for cause.

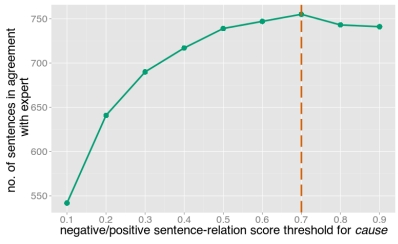

Fig. 3: Crowd & expert agreement per neg./pos. threshold for cause.

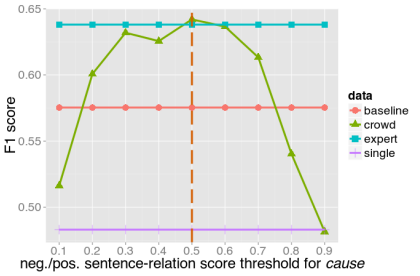

Fig. 4: F1 scores.

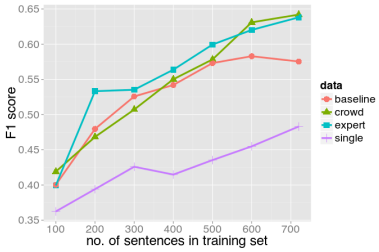

Fig. 5: Learning curves (crowd with pos./neg. threshold at 0.5).

relation score thresholds for determining positive and negative sentences. Between the thresholds of 0.6 and 0.8, the crowd out-performs the expert, reaching the maximum of 0.907 F1 score at a threshold of 0.7. This difference is significant with $p = 0.007$, measured with McNemar's test. In Fig.3 we show the number of sentences in which the crowd agrees with the expert (on both positive and negative decisions), plotted against different positive/negative thresholds for the sentence-relation score of *cause*. The maximum agreement with the expert set is at the 0.7 threshold, the same as for the annotation quality F1 score (Fig.2), with 755 sentences where crowd and expert agree.

We next wanted to verify that this improvement in annotation quality has a positive impact on the model that is trained with this data. In a cross-validation experiment, we trained the model with each of the four datasets for identifying the *cause* relation. The results of the evaluation (Fig.4) show the best performance for the crowd model when the sentence-relation threshold for deciding between negative/positive equals 0.5. Trained with this data, the classifier model achieves an F1 score of 0.642, compared to the expert-trained model which reaches 0.638. McNemar's test shows statistical significance with $p = 0.016$. This result demonstrates that the crowd provides training data that is at least as good, if not better than experts. In addition, the baseline scores an F1 of 0.575, and the single annotator shows the worst performance, scoring at 0.483. The learning curves (Fig.5) show that, above 400 sentences, the crowd consistently scores over baseline and single in F1 score. After 600 sentences, the crowd also out-performs the experts. The trend of the curve is still upward, indicating that more data is necessary to get the best performance.

Finally, we checked whether the number of workers per task was sufficient to produce a stable sentence-relation score. For the *RelEx* task, we ensured that each sentence was checked by at least 10 workers, after spam removal. The plot of the mean cosine distance between sentence vectors before and after adding the latest worker shows that the sentence vector becomes stable after 10 workers (Fig. 6). Furthermore, the annotation quality F1 score per total number of workers (Fig. 7) is also stable after 10 workers (the drop towards the end is due to sparse data – only 54 sentences had 15 or more total workers).

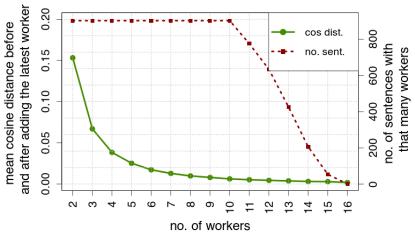

Fig. 6: Mean cosine distance for sentence vectors before and after adding the latest worker, shown per number of workers.

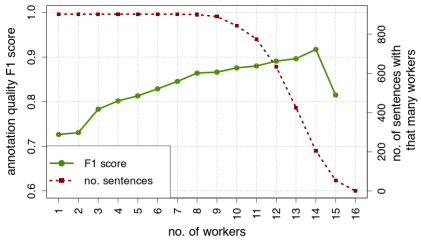

Fig. 7: Annotation quality F1 for crowd pos./neg. threshold at 0.7, shown per number of workers.

