Making the most of limited training data using distant supervision

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

Automatic recognition of relationships between key entities in text is an important problem which has many applications. Supervised machine learning techniques have proved to be the most effective approach to this problem. However, they require labelled training data which may not be available in sufficient quantity (or at all) and is expensive to produce. This paper proposes a technique that can be applied when only limited training data is available. The approach uses a form of distant supervision but does not require an external knowledge base. Instead, it uses information from the training set to acquire new labelled data and combines it with manually labelled data. The approach was tested on an adverse drug data set using a limited amount of manually labelled training data and shown to outperform a supervised approach.

1 Introduction

Relation extraction is a widely explored problem that has been applied to a range of domains (Craven and Kumlien, 1999; Agichtein and Gravano, 2000; Xu et al., 2007) using a variety of techniques (Yangarber, 2003; Bunescu and Mooney, 2006; Neumann and Schmeier, 2012). In the biomedical domain relation extraction has been used to identify a wide range of types of relation, including adverse drug effects (ADE), gene regulations and drug-drug interactions. Community evaluation exercises, such as the BioNLP Shared Task (Kim et al., 2011; Nédellec et al., 2013) or the Drug-Drug Interaction (DDI) challenge (Segura-Bedmar et al., 2013), have shown that supervised learning techniques normally produce better results than other approaches.

Supervised learning techniques rely on labeled training data but these are not available for all relations of interest and are also difficult and time-consuming to create. Other approaches may be more appropriate in situations where training data is limited or unavailable. Minimally supervised approaches, such as seed and bootstrapping techniques (Brin, 1999; Riloff and Jones, 1999; Agichtein and Gravano, 2000), are provided with a small set of seed instances (examples of related information) or patterns and acquire further examples from a large corpus by applying an iterative process. While these approaches do not require labelled training data they often suffer from low precision or semantic drift (Mintz et al., 2009). Distant supervision combines the advantages of minimally supervised and supervised approaches to relation extraction.

Distant supervision makes use of an external knowledge source that provides information about pairs of entities which are related. Sentences containing both entities in a pair are identified from a corpus and used in place of labeled training examples. For example, knowledge that *hair loss* is a drug-related adverse effect of *paroxetine* would allow further positive examples to be identified by searching for other sentences containing the same drug and side-effect. Many knowledge sources only contain positive entity pairs. Therefore negative examples are often generated using a closed-world assumption. Given the known positive entity pairs, negative entity pairs are generated by producing new combinations of entities. Negative example sentences are generated by selecting sentences containing these negative entity pairs.

The example in figure 1 shows the limitations of distant supervision since related entities might express a different relation. This can lead to examples being falsely labelled as positive examples of a relation. Classifiers trained using data generated using distant supervision do not generally

Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 12–20,
Beijing, China, July 30, 2015. ©2015 Association for Computational Linguistics

perform as well as those trained using manually labelled data. However, distant supervision allows large data sets to be generated at low cost.

> There are a few case reports on **[CONDITION:***hair loss***]** associated with tricyclic antidepressants and serotonin selective reuptake inhibitors (SSRIs), but none deal specifically with **[DRUG:***paroxetine***]**.

Figure 1: Generation of false positives by using automatically labelled data, PMID=10442258

The majority of distant supervision approaches use structured knowledge sources such as Wikipedia (Hoffmann et al., 2010) or Freebase (Mintz et al., 2009; Riedel et al., 2010; Ritter et al., 2013; Augenstein et al., 2014). However there may not be a suitable knowledge base available for a particular relation of interest. This paper addresses the problem of developing relation extraction systems in situations where only a small amount of training data is available.

We introduce a method for relation extraction that can be used when only limited amounts of training data are available. The approach is based on *distant supervision* but, rather than relying on a knowledge base, seed pairs are extracted from Medline articles. Sentences from the Medline Baseline Repository containing these seed pairs are extracted to generate a large distantly labelled training data set. Using this data manually labelled data can be extended and combined to a hybrid *mixture model* which outperforms both the supervised and the distantly supervised models.

This paper makes the following contributions: 1) introduces a method which can be used to train a relational classifier when only a small set of labelled training data is available, 2) provides a method for combining distant supervision with supervised learning methods and 3) presents distant supervision without the need of a knowledge base.

