Disagreement in Crowdsourcing and Active Learning for Better Distant Supervision Quality

ANCA DUMITRACHE, LORA AROYO, Vrije Universiteit Amsterdam
CHRIS WELTY, Google Research

1. INTRODUCTION

Distant supervision (DS) [Mintz et al. 2009; Welty et al. 2010] is a well-established semi-supervised method for performing relation extraction from text. It is based on the assumption that, given a knowledge base (KB) contains a relation between a pair of terms, then any sentence that contains that pair is likely to express the relation. This approach can generate false positives, as not every mention of a term pair in a sentence means a relation is also present [Angeli et al. 2014].

Crowdsourcing is a possible solution to correcting the mistakes of DS with human annotation, but the cost of employing annotators for large amounts of data is a bottleneck. [Angeli et al. 2014] present an active learning approach to select the most useful sentences that need human re-labeling using a query by committee – training a set of classifiers on partitions of DS data, and in the evaluation phase selecting the sentences where the models disagree on the relation classification and sending those to the crowd. However, as our experiments show, the error rate in the DS data can be so high that classifier agreement is not a reliable indication of good training data. Moreover, this approach does not take into account the disagreement that usually occurs between annotators, instead labeling data with the most commonly picked relation.

We have proposed the *CrowdTruth* [Aroyo and Welty 2014] method for crowdsourcing a ground truth for relation extraction by capturing inter-annotator disagreement. Our previous work in medical relation extraction[Dumitrache et al. 2015] has shown that harnessing disagreement in crowd annotations can be used to model the ambiguity inherent in natural language. In this paper, we present ongoing work on combining active learning with the CrowdTruth methodology for further improving the quality of DS training data. We report the results of a crowdsourcing experiment ran on 2,500 sentences from the open domain. We show that modeling disagreement can be used to identify interesting types of errors caused by ambiguity in the TAC-KBP knowledge base, and we discuss how an active learning approach can incorporate these observations to utilize the crowd more efficiently.

2. CROWDSOURCING OVERVIEW

The dataset in our experiment contains 2,500 sentences from the NIST TAC-KBP 2013 English Slotfilling data that were annotated with DS. We focused on 16 relations (Fig.1) that occur between terms of types $Person$, $Organization$ and $Location$, the same set used by [Angeli et al. 2014]. We ran a multiple-choice task on Crowdflower, asking 15 workers to annotate each sentence with the appropriate relations, or choose the option none in none of the relations apply (they were encouraged to select all that apply), and paid them $0.05 per sentence. We have made the crowd annotated data available [Dumitrache et al. 2016].

The crowd output was processed with CrowdTruth metrics [Aroyo and Welty 2014]. For every sentence, the output of each worker is a binary *worker vector* corresponding to the relation set, where the relations selected are equal to '1', and the rest to '0'. The *sentence vector* is the sum of all worker vectors for the given sentence. Then for each relation, we compute the *sentence-relation score (SRS)*

as the ratio of workers that picked that relation over the total of number of workers. This metric indicates how clearly the relation is expressed in the sentence. We also calculate the *causal power* between relation pairs (RCP) over the entire set of sentences. Given two relations i and j, $RCP(R_i, R_j) = [P(R_j|R_i) - P(R_j|\neg R_i)]/[1 - P(R_j|\neg R_i)]$, where $P(R_i)$ is the probability that relation R_i is annotated at least once in the sentence. A high RCP score indicates that relation R_i causes relation R_j to appear (the score is not symmetrical, because the opposite might not be true). We compute two types of RCP: the *micro RCP* uses the probability the relations co-occur in the worker vectors, calculating how likely one worker is to annotate two relations together; the *macro RCP* calculates the probabilities in the sentence vectors, capturing causality as a result of two relations being annotated together in the same sentence, but not necessarily by the same workers.

Using the SRS as a ground truth, we evaluated the correctness of the DS labels. We set an agreement threshold for the SRS of 0.5, equivalent to the majority vote decision, and labeled any value equal or above it as a positive, and the rest as negative. The evaluation results (Fig. 1) show *considerable variation in DS data quality according to the relation*. The $origin$ and $place\_of\_death$ relations scored particularly badly, with more than 90% false positives. With such a high error rate in some relations, it is arguable that a classifier could learn anything meaningful, raising questions about an approach like query by committee for crowd data selection. Manual investigation of the data showed that many sentences contain a $Person$ - $Location$ pair, with the KB also specifying that the person died at that location. However, in most of these cases, the sentence expresses $places\_of\_residence$ – most people die in the place that they have lived. The $origin$ relation data suffers from the same problem.

