False Positive and Cross-relation Signals in Distant Supervision Data

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

1. INTRODUCTION

Distant supervision (DS) [Mintz et al. 2009] is a well-established method for relation extraction from text, based on the assumption that when a knowledge-base contains a relation between a term pair, then sentences that contain that pair are 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 expressed [Feng et al. 2017]. Furthermore, dependencies between the semantics of the relations such as causality or contradiction are also not considered by the DS methodology. It is often assumed that these disadvantages are compensated for by the scale of the data a DS method can produce, or can be largely overcome with crowdsourced human annotation [Angeli et al. 2014].

Crowdsourcing is a well-used approach to correcting the mistakes in DS by scaling out cheap human annotation. We have been studying the problem of collecting human annotations from the crowd using the CrowdTruth methodology [Aroyo and Welty 2013]. Our method differs in that it gathers many annotations for the same examples, to better reflect properties like ambiguity, human error and spam, and the target semantics [Aroyo and Welty 2014].

This abstract describes the results of a crowdsourcing relation extraction task identifying two specific problems we have found with distant supervision training data: the widely varying degree of false positives across different TAC-KBP relation types, and the observed causal connection between relations. We expose these problems using a novel approach to gathering human annotated data, CrowdTruth [Aroyo and Welty 2014; Aroyo and Welty 2015; Aroyo and Welty 2013], analyze them, and offer preliminary heuristic and statistical approaches to incorporating them back into DS-based training, that provides better sentence-level relation extraction results, without requiring crowdsourcing on the full set of data. The full paper describing this work is [Dumitrache et al. 2017].

2. EXPERIMENTAL SETUP

We crowdsourced annotations for 2,500 sentences from the NIST TAC-KBP 2013 English Slotfilling data that were annotated with DS. We split the data in half into a test and dev set. We focused the crowd annotations on a subset of 16 relations (Fig.1). We ran a multiple-choice crowdsourcing task on Crowdflower, asking workers to annotate each sentence with the appropriate relations. The data is available online.[1] The crowd output was processed with the CrowdTruth metrics,[2]. We calculate the per-relation **false positive (FP)** rate and the **causal power (RCP)** between relation pairs over the dev set. Spam removal was performed as well, but the details of this process are not relevant for the paper.

For every sentence, the annotations of each worker form a binary *worker vector*, 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

[1] https://github.com/CrowdTruth/Open-Domain-Relation-Extraction
[2] https://git.io/v5iTB

given sentence. Then for each relation, we compute the *sentence-relation score* of relation i (SRS_i) as the ratio of workers that picked that relation over the total of number of workers. The SRS measures how clearly the relation is expressed in the sentence, and is used as a continuous truth measure in a lot of our work. In order to make our results compatible with discrete evaluation metrics (e.g. P, R, F1), we have chosen a threshold of 0.5 per relation, corresponding to the majority vote, that allows for multiple relations to be considered correct in a sentence. **FP rates** are then computed per relation on the dev set with this threshold. **RCP** [Cheng 1997] is an estimate of the probability that the presence of one relation implies the presence of another. 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 in the sentence.

In addition to the dev and test sets, we also used a training set of 235,000 sentences annotated by DS from freebase relations as a baseline for training a relation extraction classifier [Nguyen and Grishman 2015]. The evaluation was done on the held-out test set. We constructed an experiment to compare the performance of the relation extraction classifier trained on the DS baseline in comparison with 3 versions of the DS data enhanced with signals from the CrowdTruth metrics from the dev set:

(1) **DS merged:** DS is augmented by manually merging relations with symmetric RCP, and adding the implied relation in the case of asymmetric RCP.
(2) **DS_RCP:** When a relation i has a positive **DS** label for a given sentence, the labels of all other relations $j \neq i$ are updated by adding the RCP that i has over j.
(3) **DS_FP:** DS is augmented by removing relations with high FP rates in sentences where other relations also appear.

3. RESULTS

Fig. 1 shows the correctness of the DS labels on the dev set. There is *considerable variation in DS data quality across relations*. The $origin$ and $place\_of\_death$ relations scored particularly badly, with more than 90% false positives. These results were used to construct the **DS FP** dataset.

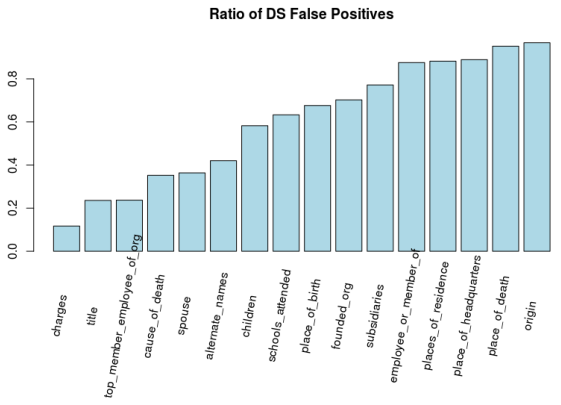

Fig. 1: DS ratio of false positive over all positive labels, using the crowd as ground truth.

