

# Crowdsourcing Semantic Label Propagation in Relation Classification

**Anca Dumitrache**

Vrije Universiteit Amsterdam

IBM CAS Benelux

anca.dmrch@gmail.com

**Lora Aroyo**

Vrije Universiteit Amsterdam

l.m.aroy@gmail.com

**Chris Welty**

Google Research, New York

cawelty@gmail.com

## Abstract

Distant supervision is a popular method for performing relation extraction from text that is known to produce noisy labels. Most progress in relation extraction and classification has been made with crowdsourced corrections to distant-supervised labels, and there is evidence that indicates still more would be better. In this paper, we explore the problem of propagating human annotation signals gathered for open-domain relation classification through the CrowdTruth methodology for crowdsourcing, that captures ambiguity in annotations by measuring inter-annotator disagreement. Our approach propagates annotations to sentences that are similar in a low dimensional embedding space, expanding the number of labels by two orders of magnitude. Our experiments show significant improvement in a sentence-level multi-class relation classifier.

## 1 Introduction

Distant supervision (DS) (Mintz et al., 2009) is a popular method for performing relation extraction from text. It is based on the assumption that, when a knowledge-base 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 expressed (Feng et al., 2017).

Recent results (Angeli et al., 2014; Liu et al., 2016) have shown strong evidence that the community needs more annotated data to improve the quality of DS data. This work explores the possibility of automatically expanding smaller human-annotated datasets to DS scale. Sterckx et al. (2016) proposed a method to correct labels of sentence dependency paths by using expert annotators, and then propagating the corrected labels to a

corpus of DS sentences by calculating the similarity between the labeled and unlabeled sentences in the embedding space of their dependency paths.

In this paper, we adapt and simplify semantic label propagation to propagate labels without computing dependency paths, and using the crowd instead of experts, which is more scalable. Our simplified algorithm propagates crowdsourced annotations from a small sample of sentences to a large DS corpus. To evaluate our approach, we perform an experiment in open domain relation classification in the English-language, using a corpus of sentences (Dumitrache et al., 2017) whose labels have been collected using the CrowdTruth method (Aroyo and Welty, 2014).

## 2 Related Work

There exist several efforts to correct DS with the help of crowdsourcing. 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. Zhang et al. (2012) show that labeled data has a statistically significant, but relatively low impact on improving the quality of DS training data, while increasing the size of the DS corpus has a more significant impact. In contrast, Liu et al. (2016) prove that a corpus of labeled sentences from a pool of highly qualified workers can significantly improve DS quality. All of these methods employ large annotated corpora of 10,000 to 20,000 sentences. In our experiment, we show that a comparatively smaller corpus of 2,050 sentences is enough to correct DS errors through semantic label propagation.

Levy et al. (2017) have shown that a small crowdsourced dataset of questions about relations can be exploited to perform zero-shot learning for relation extraction. Pershina et al. (2014) use a small dataset of hand-labeled data to generate relation-specific guidelines that are used as addi-

Figure 1: Fragment of the crowdsourcing task template.

The sentence:

“ A failure to follow through in Geneva and deliver the results we need would represent nothing short of political failure , ” **NEW ZEALAND Prime Minister JOHN KEY** said .

STEP 1: Select ALL THE STATEMENTS between the terms **JOHN KEY** and **NEW ZEALAND** that are expressed in the sentence above. (required)

