Self-Crowdsourcing Training for Relation Extraction

Azad Abad[†], **Moin Nabi**[†], **Alessandro Moschitti**
†DISI, University of Trento, 38123 Povo (TN), Italy
Qatar Computing Research Institute, HBKU, 34110, Doha, Qatar
{azad.abad,moin.nabi}@unitn.it
amoschitti@gmail.com

Abstract

One expensive step when defining crowdsourcing tasks is to define the examples and control questions for instructing the crowd workers. In this paper, we introduce a self-training strategy for crowdsourcing. The main idea is to use an automatic classifier, trained on weakly supervised data, to select examples associated with high confidence. These are used by our automatic agent to explain the task to crowd workers with a question answering approach. We compared our relation extraction system trained with data annotated (i) with distant supervision and (ii) by workers instructed with our approach. The analysis shows that our method relatively improves the relation extraction system by about 11% in F1.

1 Introduction

Recently, the Relation Extraction (RE) task has attracted the attention of many researchers due to its wide range of applications such as question answering, text summarization and bio-medical text mining. The aim of this task is to identify the type of relation between two entities in a given text. Most work on RE has mainly regarded the application of supervised methods, which require costly annotation, especially for large-scale datasets.

To overcome the annotation problem, Craven et al. (1999) firstly proposed to collect automatic annotation through Distant Supervision (DS). In the DS setting, the training data for RE is often automatically annotated utilizing an external Knowledge-Base (KB) such as Wikipedia or Freebase (Hoffmann et al., 2010; Riedel et al., 2010; Nguyen and Moschitti, 2011). Although DS has shown to be promising for RE, it also produces many noisy labels in the automatic annotated data, which deteriorate the performance of the system trained on it.

Hoffmann et al. (2011) showed that by simply adding a small set of high quality labeled instances (i.e., human-annotated training data) to a larger set of instances annotated by DS, makes the overall precision of the system significantly increases. Such level of quality of the labels usually can be obtained at low cost via crowdsourcing.

However, this finding does not hold for more complex tasks, where the annotators[1] need to have some expertise on them. For instance in RE, several works have shown that only a marginal improvement can be achieved via crowdsourcing the data (Angeli et al., 2014; Zhang et al., 2012; Pershina et al., 2014). In such papers, the well-known Gold Standard quality control mechanism was used without annotators being trained.

Very recently, despite the previous results, Liu et al. (2016) showed a larger improvement for the RE task when training crowd workers in an interactive tutorial procedure called "Gated Instruction". This approach, however, requires a set of high-quality labeled data (i.e., the Gold Standard) for providing the instruction and feedback to the crowd workers. However, acquiring such data requires a considerable amount of human effort.

In this paper, we propose to alternatively use *Silver Standard*, i.e., a high-quality automatic annotated data, to train the crowd workers. Specifically, we introduce a self-training strategy for crowd-sourcing, where the workers are first trained with simpler examples (which we assume to be less noisy) and then gradually presented with more difficult ones. This is biologically inspired by the common human process of gradual learn-

[1] From now, the both entities *annotators* and *crowd workers* refer to the same concept.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 518–523
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2082