5 Discussion

Our goal was to demonstrate that, like the crowdsourced medical entity recognition work by Zhai et al. [23], the CrowdTruth approach of having multiple annotators with precise quality scores can be harnessed to create gold standard data with a quality that rivals annotated data created by medical experts. Our results show this clearly, in fact with slight improvements, with a sizable dataset (902 sentences) on a problem (relation extraction) that *prima facie* seems to require more domain expertise. Tab.4 shows the results in more detail.

The most curious aspect of the results is that the positive/negative sentence-relation score threshold that gives the best quality annotations (Fig.2) is different from the best threshold for training the model (Fig.4). It is the lower threshold (equal to 0.5) that gives a better classification. This is most likely due to the higher recall of the lower threshold, which exposes the classifier to more positive examples. F-score is the harmonic mean between precision and recall, and does not necessarily represent the best trade-off between them, as this experiment shows. Indeed F-score may not be the best trade-off between precision and recall for the classifier. In [11], we experimented with a weighted F-score, using the CrowdTruth metrics to account for ambiguity in the sentences. Using this new metric, we found an improved performance for both crowd and expert.

Table 4: Model evaluation results for each dataset.

	precision	**recall**	**F1 score**	**accuracy**	**max. F1 score**
crowd (0.5 sent.-rel. threshold)	0.565	0.743	0.642	0.784	0.659
crowd (0.7 sent.-rel. threshold)	0.619	0.61	0.613	0.8	0.654
expert	0.672	0.604	0.638	0.818	0.679
baseline	0.436	0.844	0.575	0.674	0.622
single	0.495	0.473	0.483	0.737	0.54

It is also notable that the baseline out-performs the single annotator. This could be an indicator that the crowd can only achieve quality when accounting for the choices of multiple annotators. In addition, the recall score for baseline is notably high. This could be a consequence of how the model performs its training – one of the features it learns is the UMLS type of the terms. For *cause*, term types are often enough to accurately qualify the relation.

The learning curves (Fig.5) show we still have not reached the ideal amount of training data, especially for the CrowdTruth approach, in which the weights of sentences have less of a cumulative effect, as opposed to datasets with binary labels. In other words, when accounting for ambiguity in training, more data points are needed to reach maximum performance. A bottleneck in this analysis is the availability of expert annotations – we did not have the resources to collect a larger expert dataset, and this indeed is the main reason to consider crowdsourcing. It is also worth noting that, while the crowd annotations consistently out-perform the distant-supervision baseline, we do not yet have a fair comparison between a distant supervision approach and the CrowdTruth approach. The real value of distant supervision is that large amounts of data can be gathered rather easily and cheaply, since humans are not involved. We are working on experiments to explore the trade-off between scale, quality, and cost, based on the assumption that systems trained with either kind of data will eventually reach a performance maximum.

Finally, in Figs. 6 & 7 we observe that we need at least 10 workers to get a stable crowd score. This result goes against the general practice for building a ground truth, where per task there usually are 2 to 5 annotators. Based on our results, we believe that the general practice is wrong, and that outside of a few clear cases, the input of more annotators is necessary to capture ambiguity. Even with this added requirement, we found that crowdsourcing is still cheaper than medical experts – the cost of the experts was 50% higher.

6 Conclusion

The lack of ground truth for training and benchmarking is one of the main challenges of Clinical NLP. In addition, current methods for collecting annotation attempt to minimize disagreement between annotators, but end up failing to model the ambiguity inherent in language. We propose the CrowdTruth method for crowdsourcing ground truth while also capturing and interpreting disagreement. We used CrowdTruth to build a gold standard of 902 sentences for medical relation extraction, which was employed in training a classification model. We have shown that, with appropriate processing, the crowd performs just as well as medical experts in terms of the quality and efficacy of annotations, while being cheaper and more readily available. Our results indicate that at least 10 workers per sentence are needed to get the highest quality annotations, in contrast to the general practice of employing a small number of annotators for collecting ground truth. We plan to continue our experiments by scaling out the crowdsourcing approach, which has the possibility of performing better.

Acknowledgments

The authors would like to thank Chang Wang for support with using the medical relation extraction classifier, and Anthony Levas for help with collecting the expert annotations.