The results of the macro RCP analysis (Tab.II) shows that the $place\_of\_birth$ relation has a high causal power (0.42) over $origin$, meaning that when $place\_of\_birth$ is annotated in a sentence, $origin$ is also likely to appear. However the micro RCP (Tab.III) for this relation pair shows causal power is reduced almost by half (0.25). This means that, while the two relations might appear in the same sentence, it is not the same workers that are picking them. A similar effect occurs for the $top\_employee\_or\_member$ relation, which has some macro causal power over $founded\_organization$, but not micro. In contrast, the micro RCP shows a clear trend that a worker picking $top\_employee\_or\_member$, would also pick $employee\_or\_member$. The effect could be due to the fact that $top\_employee\_or\_member$ clearly refers to a subset of $employee\_or\_member$ examples, making the causal power between the relations less ambiguous for the workers. Moreover, even with such a clear dependency between the relations, the workers still did not agree entirely that the two relations should be annotated together. These results show us that, even though the task is multiple choice and we encouraged the workers to pick all relations that apply, the *crowd workers typically tend to pick the one relation which appears the most suitable to them* (e.g. many did not pick the top employee relation if they thought that founder fits better). This is supported by the low average number of relations a worker picks per sentence (equal to 1.26). In addition, this analysis exposes the *ambiguities in the TAC-KBP relation set*. Relation pairs like $origin$ and $place\_of\_birth$ are confusable to workers because their semantics are similar, and they are routinely expressed in similar ways in language. Therefore, it might makes sense to merge them together when building a ground truth for training a model.

3. ACTIVE LEARNING WITH CROWDTRUTH

The analysis of our crowdsourcing experiment showed the considerable variation in DS data quality according to the relation (Fig.1). This raises questions about the efficiency of the query by committee approach [Angeli et al. 2014]. Training an ensemble of classifiers on relations with a very high rate of false positives (e.g. the $place\_of\_death$ relation) is likely to replicate the same bias in every classifier.

We propose an active learning approach (Fig.2) that selects sentences for crowd labeling based on the likelihood of being false positive. Our methodology is centered around two models: a *relation extraction*

classifier and a *false positive classifier*. As opposed to the query by committee approach, a relation extraction classifier trained with DS data and evaluated with the crowd is able to produce all the sentences that are difficult to classify, without being affected by a high rate of false positives in the training data. Evaluating the classifier results as opposed to DS labels with the crowd helps identify the sentences that specifically bias the classifier – i.e. if the training data contains flaws, but the classifier is still able to produce good results, there might not be a need for crowd re-labeling. The false positive classifier is then used to learn the features of the sentences that bias relation extraction. After it is trained, the model can be used to target crowd annotation on likely false positive sentences.

The relation extraction classifier used in our methodology is a *convolutional neural network* [Nguyen and Grishman 2015], trained on DS data, and evaluated on CrowdTruth data. A separate classifier is trained for each relation. The negative sentences for a given relation are the positive sentences for all the other relations in the set, with some exceptions. In the crowdsourcing results, we have shown some relations holding causal power over others – if a relation i causes relation j to appear as well, sentences where j is expressed cannot be used as negatives for i. We also merge into one relation the tuples that overlap in semantics, because relations that are expressed in similar ways could impact classifier performance when differentiating between them.

To evaluate the relation extraction model on CrowdTruth data, we set a threshold in the sentence-relation score at 0.5 for separating between negatives and positives. This data is used to label the classifier results as true positive, false positive etc. The evaluation results are then used to train the false positive classifier. False positives from the evaluation set are labeled as positives, and the rest of the sentences as negatives. Because CrowdTruth data is more expensive to acquire than DS, we are limited in the number of training examples for the false positive classifier. Therefore, instead of another convolutional neural network, we use *logistic regression*, which can handle smaller volumes of data. The number of training examples for the second model can also be increased by reusing DS data with negative labels, which is unlikely to be classified as false positive.

After it is trained, the false positive classifier is used to pick sentences for crowdsourcing annotation, with the purpose of improving the relation extraction model. Similar to the approach of [Angeli et al. 2014], in order to perform *active learning from the crowd when re-training the first model*, we give the crowd sentences a greater weight in the model than the DS sentences, which have a greater probability of being incorrect. We do this by altering the train batches of the relation extraction model to reuse crowd data more often than DS data. This process could also be altered to take into account the CrowdTruth sentence-relation score – i.e. sentences with low scores should not hold as much weight as the high score ones, so they should be reused less often in the training batches. Finally, to evaluate the performance of our active learning approach for picking sentences to be re-labeled by the crowd, we compare it with a semi-supervised classifier where sentences are picked randomly.

Table I. : RCP for relation subset: $place\_of\_birth$ *(PoB),* $origin$ *(O),* $places\_of\_residence$ *(PoR),* $founded\_organization$ *(FO),* $employee\_or\_member$ *(EoM),* $top\_employee\_or\_member$ *(TEoM). The scores show the causal power* $RCP(R_i, R_j)$ *of relations* R_i *in the rows, over the relations* R_j *in the columns.*

Table II. : Macro RCP.