- |  |  |
|--|--|
| <input type="checkbox"/> <b>JOHN KEY</b> is an organization with the alternate name <b>NEW ZEALAND</b> | <input type="checkbox"/> headquarters of <b>JOHN KEY</b> are/were located in <b>NEW ZEALAND</b>        |
| <input type="checkbox"/> <b>NEW ZEALAND</b> is/was a subsidiary of <b>JOHN KEY</b>                     | <input type="checkbox"/> <b>JOHN KEY</b> is/was a member/employee of <b>NEW ZEALAND</b>                |
| <input type="checkbox"/> <b>NEW ZEALAND</b> was founded by <b>JOHN KEY</b>                             | <input checked="" type="checkbox"/> <b>JOHN KEY</b> is/was a top member/employee of <b>NEW ZEALAND</b> |
| <input type="checkbox"/> <b>JOHN KEY</b> is a person with the alternate name <b>NEW ZEALAND</b>        | <input type="checkbox"/> <b>JOHN KEY</b> died because of <b>NEW ZEALAND</b>                            |
| <input type="checkbox"/> <b>JOHN KEY</b> is/was charged with <b>NEW ZEALAND</b>                        | <input type="checkbox"/> <b>JOHN KEY</b> is the father/mother of <b>NEW ZEALAND</b>                    |
| <input type="checkbox"/> <b>JOHN KEY</b> is a person who lives/lived in <b>NEW ZEALAND</b>             | <input type="checkbox"/> <b>JOHN KEY</b> is a person who is/was born in <b>NEW ZEALAND</b>             |
| <input type="checkbox"/> <b>JOHN KEY</b> is a person who died in <b>NEW ZEALAND</b>                    | <input type="checkbox"/> <b>JOHN KEY</b> attended school(s) <b>NEW ZEALAND</b>                         |
| <input type="checkbox"/> <b>JOHN KEY</b> is a person originating from <b>NEW ZEALAND</b>               | <input type="checkbox"/> <b>JOHN KEY</b> is/was married to <b>NEW ZEALAND</b>                          |
| <input type="checkbox"/> <b>JOHN KEY</b> is a person with the title of <b>NEW ZEALAND</b>              | <input type="checkbox"/> none of these   |

It is important that you understand what the different statements mean. Carefully read the EXAMPLE by hovering over each statement.

tional features in the relation extraction. The label propagation method was introduced by Xiaojin and Zoubin (2002), while Chen et al. (2006) first applied it to correct DS, by calculating similarity between labeled and unlabeled examples an extensive list of features, including part-of-speech tags and target entity types. In contrast, our approach calculates similarity between examples in the word2vec (Mikolov et al., 2013) feature space, which it then uses to correct the labels of training sentences. This makes it easy to reuse by the state-of-the-art in both relation classification and relation extraction – convolutional (Ji et al., 2017) and recurrent neural network methods (Zhou et al., 2016) that do not use extensive feature sets. To evaluate our approach, we used a simple convolutional neural network to perform relation classification in sentences (Nguyen and Grishman, 2015).

### 3 Experimental Setup

#### 3.1 Annotated Data

The labeled data used in our experiments consists of 4,100 sentences: 2,050 sentences from the CrowdTruth corpus (Dumitrache et al., 2017), which we have augmented by another 2,050 sentences picked at random from the corpus of Angeli et al. (2014). The resulting corpus contains sentences for 16 popular relations from the open domain, as shown in in Figure 1,<sup>1</sup> as well as candidate term pairs and DS seed relations for each sentence. As some relations are more general than others, the relation fre-

quency in the corpus is slightly unequal – e.g. *places\_of\_residence* is more likely to be in a sentence when *place\_of\_birth* and *place\_of\_death* occur, but not the opposite.

The crowdsourcing task (Figure 1) was designed in our previous work (Dumitrache et al., 2017). We asked workers to read the given sentence where the candidate term pair is highlighted, and then pick between the 16 relations or *none of the above*, if none of the presented relations apply. The task was multiple choice and run on the Figure Eight<sup>2</sup> and Amazon Mechanical Turk<sup>3</sup> crowdsourcing platforms. Each sentence was judged by 15 workers, and each worker was paid \$0.05 per sentence.

Crowdsourcing annotations are aggregated usually by measuring the consensus of the workers (e.g. using majority vote). This is based on the assumption that a single right annotation exists for each example. In the problem of relation classification, the notion of a single truth is reflected in the fact that a majority of proposed solutions treat relations as mutually exclusive, and the objective of the classification task is usually to find the best relation for a given sentence and term pair. In contrast, the CrowdTruth methodology proposes that crowd annotations are inherently diverse (Aroyo and Welty, 2015), due to a variety of factors such as the ambiguity that is inherent in natural language. We use a comparatively large number of workers per sentences (15) in order to collect inter-annotator disagreement, which results in a more

<sup>1</sup>The *alternate\_names* relation appears twice in the list, once referring to alternate names of persons, and the other referring to organizations.

