CrowdTruth Measures for Language Ambiguity
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. A widespread use of linked data for information extraction is *distant supervision*, in which relation tuples from a data source are found in sentences in a text corpus, and those sentences are treated as training data for relation extraction systems. Distant supervision is a cheap way to acquire training data, but that data can be quite noisy, which limits the performance of a system trained with it. Human annotators can be used to clean the data, but in some domains, such as medical NLP, it is widely believed that only medical experts can do this reliably. We have been investigating the use of crowdsourcing as an affordable alternative to using experts to clean noisy data, and have found that with the proper analysis, crowds can rival and even out-perform the precision and recall of experts, at a much lower cost. We have further found that the crowd, by virtue of its diversity, can help us find evidence of ambiguous sentences that are difficult to classify, and we have hypothesized that such sentences are likely just as difficult for machines to classify. In this paper we outline CrowdTruth, a previously presented method for scoring ambiguous sentences that suggests that existing modes of truth are inadequate, and we present for the first time a set of weighted metrics for evaluating the performance of experts, the crowd, and a trained classifier in light of ambiguity. We show that our theory of truth and our metrics are a more powerful way to evaluate NLP performance over traditional unweighted metrics like precision and recall, because they allow us to account for the rather obvious fact that some sentences express the target relations more clearly than others.

1 Introduction

NLP often relies on the development of a set of gold standard annotations, or *ground truth*, for the purpose of training, testing and evaluation. Distant supervision [17] is a helpful solution that has given linked data sets a lot of attention in NLP, however the data can be noisy. Human annotators can help to clean up this noise, however for Clinical NLP 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 [6].

Furthermore, the assumption that the gold standard represents a universal and reliable model for language is flawed [4]. Disagreement between annotators is usually eliminated through overly prescriptive guidelines, resulting in data that is neither general nor reflects language's inherent ambiguity. The process of acquiring ground truth by working exclusively with domain experts is costly and non-scalable.

Crowdsourcing can be a much faster and cheaper procedure than expert annotation, and it 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.

Our approach can overcome the limitations of gathering expert ground truth, by using disagreement analysis on crowd annotations to model the ambiguity inherent in medical text. We have previously shown our approach can improve relation extraction classifier performance over annotated data provided by experts, can effectively identify low-quality workers, and identify issues with the annotation tasks themselves. In this paper we explore the hypothesis that our sentence-level metrics are providing useful information about sentence clarity, and present initial results on the value of different approaches to scoring that the traditional precision, recall, and accuracy.

2 Related Work

Crowdsourcing ground truth has shown promising results in a variety of domains. [12] 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. [14] 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 [15], Twitter entity extraction [11], multilingual entity extraction and paraphrasing [8], and taxonomy creation [9]. 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.

The role of inter-annotator disagreement when building a gold standard has previously been discussed by [19]. After empirically studying part-of-speech datasets, the authors found that inter-annotator disagreement is consistent across domains, even across languages. Furthermore, most disagreement is indicative of debatable cases in linguistic theory, rather than faulty annotation. We believe these findings manifest even more strongly for NLP tasks involving semantic ambiguity, such as relation extraction. In assessing the Ontology Alignment Evaluation Initiative (OAEI) benchmark, [7] found that disagreement between annotators (both crowd and expert) is an indicator for inherent uncertainty in the domain knowledge, and that current benchmarks in ontology alignment and evaluation are not designed to model this uncertainty.

Human annotation is a process of semantic interpretation. It can be described using the triangle of reference [13], that links together three aspects: sign (input text), interpreter (worker), referent (annotation). Ambiguity for one aspect of the triangle will propagate and affect the others – e.g. an unclear sentence will cause more disagreement between workers. Therefore, in our work, we use metrics to harness disagreement for each of the three aspects of the triangle, measuring the quality of the worker, as well as the ambiguity of the text and the task.

3 Methods

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 disorders and symptoms, 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 our metrics for capturing disagreement, and then used to train a model for relation extraction. In parallel, we also constructed a model based on data from a traditional gold standard using 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 [5]. 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. The seed relations were restricted to a set of eleven UMLS relations important for clinical decision making [21] (see Tab.1). All of the data that we have used is available online at: `http://data.crowdtruth.org/medical-relex`.

Table 1: Set of medical relations.