	PoB	O	PoR	FO	EoM	TEoM
PoB	1	0.42	-0.12	-0.14	-0.15	-0.19
O	0.17	1	0.11	-0.05	-0.05	-0.07
PoR	-0.11	0.24	1	-0.12	-0.09	-0.12
FO	-0.03	-0.02	-0.03	1	-0.02	0.03
EoM	-0.1	-0.08	-0.07	-0.07	1	0.09
TEoM	-0.43	-0.38	-0.31	0.32	0.33	1

Table III. : Micro RCP.

	PoB	O	PoR	FO	EoM	TEoM
PoB	1	0.25	0.02	-0.09	-0.1	-0.12
O	0.17	1	0.12	-0.06	-0.03	-0.05
PoR	0.03	0.21	1	-0.12	0.003	-0.04
FO	-0.02	-0.02	-0.03	1	0.008	0.009
EoM	-0.12	-0.06	0.003	0.03	1	0.29
TEoM	-0.25	-0.17	-0.08	0.06	0.5	1

REFERENCES

G. Angeli, J. Tibshirani, J. Wu, and C. Manning. 2014. Combining Distant and Partial Supervision for Relation Extraction.. In *EMNLP*. 1556–1567.

L. Aroyo and C. Welty. 2014. The Three Sides of CrowdTruth. *Journal of Human Computation* 1 (2014), 31–34. Issue 1.

A. Dumitrache, L. Aroyo, and C. Welty. 2015. Achieving Expert-Level Annotation Quality with CrowdTruth. In *Proc. of BDM2I Workshop, ISWC*.

A. Dumitrache, L. Aroyo, and C. Welty. 2016. 2210 TAC-KBD CrowdTruth Annotated Sentences. github.com/CrowdTruth/Open-Domain-Relation-Extraction. (2016).

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proc. of IJCNLP 2009: Volume 2*. ACL, 1003–1011.

T.H. Nguyen and R. Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proc. of NAACL-HLT*. 39–48.

C. Welty, J. Fan, D. Gondek, and A. Schlaikjer. 2010. Large Scale Relation Detection. In *Proc. of the NAACL HLT FAM-LbR*. 24–33.

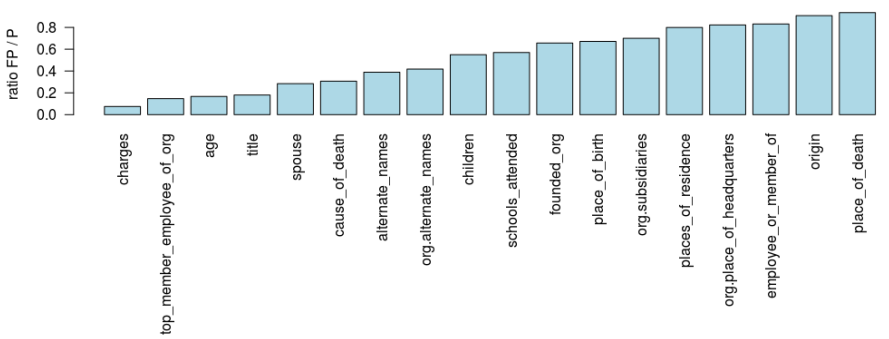

Fig. 1: DS ratio of false positive over all positive labels, using the crowd as ground truth.

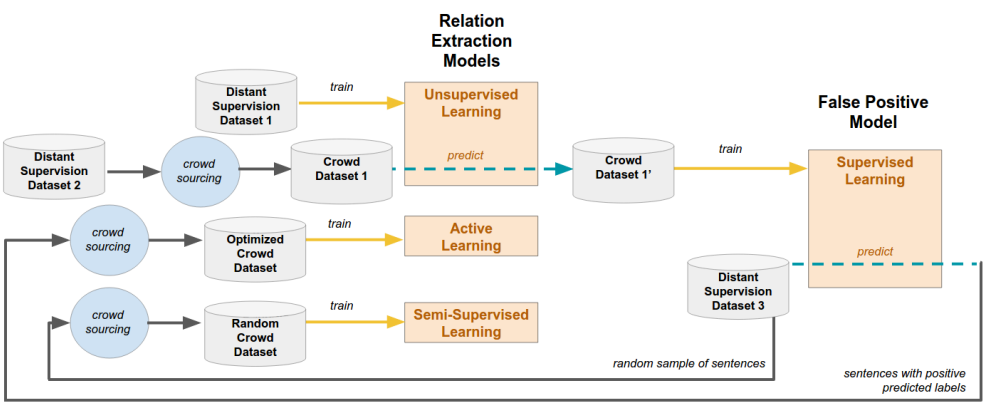

Fig. 2: Active Learning Pipeline.