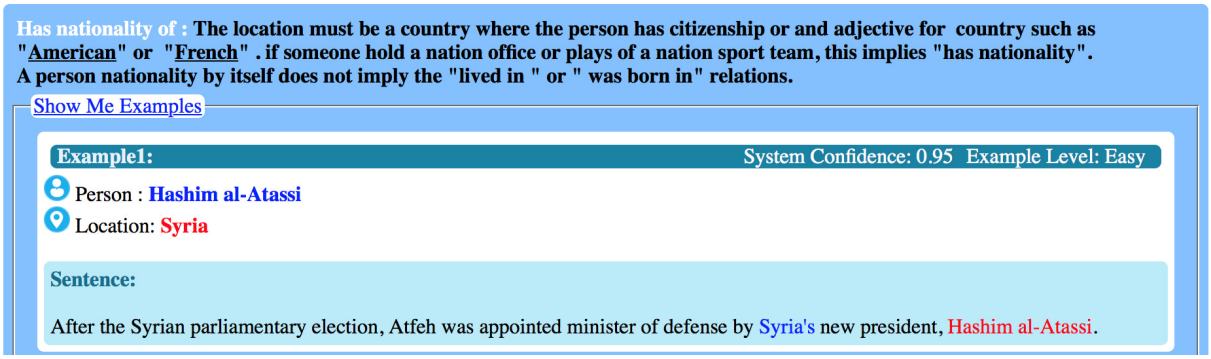

Figure 1: User Interface of crowd worker training: instruction phase

ing, starting from the simplest concepts.

Moreover, we propose an iterative human-machine co-training framework for the task of RE. The main idea is (i) to automatically select a subset of *less-noisy* examples applying an automatic classifier, (ii) training the annotators with such subset, and (iii) iterating this process after retraining the classifiers using the annotated data. That is, the educated crowd workers can provide higher quality annotations, which can be used by the system in the next iteration to improve the quality of its classification. In other words, this cycle gradually improves both system and human annotators. This is in line with the studies in human-based computational approaches, which showed that the crowd intelligence can effectively alleviate the drifting problem in auto-annotation systems (Sun et al., 2014; Russakovsky et al., 2015).

Our study shows that even without using any gold standard, we can still train workers and their annotations can achieve results comparable with the more costly state-of-the-art methods. In summary our contributions are the following:

- we introduce a self-training strategy for crowdsourcing;
- we propose an iterative human-machine co-training framework for the task of RE; and
- we test our approach on a standard benchmark, obtaining a slightly lower performance compared to the state-of-the-art methods based on Gold Standard data.

This study opens up avenues for exploiting inexpensive crowdsourcing solutions similar to ours to achieve performance gain in NLP tasks.

2 Background Work

There is a large body of work on DS for RE, but we only discuss the most related to our work and refer the reader to other recent work (Wu and Weld, 2007; Mintz et al., 2009; Bunescu, 2007; Hoffmann et al., 2010; Riedel et al., 2010; Surdeanu et al., 2012; Nguyen and Moschitti, 2011).

Many researchers have exploited the techniques of combining the DS data with small human annotated data collected via crowdsourcing, to improve the relation extractor accuracy (Liu et al., 2016; Angeli et al., 2014; Zhang et al., 2012). Angeli et al. (2014) reported a minor improvement using active learning methods to select the best instances to be crowdsourced.

In the same direction, Zhang et al. (2012) studied the effect of providing human feedback in crowdsourcing tasks and observed a minor improvement in terms of F1. At high level, our work may be viewed as employing crowdsourcing for RE. In that spirit, we are similar to these works, but with the main difference of training crowd workers to obtain higher quality annotations.

The most related paper to our work is by Liu et al. (2016), who trained the crowd workers via "Gated Instruction". They also showed that collecting higher-quality annotations can be achieved through training the workers. The produced data also improved the performance of the RE systems trained on it. Our study confirms their finding. However, unlike them, we do not employ any Gold Standard (annotated by experts) for training the annotators and instead we propose a self-training strategy to select a set of high-quality automatic annotated data (namely, Silver Standard).