<sup>2</sup><https://www.figure-eight.com/>

<sup>3</sup><https://www.mturk.com/>

fine-grained ground truth that separates between clear and ambiguous expressions of relations. This is achieved by labeling examples with the inter-annotator agreement on a continuous scale, as opposed to using binary labels.

To aggregate the results of the crowd, we use CrowdTruth metrics<sup>4</sup> (Dumitrache et al., 2018) to capture and interpret inter-annotator disagreement as quality metrics for the workers, sentences, and relations in the corpus. The annotations of one worker over one sentence are encoded as a binary worker vector with 17 components, one for each relation and including *none*. The quality metrics for the workers, sentences and relations, are based on average cosine similarity over the worker vectors – e.g. the quality of a worker  $w$  is given by the average cosine similarity between the worker vector of  $w$  and the vectors of all other workers that annotated the same sentences. These metrics are mutually dependent (e.g. the sentence quality is weighted by the relation quality and worker quality), the intuition being that low quality workers should not count as much in determining sentence quality, and ambiguous sentences should have less of an impact in determining worker quality, etc.

We reused these scores in our experiment, focusing on the **sentence-relation score** ( $srs$ ), representing the degree to which a relation is expressed in the sentence. It is the ratio of workers that picked the relation to all the workers that read the sentence, weighted by the worker and relation quality. A higher  $srs$  should indicate that the relation is more clearly expressed in a sentence.

### 3.2 Propagating Annotations

Inspired by the semantic label propagation method (Sterckx et al., 2016), we propagate the vectors of  $srs$  scores on each crowd annotated sentence to a much larger set of distant supervised (DS) sentences (see datasets description in Section ), scaling the vectors linearly by the distance in low dimensional word2vec vector space (Mikolov et al., 2013). One of the reasons we chose the CrowdTruth set for this experiment is that the annotation vectors give us a score *for each relation* to propagate to the DS sentences, which have only one binary label.

Similarly to Sultan et al. (2015), we calculate the vector representation of a sentence as the average over its word vectors, and like Sterckx

et al. (2016) we get the similarity between sentences using cosine similarity. Additionally, we restrict the sentence representation to only contain the words between the term pair, in order to reduce the vector space to the one that is most likely to express the relations. For each sentence  $s$  in the DS dataset, we find the sentence  $l'$  from the crowd annotated set that is most similar to  $s$ :  $l' = \arg \max_{l \in \text{Crowd}} \cos\_sim(l, s)$ . The score for relation  $r$  of sentence  $s$  is calculated as the weighted average between the  $srs(l', r)$  and the original DS annotation, weighted by the cosine similarity to  $s$  ( $\cos\_sim(s, s) = 1$  for the DS term, and  $\cos\_sim(s, l')$  for the  $srs$  term):

$$DS^*(s, r) = \frac{DS(s, r) + \cos\_sim(s, l') \cdot srs(l', r)}{1 + \cos\_sim(s, l')} \quad (1)$$

where  $DS(s, r) \in \{0, 1\}$  is the original DS annotation for the relation  $r$  on sentence  $s$ .

### 3.3 Training the Model

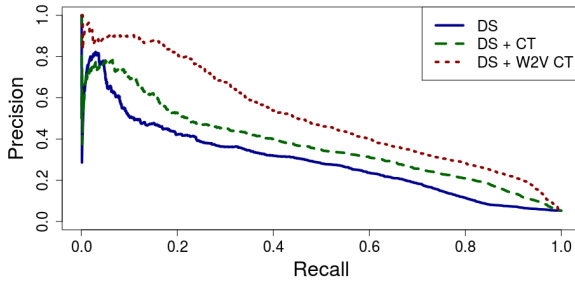
The crowdsourced data is split evenly into a dev and a test set of 2,050 sentences each chosen at random. In addition, we used a training set of 235,000 sentences annotated by DS from freebase relations, used in Riedel et al. (2013).

The relation classification model employed is based on Nguyen and Grishman (2015), who implement a convolutional neural network with four main layers: an embedding layer for the words in the sentence and the position of the candidate term pair in the sentence, a convolutional layer with a sliding window of variable length of 2 to 5 words that recognizes n-grams, a pooling layer that determines the most relevant features, and a softmax layer to perform classification.