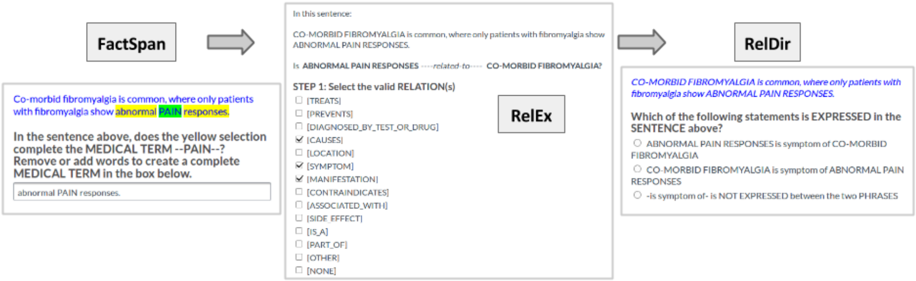
Relation	Corresponding UMLS relation(s)	Definition	Example
treat	may treat	therapeutic use of a drug	penicillin treats infection
prevent	may prevent	preventative use of a drug	vitamin C prevents influenza
diagnosis	may diagnose	diagnostic use of an ingredient, test or a drug	RINNE test is used to diagnose 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

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 is performed in a workflow of three tasks (Fig.1). The sentences were pre-processed to determine whether the terms found with distant supervision are complete or not; identifying complete medical terms is difficult, and the automated method left a number of terms still incomplete, which was a significant source of error for the crowd in subsequent stages, so the incomplete terms were sent 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. We also added four new relations (e.g. *associated with*), to account for weaker, more general links between the terms (see Tab.1). The workers were able to read the definition of each relation, and could choose any number of relations per sentence. There were options for the cases when the terms were related, but not by those we provided (*other*), and for no relation between the terms (*none*). Finally, the results from *RelEx* were passed to another crowdsourcing task (*RelDir*) to determine the direction of the relation with regards to the two extracted terms. (*FactSpan* and *RelDir*) were added to the basic *RelEx* task to correct the most common sources of errors from the crowd.

All three crowdsourcing tasks were run on the CrowdFlower platform [4] with 10-15 workers per sentence, to allow for a distribution of perspectives. 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.



Co-morbid fibromyalgia is common, where only patients with fibromyalgia show abnormal PAIN responses.

In the sentence above, does the yellow selection complete the MEDICAL TERM --PAIN--? Remove or add words to create a complete MEDICAL TERM in the box below.

abnormal PAIN responses.

In this sentence:

CO-MORBID FIBROMYALGIA is common, where only patients with fibromyalgia show ABNORMAL PAIN RESPONSES.

Is ABNORMAL PAIN RESPONSES ----*related-to*---- CO-MORBID FIBROMYALGIA?

STEP 1: Select the valid RELATION(s)

☐ [TREATS]
☐ [PREVENTS]
☐ [DIAGNOSED_BY_TEST_OR_DRUG]
☑ [CAUSES]
☐ [LOCATION]
☑ [SYMPTOM]
☑ [MANIFESTATION]
☐ [CONTRAINDICATES]
☐ [ASSOCIATED_WITH]
☐ [SIDE_EFFECT]
☐ [IS_A]
☐ [PART_OF]
☐ [OTHER]
☐ [NONE]

CO-MORBID FIBROMYALGIA is common, where only patients with fibromyalgia show ABNORMAL PAIN RESPONSES.

Which of the following statements is EXPRESSED in the SENTENCE above?

○ ABNORMAL PAIN RESPONSES is symptom of CO-MORBID FIBROMYALGIA
○ CO-MORBID FIBROMYALGIA is symptom of ABNORMAL PAIN RESPONSES
○ -is symptom of- is NOT EXPRESSED between the two PHRASES

Fig. 1: CrowdTruth Workflow for Medical Relation Extraction on CrowdFlower [10].

3.3 Metrics

For each crowdsourcing task in the workflow, the crowd output was processed with our metrics – a set of general-purpose crowdsourcing metrics [3]. These metrics attempt to model the crowdsourcing process based on the triangle of reference [18], with the vertices being the input sentence, the worker, and the target relations. Our theory is that 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. A low quality worker is more likely to disagree with the other workers, and this can impact the apparent quality of the sentence. If one of the target relations is itself ambiguous, it will be difficult to identify and will generate disagreement that may have nothing to do with the quality of sentences or workers. Our metrics account for this by isolating the signals from the workers, sentences, and the target relations, and more accurately evaluating each. In previous work we have validated this premise in several empirical studies [3].

