# Held-out versus Gold Standard: Comparison of Evaluation Strategies for Distantly Supervised Relation Extraction from Medline abstracts

**Roland Roller** and **Mark Stevenson**
Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello
S1 4DP Sheffield, England
roland.roller,mark.stevenson@sheffield.ac.uk

## Abstract

Distant supervision is a useful technique for creating relation classifiers in the absence of labelled data. The approaches are often evaluated using a held-out portion of the distantly labelled data, thereby avoiding the need for lablelled data entirely. However, held-out evaluation means that systems are tested against noisy data, making it difficult to determine their true accuracy. This paper examines the effectiveness of using held-out data to evaluate relation extraction systems by comparing the results that are produced with those generated using manually labelled versions of the same data. We train classifiers to detect two UMLS Metathesaurus relations (*may-treat* and *may-prevent*) in Medline abstracts. A new evaluation data set for these relations is made available. We show that evaluation against a distantly labelled gold standard tends to overestimate performance and that no direct connection can be found between improved performance against distantly and manually labelled gold standards.

## 1 Introduction

Relation extraction is a popular topic in the biomedical domain and has been the subject of several challenges (e.g. DDI challenge (Segura-Bedmar et al., 2013), BioNLP Shared Task (Nédellec et al., 2013)). Many approaches rely on supervised learning techniques using manually labelled training data. However, the creation of annotated training data is time-consuming, expensive and often requires expert knowledge.

Distant supervision (self-supervised learning) is a widely applied technique for training relation extraction systems (Wu and Weld, 2007; Krause et al., 2012; Roth and Klakow, 2013; Ritter et al., 2013; Vlachos and Clark, 2014) that avoids the need for annotated training data. Training examples are annotated automatically using a knowledge base. Facts from the knowledge base are matched against text and used as training examples. For example, a knowledge base may assert that the entity pair *CONDITION("hair loss")-DRUG("paroxetine")* is an instance of the relationship *adverse-drug effect*. Distant supervision approaches normally assume that sentences containing both entities assert the relation between them and, consequently, the following sentence would be used as a positive example of the *adverse-drug effect* relation:

> *"Findings on discontinuation and rechallenge supported the assumption that the **hair loss** was a side effect of the **paroxetine**." (PMID=10442258)*

However, this assumption does not always hold which can lead to sentences containing entity pairs being mistakenly identified as asserting a particular relation between them. For example, the following sentence contains the same entity pair but does not assert the *adverse-drug effect* relation:

> *"There are a few case reports on **hair loss** associated with tricyclic antidepressants and serotonin selective reuptake inhibitors (SSRIs), but none deal specifically with **paroxetine**." (PMID=10442258)*

Consequently, data annotated using distant supervision is noisy and unlikely to be of as high



*Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 97–102,
Lisbon, Portugal, 17 September 2015. ©2015 Association for Computational Linguistics.

quality as manually labelled data. Despite this distantly supervised relation extraction provides reasonable results compared to those based on supervised learning (see e.g. in (Thomas et al., 2011)).

Distant supervision allows relation extraction systems to be created without manually labelled data. However, this raises the issue of how such a system can be evaluated. Previous approaches have carried out evaluation using existing data sets labelled with examples of the target relation (Bellare and Mccallum, 2007; Nguyen and Moschitti, 2011; Min et al., 2013) or a similar relation (Thomas et al., 2011; Roller and Stevenson, 2014). However, in the majority of scenarios the best use for any labeled data available is as training data. Others, such as Craven and Kumlien (1999), generated their own gold standard to annotate relevant relations of their knowledge base. But the effort required to generate manually labelled evaluation data somewhat negates the benefit of reduced development time provided by distant supervision.

An alternative approach, which does not require any labelled data, is held-out evaluation. This approach splits facts from the knowledge base into two parts: one to generate distantly supervised training data and the other to generate distantly supervised evaluation data (Mintz et al., 2009; Riedel et al., 2010; Hoffmann et al., 2010; Roller et al., 2015).