We have adapted this model to be both multi-class and multi-label – we use a sigmoid cross-entropy loss function instead of softmax cross-entropy, and the final layer is normalized with the sigmoid function instead of softmax – in order to make it possible for more than one relation to hold between two terms in one sentence. The loss function is computed using continuous labels instead of binary positive/negative labels, in order to accommodate the use of the  $srs$  in training. The features of the model are the word2vec embeddings of the words in the sentences, together with the position embeddings of the two terms that express the relation. The word embeddings are ini-

<sup>4</sup><https://github.com/CrowdTruth/CrowdTruth-core>

Figure 2: Precision / Recall curve, calculated for each sentence-relation pair.



tialized with 300-dimensional word2vec vectors pre-trained on the Google News corpus<sup>5</sup>. Both the position and word embeddings are nonstatic and become optimized during training of the model. The model is trained for 25,000 iterations, after the point of stabilization for the train loss. The values of the other hyper-parameters are the same as those reported by Nguyen and Grishman (2015). The model was implemented in Tensorflow (Abadi et al., 2016), and trained in a distributed manner on the DAS-5 cluster (Bal et al., 2016).

For our experiment, we split the crowd data into a dev and a test set of equal size, and compared the performance of the model on the held-out test set when trained by the following datasets:

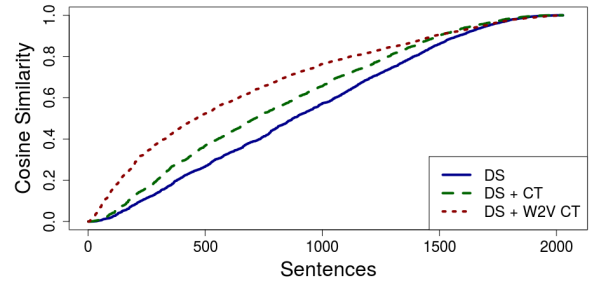
1. **DS**: The 235,000 sentences annotated by DS.
2. **DS + CT**: The 2,050 crowd dev annotated sentences added directly to the DS dataset.
3. **DS + W2V CT**: The DS\* dataset (Eq. 1), with relation scores propagated over the 2,050 crowd dev sentences.

## 4 Results and Discussion

To evaluate the performance of the models, we calculate the micro precision and recall (Figure 2), as well as the cosine similarity per sentence with the test set (Figure 3). In order to calculate the precision and recall, a threshold of 0.5 was set in the *srs*, and each sentence-relation pair was labeled either as positive or negative. However, for calculating the cosine similarity, the *srs* was used without change, in order to better reflect the degree of agreement the crowd had over annotating each example. We observe that **DS + W2V CT**, with a

<sup>5</sup><https://code.google.com/archive/p/word2vec/>

Figure 3: Distribution of sentence-level cosine similarity with test set values.



precision/recall  $AUC = 0.512$ , significantly outperforms **DS** (P/R  $AUC = 0.294$ ). **DS + CT** (P/R  $AUC = 0.372$ ) also does slightly better than **DS**, but not enough to compete with the semantic label propagation method. The cosine similarity result (Figure 3) shows that **DS + W2V CT** also produces model predictions that are closer to the different agreement levels of the crowd. Take advantage of the agreement scores in the CrowdTruth corpus, the cosine similarity evaluation allows us to assess relation confidence scores on a continuous scale. The crowdsourcing results and model predictions are available online.<sup>6</sup>

One reason for which the semantic label propagation method works better than simply adding the correctly labeled sentences to the train set is the high rate of incorrectly labeled examples in the DS training data. Figure 4 shows that some relations, such as *origin* and *places\_of\_residence*, have a ratio of over 0.8 false positive sentences, meaning that a vast majority of training examples are incorrectly labeled. The success of the **DS + W2V CT** comes in part because the method relabels all sentences in DS. Adding correctly labeled sentences to the train set would require a significantly larger corpus in order to correct the high false positive rate, but semantic label propagation only requires a small corpus (two orders of magnitude smaller than the train set) to achieve significant improvements.