The RCP analysis (Tab. III) shows that the $place\_of\_birth$ relation has a high causal power over $origin$, meaning that when $place\_of\_birth$ is annotated in a sentence, $origin$ is also likely to appear, with

the inverse causal power. This high co-causality seems to indicate a confusion between the two relations. In the crowd data we also observed a co-causality for $employee\_or\_member$ and $top\_employee\_or\_member$, with only a slight preference in the data for what we expect to be the "correct" causal direction (that $top\_employee\_or\_member$ causes $employee\_or\_member$). These results were used to construct the **DS merged** dataset.

Table I. : Crowd-based RCP

	PoB	O	PoR	PoD	FO	EoM	TEoM
PoB	1	**0.64**	0.17	-0.12	-0.19	-0.2	-0.21
O	**0.88**	1	0.31	-0.16	-0.29	-0.22	-0.22
PoR	0.42	**0.56**	1	**-0.1**	-0.59	0.12	0.13
PoD	-0.03	-0.03	-0.01	1	-0.04	-0.05	-0.05
FO	-0.07	-0.07	-0.09	-0.06	1	0.1	0.13
EoM	-0.45	-0.36	0.11	-0.47	0.62	1	0.82
TEoM	-0.5	-0.38	0.13	-0.45	0.86	**0.86**	1

Table II. : DS-based RCP.

	PoB	O	PoR	PoD	FO	EoM	TEoM
PoB	1	**-0.6**	0.55	-0.14	-0.54	-0.48	-0.57
O	**-0.02**	1	-0.11	-0.16	-0.16	0.19	-0.15
PoR	0.65	**-0.33**	1	**0.45**	-0.7	-0.68	-0.75
PoD	-0.06	-0.18	0.17	1	-0.18	-0.13	-0.19
FO	-0.08	-0.06	-0.09	-0.06	1	0.09	0.09
EoM	-0.35	0.35	-0.42	-0.21	0.46	1	0.66
TEoM	-0.16	-0.1	-0.17	-0.12	0.34	**0.24**	1

Table III. : RCP for relation subset: $place\_of\_birth$ (PoB), $origin$ (O), $places\_of\_residence$ (PoR), $place\_of\_death$ (PoD), $founded\_organization$ (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. Significant changes between crowd annotation based causal power and distant supervision are in bold.

The relation extraction evaluation results shown in Tab.IV are not overwhelming, but highly indicative. The manually merged DS shows a huge improvement across the board over the baseline, with the overall highest P and F1. Augmenting DS with RCP was comparable in precision to the baseline, but scored a huge win in recall. Correcting DS with FP did not impact the results over the baseline, mainly because there were not many $place\_of\_death$ relations in the DS data nor the test set, and any improvement did not impact the overall result. We are confident that more systematic treatment of false positive rates will improve performance.

	Precision	**Recall**	**F1 score**
DS	0.19	0.22	0.2
DS merged	**0.43**	0.33	**0.37**
DS_RCP	0.19	**0.48**	0.27
DS_FP	0.21	0.22	0.21

Table IV. : Relation extraction evaluation at 20,000 training steps.

We have shown that (1) there is considerable headroom in cross-relation signals, and a more robust approach holds promise to eliminate manual analysis, and work as part of an overall pipeline that includes partial crowd data, and (2) there is a very significant variation in the false positive rate in distant supervision data, and it seems extremely likely that this can be exploited to improve training. Currently, we are considering experiments that take advantage of another aspect of our CrowdTruth method: the identification of ambiguity in sentences where workers do not agree on the outcome. We believe a more continuous truth measure as opposed to the rather arbitrary discrete measure will be productive.

REFERENCES

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

Lora Aroyo and Chris Welty. 2013. Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *Web Science 2013. ACM* (2013).

L. Aroyo and C. Welty. 2014. The Three Sides of CrowdTruth. *J. of Hum. Comp.* 1 (2014), 31–34. Issue 1.

Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine* 36, 1 (2015), 15–24.

Patricia W. Cheng. 1997. From Covariation to Causation: A Causal Power Theory. *Psychological Review* 104, 2 (1997), 367–405.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2017. False Positive and Cross-relation Signals in Distant Supervision Data. In *AKBC@ NIPS*.

Xiaocheng Feng, Jiang Guo, Bing Qin, Ting Liu, and Yongjie Liu. 2017. Effective Deep Memory Networks for Distant Supervised Relation Extraction. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, Carles Sierra (Ed.). ijcai.org, 4002–4008. DOI:http://dx.doi.org/10.24963/ijcai.2017/559

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.