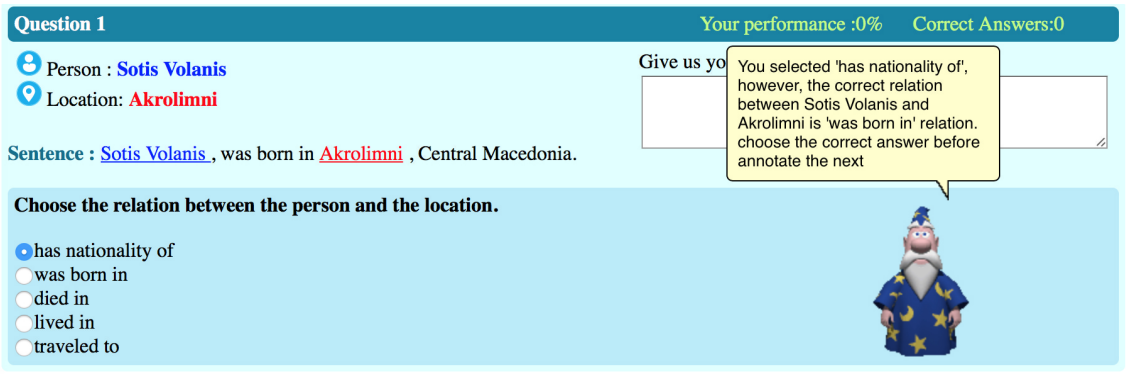

Figure 2: User Interface of crowd worker training: interactive QA phase

3 Self-Crowdsourcing Training

In this section we first explain, our proposed method for automatically identifying high-quality examples (i.e., Silver Standard) to train the crowd workers and collect annotations for the lower-quality examples. We then explain the scheme designed for crowd worker training and annotation collection.

3.1 Silver Standard Mining

The main idea of our approach to Self-Crowdsourcing training is to use the classifier's score for gradually training the crowd workers, such that the examples and labels associated with the highest prediction values (i.e., the most reliable) will be used as Silver Standard.

More in detail, our approach is based on a noisy-label dataset, DS, whose labels are extracted in a distant supervision fashion and CS a dataset to be labeled by the crowd. The first step is to divide CS into three parts: CS_I, which is used to create the instructions for the crowd workers; CS_Q, which is used for asking questions about sentence annotations; and CS_A, which is used to collect the labels from annotators, after they have been trained.

To select CS_I, we train a classifier C on DS, and then used it to label CS examples. In particular, we used MultiR framework (Hoffmann et al., 2011) to train C, as it is a widely used framework for RE. Then, we sort CS in a descending order according to the classifier prediction scores and select the first N_i elements, obtaining CS_I.

Next, we select the N_q examples of $CS \setminus CS_I$ with the highest score to create the set CS_Q. Note that the latter contains highly-reliable classifier annotations but since the scores are lower than for CS_I examples, we conjecture that they may be more difficult to be annotated by the crowd workers.

Finally, CS_A is assigned with the remaining examples, i.e., $CS \setminus CS_I \setminus CS_Q$. These have the lowest confidence and should therefore be annotated by crowd workers. N_i and N_q can be tuned on the task, we set both to 10% of the data.

3.2 Training Schema

We conducted *crowd worker training* and *annotation collection* using the well-known Crowdflower platform[2]. Given CS_I and CS_Q (see Section 3.1), we train the annotators in two steps:

(i) **User Instruction**: first, a definition of each relation type (borrowed from TAC-KBP official guideline) is shown to the annotators. This initial training step provides the crowd workers with a big picture of the task. We then train the annotators showing them a set of examples from CS_I (see Fig. 1). The latter are presented in the reverse order of difficulty. The ranked list of examples provided by our self-training strategy facilitates the gradual education of the annotators (Nosofsky, 2011). This gives us the benefit of training the annotators with any level of expertise, which is a crucial property of crowdsourcing when there is no clue about the workers' expertise in advance.

(ii) **Interactive QA**: after the initial step, we challenge the workers in an interactive QA task with multiple-choice questions over the sentence annotation (see Fig. 2). To accomplish that, we designed an artificial agent that interacts with the crowd workers: it corrects their mistakes and makes them reasoning on why their answer was wrong. Note that, to have a better control of the