The remainder of the paper is structured as follows. The next section presents the background on relation extraction from biomedical documents. Section 3 introduces the data set which is used for the experiments. The techniques for generating the distantly supervised training data and relational classifier are described in sections 4 and 5. Section 6 describes the experiment and the results. Conclusions are presented in section 7.

2 Related Work

Supervised learning techniques are popular and efficient approaches to detecting relations between entities in natural language. Results using supervised learning methods tend to improve as more training data is available. However the generation of labelled data is cumbersome, expensive and time-consuming. It often requires expert knowledge in restricted domains, such as biomedicine. A new labelled data set is required for each target relation.

In recent years, distant supervision has become very popular. Rather than using manually annotated data, distant supervision uses knowledge about which entity pairs are instances of the target relation to generate automatically labelled data which is used to train a relational classifier. Craven and Kumlien (1999) introduced distant supervision for relation extraction. The authors used the Yeast Protein Database (YPD) as source of knowledge and mapped this information to PubMed articles to generate training examples. The technique has been widely applied particularly outside the medical domain. Many approaches such as (Mintz et al., 2009; Sun et al., 2011; Hoffmann et al., 2011; Krause et al., 2012; Xu et al., 2013) focus on approaches using Freebase as knowledge source to generate automatically labelled data. In recent years distant supervision has also become more popular in the biomedical domain beeing used to detect protein-protein interactions using IntAct (Thomas et al., 2011), protein-residue associations with PDB (Ravikumar et al., 2012) or relationships of the National Drug File-Reference Terminology (NDF-RT) using the UMLS Metathesaurus (Roller and Stevenson, 2014). Liu et al. (2014) focus on the detection of genes in brain regions from literature using the UMLS Semantic Network and Ellendorff et al. (2014) uses the Comparative Toxicogenomics Database (CTD) to detect interactions between genes and chemicals.

The distantly supervised methods of Nguyen and Moschitti (2011) and Pershina et al. (2014) differ slightly from many other approaches. Both combine supervised and distantly supervised models. Nguyen and Moschitti (2011) use a support vector machine and combine the supervised and the distantly supervised classifier with a linear combination. Pershina et al. (2014) instead integrate the manually labelled data directly within their distantly supervised multi-learning approach.

Both approaches show that a combination of a large set of distantly supervised (noisy) data with manually labelled examples can improve the classification results. The combination of noisy data and hand-selected training examples is also used in this paper.

3 Data

The experiments in this work uses the ADE data set (Gurulingappa et al., 2012b) which contains examples of adverse drug effects (ADE). An ADE is a response of a drug which is noxious and unintended, and which occurs at doses normally used in humans for the prophylaxis, diagnosis, therapy of disease, or for the modification of physiological function[1] (Gurulingappa et al., 2012b). ADEs contribute to one of the most common causes of death in industrialised nations and are the fourth leading cause of death in the U.S. (Giacomini et al., 2007). To reduce this risk the side-effects of drugs need to be detected and made publicly available as quickly as possible.

The ADE data set consists of Medline case reports examined by three human annotators. Sentences in these case reports containing adverse effects between *drugs* and *conditions* were extracted and entities annotated to generate the data set. An example relation between a drug and a condition from this data set is shown in figure 2. According to the given sentence the condition *pseudoporphyria* is caused by the two drugs *naproxen* and *oxaprozin*.

METHODS: We report two cases of **[CONDITION:***pseudoporphyria***]** caused by **[DRUG:***naproxen***]** and **[DRUG:***oxaprozin***]**.

Figure 2: Example of a drug-related adverse effect taken from PMID=10082597

The ADE corpus only contains examples of positive relations. Negative examples are also required to set-up a meaningful ADE prediction task and to train a supervised ADE classifier. A set of negative examples were generated using the following process.

Named entity recognition is applied to detect drugs and conditions. MetaMap[2] (Aronson and Lang, 2010) was run on the unannotated sentences in the ADE corpus to detect biomedical concepts from the UMLS. MetaMap provides different possible UMLS concept mappings and we select the best (highest ranked) mapping. Each biomedical concept detected by MetaMap now refers to a unique UMLS CUI thereby allowing identical concepts to be merged and assigned semantic types. Using the same approach as Kang et al. (2014), sentences containing concepts with semantic types which belong to the two groups “Chemicals & Drugs” and “Disorders” are extracted and considered as negative examples. Nested relations are not included in our data set.