This approach is often combined with a manual evaluation in which a subset of the predictions is selected to be examined in more detail. For example, Riedel et al. (2010) supplemented the held-out evaluation of their distant supervision approach for Freebase by selecting the top 1000 facts it predicted and evaluating them manually. Others such as Surdeanu et al. (2012) and Intxaurrondo et al. (2013) work with the same knowledge base and are able to re-use the manually labelled data generated by Riedel et al. (2010). However, this data is only available for some Freebase relations and evaluation data has to be generated for each new relation. Approaches such as Takamatsu et al. (2012), Zhang et al. (2013) and Augenstein et al. (2014) combine a held-out evaluation with a manual evaluation of a randomly chosen subset or the top-k predictions. This technique is a more reliable evaluation method but requires more effort including (potentially) domain knowledge and needs to be repeated for each version of the classifier.

Held-out evaluation using distantly labelled data is a simple and quick technique for estimating the accuracy of distantly supervised relation extraction systems. However, this evaluation data is noisy and it is unclear what effect this has on the accuracy of performance estimates.

The issue is explored in this paper by evaluating relation extraction systems for two biomedical relations using both manually and distantly labelled data. We automatically generate labelled held-out data and then carry out a manual annotation to allow direct comparison. A distantly supervised classifier is trained and evaluated on both data sets. Similar as in Xu et al. (2013) we show that a large portion of the labels generated by distant supervision for the two relations are incorrect. However we find that evaluating classifiers using held-out distantly supervised data tends to overestimate performance compared to manually labelled data and that improvements in performance observed in evaluation against distantly supervised data are not necessarily reflected in improved results when measured against manually labelled data. To the best of our knowledge this is the first direct comparison of evaluating distantly supervised classifiers against distantly and manually labelled gold standards. Analysis in previous work has been restricted to determining the true labels for a set of positively predicted labels.

The remainder of this paper is structured as follows. The next section 2 describes the creation of the distantly supervised data and a manually labelled subset. A comparison of the automatically and manually generated labels is carried out in Section 3. Sections 4 evaluates a relation extraction system using different data sets and compares the performance obtained. The paper concludes with section 5.

## 2 Data Generation

A large set of distantly labelled examples was generated (Section 2.1). A small portion of these were used as held-out test data. This data set was also manually annotated (Section 2.2).

### 2.1 Distant labelling

Distantly labelled examples are generated using the Unified Medical Language System (UMLS) Metathesaurus as a knowledge source. UMLS is a large biomedical knowledge base which contains information about millions of medical concepts and the relations between them, making it well

| | | distantly labelled (DL) | | | | | |
|---|---|---|---|---|---|---|---|
| | | may-treat | | | may-prevent | | |
| | | pos | neg | # | pos | neg | # |
| manually labelled (ML) | pos | **106** | 67 | 173 | **85** | 54 | 139 |
| | neg | 94 | **133** | 227 | 115 | **146** | 261 |
| | | 200 | 200 | | 200 | 200 | |

Table 1: Comparison of manual and distantly labelled annotations

suited for distant supervision. Two biomedical relations (*may-treat* and *may-prevent*) were selected from UMLS. These relations describe connections between a pharmacological substance (e.g. drug) and a disease. For example, the following sentence expresses a *may-prevent* relationship between the entities *fluoride* and *dental caries*:

> *"Although **fluoride** is clearly a major reason for the decline in the prevalence of **dental caries**, there are no studies of the incremental benefit of in-office fluoride treatments for low-risk patients exposed to fluoridated water and using fluoridated toothpaste." (PMID=10698247)*

Training data for the two relations was generated from approximately 1 million biomedical abstracts from Medline[1] annotated with UMLS concepts by MetaMap[2] (Aronson and Lang, 2010). Sentences containing concepts that are identified as being related in the UMLS's `MRREL` table were selected and used as positive examples.[3] Negative examples were generated using a closed word assumption: pairs of concepts that are not listed as being related in UMLS for a given relation are considered to be negative examples of that relation. Such pairs are generated by considering all possible pairs from a particular relation and creating new pairs from the set of entities.