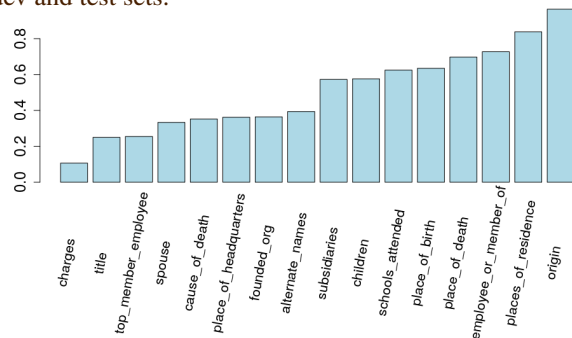
## 5 Conclusion and Future Work

This paper explores the problem of propagating human annotation signals in distant supervision data for open-domain relation classification. Our approach propagates human annotations to sentences that are similar in a low dimensional em-

<sup>6</sup><https://github.com/CrowdTruth/Open-Domain-Relation-Extraction>



Figure 4: DS false positive ratio in combined crowd dev and test sets.



bedding space, using a small crowdsourced dataset of 2,050 sentences to correct training data labeled with distant supervision. We present experimental results from training a relation classifier, where our method shows significant improvement over the DS baseline, as well as just adding the labeled examples to the train set.

Unlike Sterckx et al. (2016) who employ experts to label the dependency path representation of sentences, our method uses the general crowd to annotate the actual sentence text, and is thus easier to scale and not dependent on methods for extracting dependency paths, so it can be more easily adapted to other languages and domains. Also, since the semantic label propagation is applied to the data before training is completed, this method can easily be reused to correct train data for any model, regardless of the features used in learning. In our future work, we plan to use this method to correct training data for state-of-the-art models in relation classification, but also relation extraction and knowledge-base population.

We also plan to explore different ways of collecting and aggregating data from the crowd. CrowdTruth (Dumitrache et al., 2017) proposes capturing ambiguity through inter-annotator disagreement, which necessitates multiple annotators per sentence, while Liu et al. (2016) propose increasing the number of labeled examples added to the training set by using one high quality worker per sentence. We will compare the two methods to determine whether quality or quantity of data are more useful for semantic label propagation. To achieve this, we will investigate whether disagreement-based metrics such as sentence and relation quality can also be propagated through the training data.

## References

- Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283.
- Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D Manning. 2014. Combining distant and partial supervision for relation extraction. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1556–1567.
- Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *Journal of Human Computation*, 1:31–34.
- Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.
- Henri Bal, Dick Epema, Cees de Laat, Rob van Nieuwpoort, John Romein, Frank Seinstra, Cees Snoek, and Harry Wijshoff. 2016. A medium-scale distributed system for computer science research: Infrastructure for the long term. *Computer*, 49(5):54–63.
- Jinxu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. 2006. Relation extraction using label propagation based semi-supervised learning. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-44*, pages 129–136, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Anca Dumitrache, Lora Aroyo, and Chris Welty. 2017. False positive and cross-relation signals in distant supervision data. In *Proceedings of the 6th Workshop on Automated Knowledge Base Construction*.
- Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. *arXiv preprint arXiv:1808.06080*.
- 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*, pages 4002–4008. ijcai.org.
- Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao, et al. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *AAAI*, pages 3060–3066.
- Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *CoNLL 2017*, page 333.

- Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H Lin, Xiao Ling, and Daniel S Weld. 2016. Effective crowd annotation for relation extraction. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 897–906.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 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-Volume 2*, pages 1003–1011. Association for Computational Linguistics.
- Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48.
- Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. 2014. Infusion of labeled data into distant supervision for relation extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 732–738.
- Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84.
- Lucas Sterckx, Thomas Demeester, Johannes Deleu, and Chris Develder. 2016. Knowledge base population using semantic label propagation. *Knowledge-Based Systems*, 108(C):79–91.
- Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. DLS @ CU: Sentence Similarity from Word Alignment and Semantic Vector Composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 148–153.
- Zhu Xiaojin and Ghahramani Zoubin. 2002. Learning from labeled and unlabeled data with label propagation. *Technical Report CMU-CALD-02-107*, Carnegie Mellon University.
- Ce Zhang, Feng Niu, Christopher Ré, and Jude Shavlik. 2012. Big data versus the crowd: Looking for relationships in all the right places. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 825–834. Association for Computational Linguistics.
- Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 207–212.