In this paper we focus specifically on sentence quality, to evaluate our claim that low quality sentences are difficult to annotate, and likewise difficult for ma-

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

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 2: 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.*

chines to process. To measure this effect, we begin with a simple representation of the crowd output from the *RelEx* task:

- *annotation vector:* the annotations of one worker for one sentence. For each worker i their solution to a task on a sentence s is the vector $W_{s,i}$. If the worker selects a relation, 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, *none* and *other*.
- *sentence vector:* For every sentence s, we sum the annotation vectors for all workers on the given task: $V_s = \sum_i W_{s,i}$.

The sentence vector is a simple representation of the annotations on a sentence, and leads to the *sentence-relation score*, which measures, for each relation, the degree to which a sentence vector diverges from perfect agreement on that relation. It is simply 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. The purpose of the experiments is to provide evidence that the *srs* is measuring the clarity, or inversely the ambiguity, of a sentence with respect to a particular relation, and that sentences with low scores present difficulty for the crowd, experts, and machines alike.

We use a two-step process to eliminate low-quality worker annotations. We run the sentence metrics and filter out sentences whose quality score is one standard deviation below the mean, then we run our worker metrics [2] on the remaining sentences and filter out all workers below a trained threshold. The purpose of the first step is to ensure the worker quality scores are not adversely impacted by confusing sentences. We remove all low quality worker annotations and re-evaluate the sentence metrics on all sentences.

3.4 Training the model

At the highest level our research goal is to investigate crowdsourcing as a way to gather human annotated data for training and evaluating cognitive systems. In these experiments we were specifically gathering annotated data for a sentence-level relation extraction classifier [21]. This classifier is trained per individual relation, by feeding it both *positive* and *negative* examples. 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 $[-1, 0)$.

To test our approach, we gathered four annotated data sets and trained classifier models for the *cause* relation using five-fold cross-validation over the 902 sentences:

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. 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. 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 (see below). An example of how the scores were processed is given in Tab.2.

In order to directly compare the expert to the crowd annotations, it was necessary to annotate precisely the same sentences using each method, and train the classifier on each set. The limitation on batch size came from the availability of our experts, we were only able to use them for 902 sentences. In a batch this small, we found that the sentence-relation score, which ranged between [0,1] and rarely assigned a weight of 1, diluted the positive signal too much in comparison to the expert scores which were simply 0 or 1. We experimented, on a different data set, with rescaling the scores and selected the range that yielded the highest quality score, specified above.

3.5 Evaluation setup

In order for a meaningful comparison between the crowd and expert models, we verified the sentences to provide a *ground truth* – a discrete positive or negative label on each sentence used in evaluation (for training, only the scores from the respective data set were used). While the main purpose of this work is to move beyond discrete labels for truth, we needed a reference standard to establish that our approach is at least as good as the accepted practice. To produce this reference standard, we first selected the positive/negative threshold for *sentence-relation score* in the *crowd* dataset that yielded the highest agreement between the crowd and the experts, and then accepted all 755 sentences where the experts and crowd agreed as true positives. The remaining sentences were manually evaluated and assigned either a positive, negative, or ambiguous value. The ambiguous cases were subsequently removed resulting in 902 sentences. In this way we created reliable, unbiased test scores, to be used in the evaluation of the models. In some sense, removing the ambiguous cases penalizes our approach, which is designed specifically to help deal with them, but again we want to first establish our approach is at least as good as accepted practice.

4 Results

4.1 Preliminary experiments

As reported in [10] and summarized here, we compared each of the four datasets to our vetted reference standard, to determine the quality of the *cause* relation annotations, as shown in Fig.2. As expected, the baseline data was the lowest quality, 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-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 a maximum of 0.907 F1 score at

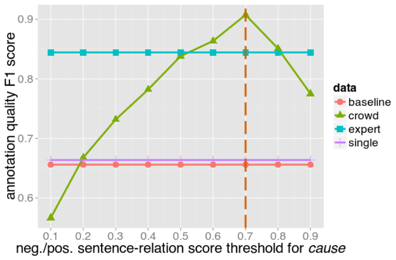


Fig. 2: Annotation quality F1 score per negative/positive threshold for cause.

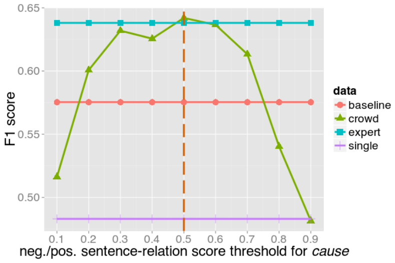


Fig. 3: Classifier F1 scores when trained with each dataset.

a threshold of 0.7. This difference is significant with $p = 0.007$, measured with McNemar's test [16].

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 (discussed in more detail in [10]). The results of the evaluation (Fig.3) 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.