References

1. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium. p. 17. AMIA (2001)
2. Aroyo, L., Welty, C.: Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. Web Science 2013. ACM (2013)
3. Aroyo, L., Welty, C.: Measuring crowd truth for medical relation extraction. In: AAAI 2013 Fall Symposium on Semantics for Big Data (2013)
4. Aroyo, L., Welty, C.: The Three Sides of CrowdTruth. Journal of Human Computation 1, 31–34 (2014)
5. Aroyo, L., Welty, C.: Truth is a lie: Crowd truth and the seven myths of human annotation. AI Magazine 36(1), 15–24 (2015)
6. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research 32(suppl 1), D267–D270 (2004)
7. Burger, J.D., Doughty, E., Bayer, S., Tresner-Kirsch, D., Wellner, B., Aberdeen, J., Lee, K., Kann, M.G., Hirschman, L.: Validating candidate gene-mutation relations in medline abstracts via crowdsourcing. In: Data Integration in the Life Sciences. pp. 83–91. Springer (2012)
8. Chapman, W.W., Nadkarni, P.M., Hirschman, L., D'Avolio, L.W., Savova, G.K., Uzuner, O.: Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the AMIA 18(5), 540–543 (2011)
9. Chen, D.L., Dolan, W.B.: Building a persistent workforce on mechanical turk for multilingual data collection. In: Proceedings of The 3rd HCOMP (2011)
10. Chilton, L.B., Little, G., Edge, D., Weld, D.S., Landay, J.A.: Cascade: crowdsourcing taxonomy creation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1999–2008. CHI '13, ACM, New York, NY, USA (2013)
11. Dumitrache, A., Aroyo, L., Welty, C.: CrowdTruth Measures for Language Ambiguity: The Case of Medical Relation Extraction. In: Proceedings of the 2015 International Workshop on Linked Data for Information Extraction (LD4IE-2015), 14th International Semantic Web Conference (2015)
12. Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., Dredze, M.: Annotating named entities in Twitter data with crowdsourcing. In: In Proc. NAACL HLT. pp. 80–88. CSLDAMT '10, ACL (2010)
13. Hovy, D., Plank, B., Søgaard, A.: Experiments with crowdsourced re-annotation of a POS tagging data set. In: Proceedings of the 52nd Annual Meeting of the ACL (Volume 2: Short Papers). pp. 377–382. ACL, Baltimore, Maryland (June 2014)
14. Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., van der Ploeg, J., Romaszko, L., Aroyo, L., Sips, R.J.: CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data. In: The Semantic Web–ISWC 2014, pp. 486–504. Springer (2014)

15. Kondreddi, S.K., Triantafillou, P., Weikum, G.: Combining information extraction and human computing for crowdsourced knowledge acquisition. In: 30th International Conference on Data Engineering. pp. 988–999. IEEE (2014)
16. Lee, J., Cho, H., Park, J.W., Cha, Y.r., Hwang, S.w., Nie, Z., Wen, J.R.: Hybrid entity clustering using crowds and data. The VLDB Journal 22(5), 711–726 (2013)
17. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2. pp. 1003–1011. ACL (2009)
18. Mortensen, J.M., Musen, M.A., Noy, N.F.: Crowdsourcing the verification of relationships in biomedical ontologies. In: AMIA Annual Symposium Proceedings. vol. 2013, p. 1020. AMIA (2013)
19. Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W.: Natural language processing: an introduction. Journal of the AMIA 18(5), 544–551 (2011)
20. Ogden, C.K., Richards, I.: The meaning of meaning. Trubner & Co, London (1923)
21. Soberón, G., Aroyo, L., Welty, C., Inel, O., Lin, H., Overmeen, M.: Measuring Crowd Truth: Disagreement Metrics Combined with Worker Behavior Filters. In: 1st International Workshop on Crowdsourcing the Semantic Web, 12th International Semantic Web Conference (2013)
22. Wang, C., Fan, J.: Medical relation extraction with manifold models. In: 52nd Annual Meeting of the ACL, vol. 1. pp. 828–838. ACL (2014)
23. Zhai, H., Lingren, T., Deleger, L., Li, Q., Kaiser, M., Stoutenborough, L., Solti, I.: Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. JMIR 15(4) (2013)