## 2.2 Test Data

A set of 400 distantly labelled sentences were randomly selected for each relation to generate held-out test data. Although the distantly labelled data contains more negatively labelled sentences than positive ones, equal numbers of positive and negative examples (200 of each) are selected in order to ensure that a sufficient number of positive instances are included in the data set. The sentences in this data set were selected so that none of the instance pairs occur in the data used for training. We refer to this data set as **DL** (Distantly Labelled).

The DL data set was then manually annotated. Two annotators were recruited, both of whom were studying graduate degrees in subjects related to medicine at our institution. Given a sentence with a highlighted pharmacological substance and a highlighted disease, the annotators had to determine whether a sentence expresses the relationship of interest between two presented entities or not. The annotators were not shown the labels generated by the distant supervision process. The annotators were asked to only label sentences as positive if it contains a clear indication that the pharmacological substance either treats or prevents the disease. For example, the following sentence mentions that a study has been carried out to determine whether the drug *voriconazole* treats *paracoccidioidomycosis*:

> *"A pilot study was conducted to investigate the efficacy, safety, and tolerability of **voriconazole** for the long-term treatment of acute or chronic **paracoccidioidomycosis**, with itraconazole as the control treatment." (PMID=17990229)*

However, the sentence does not contain any indication that the drug successfully treats the disease and should therefore be annotated as a negative example of the relation.

The annotators were asked to label all 400 sentences and then re-examine any for which there was disagreement. Inter-annotator agreement (Cohen, 1960) after this stage was of $\kappa = 0.91$ for *may-treat* and $\kappa = 0.94$ for *may-prevent*. Remaining disagreements were resolved by one of the authors based on comments provided by both annotators and the annotation guidelines. The manually annotated version of the data set is referred to as **ML** (Manually Labelled).[4]

[1] http://mbr.nlm.nih.gov

[2] MetaMap annotations use UMLS release 2011AB, http://mbr.nlm.nih.gov/Download/MetaMapped_Medline/

[3] The UMLS's `MRREL` table contains information about related Concept Unique Identifiers (CUIs).

[4] The annotated corpus and further details about the annotation process are available here: https://sites.google.com/site/umlscorpus/home.

| | *may-prevent* | | | | | | *may-treat* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | evaluation on DL | | | evaluation on ML | | | evaluation on DL | | | evaluation on ML | | |
| # | prec | rec | f1 | prec | rec | f1 | prec | rec | f1 | prec | rec | f1 |
| 2000 | 33.33 | **21.95** | 26.47 | **44.44** | 20.34 | **27.91** | 44.97 | **54.03** | 49.08 | **48.32** | 51.43 | **49.83** |
| 4000 | 27.27 | 14.63 | 19.05 | **40.91** | **15.25** | **22.22** | 46.32 | **50.81** | **48.46** | 46.32 | 45.00 | 45.65 |
| 6000 | 38.89 | **17.07** | **23.73** | 38.89 | 11.86 | 18.18 | **54.05** | **64.52** | **58.82** | 51.35 | 54.29 | 52.78 |
| 8000 | 47.62 | **24.39** | **32.26** | **57.14** | 20.34 | 30.00 | **57.03** | **58.87** | **57.94** | 53.91 | 49.29 | 51.49 |
| 10000 | 44.44 | **39.02** | 41.56 | **58.33** | 35.59 | **44.21** | **61.40** | **56.45** | **58.82** | 53.51 | 43.57 | 48.03 |
| 12000 | 58.33 | **34.15** | **43.08** | 58.33 | 23.73 | 33.73 | **65.05** | **54.03** | **59.03** | 53.40 | 39.29 | 45.27 |
| 14000 | **52.38** | **53.66** | **53.01** | 50.00 | 35.59 | 41.58 | **68.89** | **50.00** | **57.94** | 57.78 | 37.14 | 45.22 |
| 16000 | **70.83** | **41.46** | **52.31** | 58.33 | 23.73 | 33.73 | **66.02** | **54.84** | **59.91** | 55.34 | 40.71 | 46.91 |