[2]www.crowdflower.com

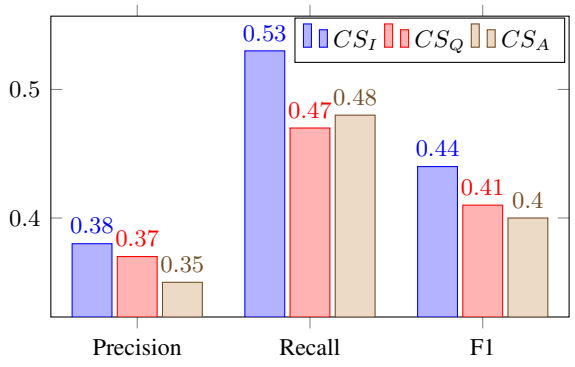

Figure 3: Accuracy of different CS partitions

worker training, we perform a selection of the sentences in CS_Q to be used for questioning in a category-wise fashion. Meaning that, we select the subsets of examples for each class of relation separately. We observed in practice that initially a lot of examples are classified as "No Relation". This is due to a difficulty of the task for the DS-based model. Thus, we used them in CS_A.

4 Experimental Setup

In this section, we first introduce the details of the used corpora, then explain the feature extraction and RE pipeline and finally present the experiments and discuss the results in detail.

4.1 Corpora

We used TAC-KBP newswires, one of the most well-known corpora for RE task. As DS, we selected 700K sentences automatically annotated using Freebase as an external KB. We used the active learning framework proposed by Angeli et al. (2014) to select CS. This allowed us to select the best sentences to be annotated by humans (sampleJS). As a result, we obtained 4,388 sentences. We divided the CS sentences in CS_I, CS_Q and CS_A, with 10%, 10% and 80% split, respectively. We requested at least 5 annotations for each sentence.

Similarly to Liu et al. (2016), we restricted our attention to 5 relations between *person* and *location*[3]. For both DS and CS, we used the publicly available data provided by Liu et al. (2016). Ultimately, 221 crowd workers participated to the task with minimum 2 and maximum 400 annotations per crowd worker. To evaluate our model, we randomly selected 200 sentences as test set and had a domain expert to manually tag them using the TAC-KBP annotation guidelines.

[3]*Nationality, Place-of-birth, Place-of-resident, Place-of-death, Traveled-to*

Model	Pr.	Rec.	F1
DS-only	0.43	0.52	0.47
Our Method	0.50	0.54	0.52
Gated Instruction	0.53	0.57	0.55

Table 1: Evaluation of the impact of the CS_A label quality in the RE task.

4.2 Relation Extraction Pipeline

We used the relation extractor, MultiR (Hoffmann et al., 2010) along with lexical and syntactic features proposed by Mintz et al. (2009) such as: (i) Part of Speech (POS); (ii) windows of k words around the matched entities; (iii) the sequences of words between them; and (iv) finally, dependency structure patterns between entity pairs. These yield low-Recall as they appear in conjunctive forms but at the same time they produce a high Precision.

4.3 Experimental Results

In the first set of experiments, we verified the quality of our Silver Standard set used in our self-training methods. For this purpose, we trained MultiR on CS_I, CS_Q and CS_A and evaluate them on our test set. Figure 3 illustrates the results in terms of Precision, Recall and F1 for each partition separately. They suggest that, the extractors trained on CS_I and CS_Q are significantly better than the extractor trained on the lower part of CS, i.e., CS_A, even if the latter is much larger than the other two (80% vs. 10%).

In the next set of experiments, we evaluated the impact of adding a small set of crowdsourced data to a large set of instances annotated by Distant Supervision. We conducted the RE experiments in this setting, as this allowed us to directly compare with Liu et al. (2016). Thus, we used CS_A annotated by our proposed method along with the noisy annotated DS to train the extractor.

We compared our method with (i) the *DS-only* baseline and (ii) the state of the art, *Gated Instruction* (GI) strategy (Liu et al., 2016). We emphasize that the same set of examples (both DS and CS) are used in this experiment and just replaced the GI annotations with the annotations collected using our proposed framework.

Models	DS-only	Our Model	GI
Accuracy	56%	82%	91%

Table 2: Annotation Accuracy of crowd workers

The results are shown in Table 1. Our method improves the *DS-only* baseline by 7%, 5% and 2% (absolute) in Precision, Recall and F1, respectively. This improvement clearly confirms the benefit of our fully automatic approach to crowdsourcing in RE task.