4.2 Results and Discussion

We believe the discrete notion of truth is obsolete and should be replaced by something more flexible. For the purposes of semantic interpretation tasks for which crowdsourcing is appropriate, we propose our annotation-level metrics as a suitable replacement. In this case, the $sentence - relationscore$ gives a real-valued score that measures the degree to which a particular sentence expresses a particular relation between two terms. We believe the preliminary experiments demonstrate the approach is sound. Our primary results evaluate the sentence-relation score as a measure of the *clarity with which a sentence expresses the relation.* To this end, we define the following metrics:

- *sentence weight:* For a given positive/negative threshold τ, if $srs(s) \geq \tau$ for sentence s then the sentence weight $w_s = srs(s)$, otherwise $w_s = 1 - srs(s)$.
- *weighted precision:* We collect true and false positives and negatives in the standard way based on the vetted reference standard, such that $tp(s) = 1$ iff s is a true positive, and 0 otherwise, similarly for fp, tn, fn. Where normally $p = tp/(tp + fp)$, weighted precision $p' = \dfrac{\sum_s w_s tp(s)}{\sum_s w_s(tp(s) + fp(s))}$.
- *weighted recall:* Where normally $r = tp/(tp + fn)$, weighted recall $r' = \dfrac{\sum_s w_s tp(s)}{\sum_s w_s(tp(s) + fn(s))}$.
- *weighted f-measure:* Is the harmonic mean of weighted precision and recall: $f1' = 2p'r1/(p' + r')$

If the srs metric is a true measure of clarity, then we would expect it to be more likely for low clarity sentences to be wrong, and less likely for high clarity sentences, and this should be revealed in an overall increase of the weighted scores over the unweighted. In Tab. 3, we show a comparison of five data sets. In the first two columns, the annotation quality of each data set is shown, comparing the F1 to the weighted F1'. The F1' scores are higher in all cases, revealing that human annotators are indeed having trouble correctly annotating these sentences. The baseline scores are the least affected by the weighting, which also fits with our intuition since the baseline does not use human judgment at all.

Table 3: Model evaluation results for each dataset.

	Annotation Quality		Classifier Performance					
	F1	**F1'**	**F1**	**F1'**	**P**	**P'**	**R**	**R'**
crowd@.5	0.838	0.933	0.642	0.687	0.565	0.632	0.743	0.754
crowd@.7	0.907	0.963	0.613	0.646	0.620	0.678	0.611	0.622
baseline	0.656	0.689	0.575	0.606	0.436	0.474	0.845	0.842
single	0.664	0.734	0.483	0.507	0.496	0.545	0.473	0.478
expert	0.844	0.861	0.638	0.658	0.672	0.711	0.605	0.616

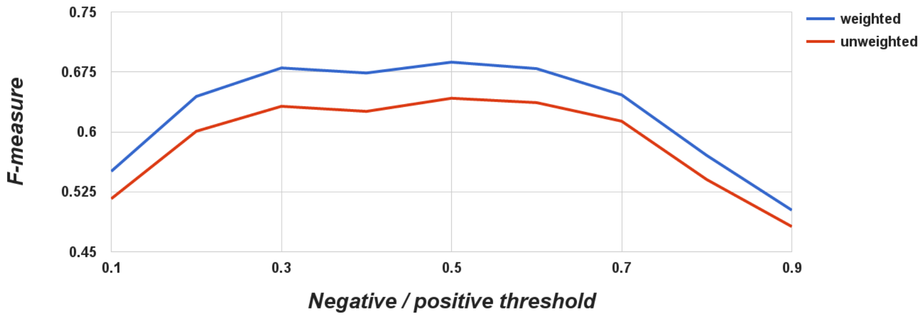


Fig. 4: Comparison of weighted to non-weighted F1 scores for the crowd-trained classifier at different thresholds.

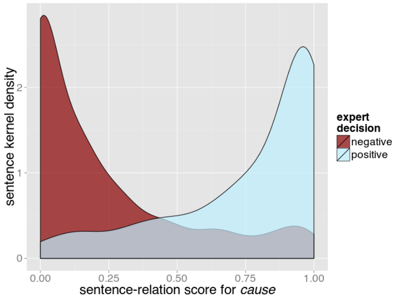


Fig. 5: Density of cause sent.-rel. score over the expert data.

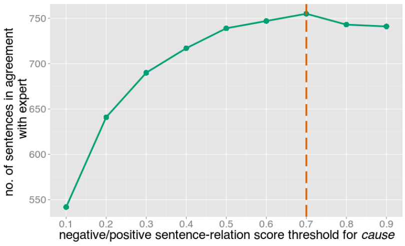


Fig. 6: Crowd & expert agreement per neg./pos. threshold for cause.