Table 2: Results for relation extraction system evaluated against DL and ML data sets

## 3 Label Comparison

Table 1 shows differences in the annotations for the two techniques for labelling that data. The ML data set for *may-treat* contains 173 positive and 227 negative examples, whereas the ML data set for *may-prevent* contains 139 positives and 261 negatives examples. A comparison of the DL and ML data sets shows that 40.25% of the labels changed for *may-treat* and 39.75% for *may-prevent*. The distant supervision process generated more false positives than false negatives for both relations.

If we assume that we have a classifier that is able to identify the *may-treat* and *may-prevent* relations with perfect accuracy then performance on the ML data sets would be precision=1.0, recall=1.0 and f-score=1.0. However, the false labels on the DL data sets would lead to performance of the same classifiers being estimated as precision=0.61, recall=0.53 and f-score=0.57 for *may-treat* and precision=0.61, recall=0.43 and f-score=0.50 for *may-prevent*. Hence, the two data sets may provide quite different estimates of system performance and we explore this in more detail in the next section.

## 4 Relation Extraction

A distantly supervised relation classifier was evaluated using manually and distantly labelled versions of the test data. Classifiers were trained for both relations and evaluated using both data sets (DL and ML). The evaluation was carried out using entity level evaluation, i.e. precision and recall are computed based on the proportion of correctly identified entity pairs which occur in sentences labeled as positive examples (according to the annotations contained within DL or ML). Entity level evaluation is commonly used to evaluate distantly supervised relation extraction systems. Similar results have been observed using the alternative approach of sentence level evaluation in which precision and recall are computed by examining the prediction for each sentence.

We use MultiR (Hoffmann et al., 2010), a multi-instance learning system that has been shown to provide state of the art results for distantly supervised relation extraction. The features used are those described by Surdeanu et al. (2011). The system is trained using distantly labelled examples (Section 2.1) of the *may-treat* and *may-prevent* relations containing equal numbers of positive and negative instances. The number of training examples is varied from 2,000 to 16,000 in increments of 2,000.

Results are shown in Table 2. Highlighted figures indicate the data set (DL or ML) against which the highest score was obtained for each metric (prec., rec. and f1) and configuration (relation and number of training examples). In general increasing the amount of training data leads to improved results on the DL data. In particular an increase in precision is observed when there is more training data. However, a different pattern is observed for the ML data and increasing the amount of training data does not always lead to an improvement in the f1-score. Results also show that the performance estimates obtained using the DL and ML data sets are only loosely associated. The results are similar for smaller training data sets but diverge as the amount of training data increases.

The table also shows that for both relations the performance estimates using the DL data are in general higher than those obtained using ML. This

trend becomes more pronounced as the amount of training data used increases. The most likely reason for this difference is that the classifiers are trained using distantly supervised data and therefore model the labels in the DL data set more closely than those in found in ML.

These results demonstrate that evaluation using distantly labelled gold standard data tends to overestimate performance. In some cases the discrepancy is large (up to 18.58 for *may-prevent* and 13.76 for *may-treat*). However, it does not seem to be consistent or particularly predictable. Consequently, improving the performance of a relation extraction system relative to distantly labelled evaluation data does not necessarily imply an increase in performance when measured against a manually annotated gold-standard.

## 5 Conclusion

This paper explored the effect of evaluating biomedical relation extraction systems using held-out test data annotated using distant supervision. Test data for two biomedical relations was annotated using distant supervision and also manually annotated. The manual and automatic labels differed for a large portion of the sentences. A distantly supervised relation extraction system was also evaluated using both data sets. We found that evaluation using held-out distantly supervised data tended to overestimate performance and that the connection between improved performance against distantly and manually labelled data was unclear. The use of held-out distantly labelled data is a cheap and efficient way to evaluate relation extraction systems, however this analysis demonstrates that the results obtained should be treated with some caution and, ideally, systems should also be evaluated against manually labelled data.