Additionally, our model is just 3% lower than the GI method in terms of F1. In both our method and GI, the crowd workers are trained before enrolling in the main task. However, GI trains annotators using Gold Standard data, which involves a higher level of supervision with respect to our method. Thus our self-training method is potentially effective and an inexpensive alternative to GI.

We also analyzed the accuracy of the crowd workers in terms of the quality of their annotations. For this purpose, we randomly selected 100 sentences from CS_A and then had them manually annotated by an expert. We compared the accuracy of the annotations collected with our proposed approach with those provided by DS-only baseline and the GI method. Table 2 shows the results: the annotations performed by workers trained with our method are just slightly less accurate than the annotations produced by annotators trained with GI. This outcome is inline with the positive impact of our good quality annotation on the RE performance.

5 Conclusion

In this paper, we have proposed a self-training strategy for crowdsourcing as an effective alternative to train annotators with Gold Standard. Our experimental results show that the annotations of workers trained with our method are accurate and produce a good performance when used in learning algorithms for RE. Our study suggests that automatically training annotators can replace the popular consensus-based filtering scheme. Our method achieves this goal through an inexpensive training procedure.

In the future, it would be interesting to study if our method generalizes to other difficult or even simpler tasks. In particular, our approach opens up many research directions on how to best train workers or best select data for them, similarly to what active learning methods have been doing for training machines.

Acknowledgement

This work has been partially supported by the EC project CogNet, 671625 (H2020-ICT-2014-2, Research and Innovation action). Many thanks to the anonymous reviewers for their valuable suggestions.

References

Gabor Angeli, Julie Tibshirani, Jean Y. Wu, and Christopher D. Manning. 2014. Combining distant and partial supervision for relation extraction. In *In Proceedings of EMNLP*. pages 1556–1567.

Razvan C. Bunescu. 2007. Learning to extract relations from the web using minimal supervision. In *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL07)*.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, pages 77–86. http://dl.acm.org/citation.cfm?id=645634.663209.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 541–550. http://dl.acm.org/citation.cfm?id=2002472.2002541.

Raphael Hoffmann, Congle Zhang, and Daniel S. Weld. 2010. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '10, pages 286–295. http://dl.acm.org/citation.cfm?id=1858681.1858711.

Angli Liu, Jonathan Bragg Xiao Ling Stephen Soderland, and Daniel S Weld. 2016. Effective crowd annotation for relation extraction. In *Association for Computational Linguistics*. NAACL-HLT 2016.

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. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '09, pages 1003–1011. http://dl.acm.org/citation.cfm?id=1690219.1690287.

Truc-Vien T. Nguyen and Alessandro Moschitti. 2011. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 277–282. http://dl.acm.org/citation.cfm?id=2002736.2002794.

Robert M Nosofsky. 2011. The generalized context model: An exemplar model of classification. *Formal approaches in categorization* pages 18–39.

Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. 2014. Infusion of labeled data into distant supervision for relation extraction. In *Proceedings of the 2014 Conference of the Association for Computational Linguistics (ACL 2014)*. Association for Computational Linguistics, Baltimore, US.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*. Springer-Verlag, Berlin, Heidelberg, ECML PKDD'10, pages 148–163. http://dl.acm.org/citation.cfm?id=1889788.1889799.

Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. 2015. Best of both worlds: Human-machine collaboration for object annotation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chong Sun, Narasimhan Rampalli, Frank Yang, and AnHai Doan. 2014. Chimera: Large-scale classification using machine learning, rules, and crowdsourcing. *Proc. VLDB Endow.* 7(13):1529–1540. https://doi.org/10.14778/2733004.2733024.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP-CoNLL '12, pages 455–465. http://dl.acm.org/citation.cfm?id=2390948.2391003.

Fei Wu and Daniel S. Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. ACM, New York, NY, USA, CIKM '07, pages 41–50. https://doi.org/10.1145/1321440.1321449.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '12, pages 825–834. http://dl.acm.org/citation.cfm?id=2390524.2390640.