The next six columns in each row show classifier performance when trained by that dataset. The first pair of columns compare the F1 to F1', and for interest the final four columns show the precision and recall. In all cases the classifier F1' is greater than F1, indicating that, as with humans, machines have trouble correctly interpreting sentences with a low *srs*. The only weighted metric that does not increase is the baseline recall, again this is justified as the baseline does not actually require any interpretation.

In Fig. 4 we show how the classifier performs throughout the possible thresholds, the weighted scores are consistently higher.

We also analyzed the data to understand the overlap between the crowd scores and the experts. In Fig.5 we compared the frequency of sentences with *cause* annotations at different sentence-relation scores (measured with kernel density estimation [20]) to the expert annotations of the same sentences. The result shows high agreement between the crowd and the expert – a low sentence-relation score is highly correlated with a negative expert decision, and a high score is highly correlated with a positive expert decision. In Fig.6 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 in common between crowd and expert. The remaining 147 sentences were manually evaluated to build the test partition.

5 Conclusion

A widespread use of linked data for information extraction is *distant supervision*, in which relation tuples from a data source are found in sentences in a text corpus, and those sentences are treated as training data for relation extraction systems. Distant supervision is a cheap way to acquire training data, but that data can be quite noisy, which limits the performance of a system trained with it. Human annotators can be used to clean the data, but in some domains, such as medical NLP, it is widely believed that only medical experts can do this reliably. Current methods for collecting this human annotation attempt to minimize disagreement between annotators, but end up failing to capture the ambiguity inherent in language. We believe this is a vestige of an antiquated notion of truth being a discrete property, and have developed a powerful new method for representing truth.

In this paper we have presented results that show that using a larger number of workers per example – up to 15 – can form a more accurate model of truth at the sentence level, and significantly improve the quality of the annotations. It also benefits systems that use this annotated data, such as machine learning systems, significantly improving their performance with higher quality data. Our primary result is to show that our scoring metric for sentence quality in relation extraction supports our hypothesis that higher quality sentences are easier to classify – for crowd workers, experts, and machines, and our model of truth allows us to more faithfully capture the ambiguity that is inherent in language and human interpretation.

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. American Medical Informatics Association (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.: The Three Sides of CrowdTruth. Journal of Human Computation 1, 31–34 (2014)
4. Aroyo, L., Welty, C.: Truth is a lie: Crowd truth and the seven myths of human annotation. AI Magazine 36(1), 15–24 (2015)
5. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research 32(suppl 1), D267–D270 (2004)
6. 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 American Medical Informatics Association 18(5), 540–543 (2011)
7. Cheatham, M., Hitzler, P.: Conference v2. 0: An uncertain version of the OAEI Conference benchmark. In: The Semantic Web–ISWC 2014, pp. 33–48. Springer (2014)
8. Chen, D.L., Dolan, W.B.: Building a persistent workforce on mechanical turk for multilingual data collection. In: Proceedings of The 3rd Human Computation Workshop (HCOMP 2011) (2011)
9. 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)
10. Dumitrache, A., Aroyo, L., Welty, C.: Achieving expert-level annotation quality with CrowdTruth: the case of medical relation extraction. In: Proceedings of the 2015 International Workshop on Biomedical Data Mining, Modeling, and Semantic Integration (BDM2I-2015), 14th International Semantic Web Conference (2015)
11. 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, Association for Computational Linguistics (2010)
12. 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 Association for Computational Linguistics (Volume 2: Short Papers). pp. 377–382. Association for Computational Linguistics, Baltimore, Maryland (June 2014)
13. Knowlton, J.Q.: On the definition of "picture". AV Communication Review 14(2), 157–183 (1966)
14. 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)
15. 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)
16. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12(2), 153–157 (1947)
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. Association for Computational Linguistics (2009)

18. Ogden, C.K., Richards, I.: The meaning of meaning. Trubner & Co, London (1923)
19. Plank, B., Hovy, D., Søgaard, A.: Linguistically debatable or just plain wrong? In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 507–511. Association for Computational Linguistics, Baltimore, Maryland (June 2014)
20. Silverman, B.W.: Density estimation for statistics and data analysis, vol. 26. CRC press (1986)
21. Wang, C., Fan, J.: Medical relation extraction with manifold models. In: 52nd Annual Meeting of the ACL, vol. 1. pp. 828–838. Association for Computational Linguistics (2014)