The results presented here were obtained for two biomedical relations. In future we plan to extend our analysis to a wider set of relations.

## Acknowledgments

The authors are grateful to the Engineering and Physical Sciences Research Council (EP/J008427/1) for funding the research described in this paper.

## References

A. Aronson and F. Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Association*, 17(3):229–236.

Isabelle Augenstein, Diana Maynard, and Fabio Ciravegna. 2014. Relation extraction from the web using distant supervision. In *Proceedings of the 19th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2014)*, Linköping, Sweden, November.

Kedar Bellare and Andrew Mccallum. 2007. Learning Extractors from Unlabeled Text using Relevant Databases. In *Sixth International Workshop on Information Integration on the Web (IIWeb)*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 77–86. AAAI Press.

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*, ACL '10, pages 286–295, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ander Intxaurrondo, Mihai Surdeanu, Oier Lopez de Lacalle, and Eneko Agirre. 2013. Removing noisy mentions for distant supervision. *Procesamiento del Lenguaje Natural*, 51:41–48.

Sebastian Krause, Hong Li, Hans Uszkoreit, and Feiyu Xu. 2012. Large-scale learning of relation-extraction rules with distant supervision from the web. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I*, ISWC'12, pages 263–278, Berlin, Heidelberg. Springer-Verlag.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782, Atlanta, Georgia, June. Association for Computational Linguistics.

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*, ACL '09, pages 1003–1011,

Stroudsburg, PA, USA. Association for Computational Linguistics.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7, Sofia, Bulgaria, August. Association for Computational Linguistics.

Truc-Vien T. Nguyen and Alessandro Moschitti. 2011. Joint distant and direct supervision for relation extraction. In *Proceedings of The 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, pages 732–740. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD '10)*.

Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. 2013. Modeling missing data in distant supervision for information extraction. In *Association for Computational Linguistics Vol. 1 (TACL)*.

Roland Roller and Mark Stevenson. 2014. Applying umls for distantly supervised relation detection. In *Proceedings of the Fifth International Workshop on Health Text Mining and Information Analysis (Louhi 2014)*, Gothenburg, Sweden.

Roland Roller, Eneko Agirre, Aitor Soroa, and Mark Stevenson. 2015. Improving distant supervision using inference learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 273–278, Beijing, China, July. Association for Computational Linguistics.

Benjamin Roth and Dietrich Klakow. 2013. Combining generative and discriminative model scores for distant supervision. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 24–29, Seattle, Washington, USA, October. Association for Computational Linguistics.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero Zazo. 2013. Semeval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Mihai Surdeanu, David McClosky, Mason Smith, Andrey Gusev, and Christopher Manning. 2011. Customizing an information extraction system to a new domain. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, pages 2–10, Portland, Oregon, USA, June. Association for Computational Linguistics.

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*, EMNLP-CoNLL '12, pages 455–465, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 721–729, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philippe Thomas, Illés Solt, Roman Klinger, and Ulf Leser. 2011. Learning protein protein interaction extraction using distant supervision. In *Proceedings of Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing*, pages 34–41.

Andreas Vlachos and Stephen Clark. 2014. Application-driven relation extraction with limited distant supervision. In *Proceedings of the First AHA!-Workshop on Information Discovery in Text*, pages 1–6, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

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

Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. 2013. Filling knowledge base gaps for distant supervision of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 665–670, Sofia, Bulgaria, August. Association for Computational Linguistics.

Xingxing Zhang, Jianwen Zhang, Junyu Zeng, Jun Yan, Zheng Chen, and Zhifang Sui. 2013. Towards accurate distant supervision for relational facts extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 810–815, Sofia, Bulgaria, August. Association for Computational Linguistics.