Training and evaluation sets were then generated. The set of utilised ADE abstracts consists of 1644 publications. 200 abstracts were removed to be used to create training data and the remainder used to form the evaluation set. The training data is created by extracting all positive and negative labelled sentences from the 200 abstracts. In order to provide reliable results we run the same experiment 5 times. Each time we randomly choose a different selection of 200 training and 1444 test abstracts.

4 Automatic Generation of Annotated Training Data

Many of the previous approaches to distant supervision use information about related instances (e.g. drugs and known adverse effects) to automatically generate training data. In the majority of cases this information is obtained from a knowledge base. We employ an alternative approach and make use of information from a small set of abstracts. For example, the sentence shown in figure 2 suggests that there are cases when the drugs *oxaprozin* and *naproxen* cause *pseudoporphyria*. Consequently sentences containing these two *drug-condition* entity pairs (i.e. *oxaprozin-pseudoporphyria* and *naproxen-pseudoporphyria*) are extracted and treated as positive examples.

The data is generated by applying a three stage process (see Figure 3).

1) *Map CUIs to the related entities in the training data set.* We begin by normalising medical concepts. Medical terms can occur in literature with different names, using a different spelling or abbreviations. For instance *Naproxen* can be also described as *Methoxypropiocin*, *MNPA* or *6-Methoxy-alpha-methyl-2-naphthaleneacetic Acid.* UMLS maps these different names to the same

[1] World Health Organization (WHO) glossary of terms used in Pharmacovigilance.

[2] http://metamap.nlm.nih.gov/, MetaMap version 13 with UMLS 2013AA

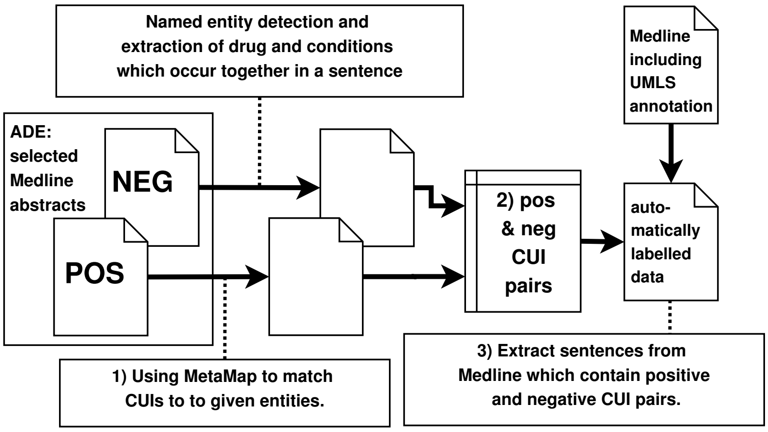
Named entity detection and extraction of drug and conditions which occur together in a sentence
Medline including UMLS annotation
ADE: selected Medline abstracts
NEG
POS
2) pos & neg CUI pairs
auto-matically labelled data
1) Using MetaMap to match CUIs to to given entities.
3) Extract sentences from Medline which contain positive and negative CUI pairs.

Figure 3: Automatic generation of training data for ADE relations

CUI, *C0027396*. We run MetaMap with the same configuration on the sentences containing positive examples. In many cases it is possible to assign a MetaMap annotation to the existing related entities.

We only assign a CUI to an entity if MetaMap identifies a CUI that can be mapped to the entity in its full length (not only a substring). Negative training examples already include CUI information for each entity (see section 3).

2) *Extract a set of positive and negative seed instance pairs.* In the next step, we extract all CUI pairs from the positive ADE examples and add them to a set of positive instance pairs P. We also extract CUI pairs of negative ADE examples and add them to a negative instance pair set N. Each CUI pair which occurs in both sets (P and N), is removed from N. Considering the 200 training abstracts of the first setup (of five) it is possible to extract 310 different positive CUIs pairs and 869 negative CUI pairs. 12 CUI pairs occur in both sets. Therefore the number of different CUI pairs in N is reduced to 857.

3) *Extract sentences containing positive and negative seed instances from abstracts.* The distantly labelled training data is generated using the Medline Baseline Repository (MBR)[3], a large collection of biomedical abstracts annotated using MetaMap[4]. We use 3,000,000 abstracts published between 1997-2003. Then sentences from this subset containing positive and negative CUI pairs are extracted and labelled as positive and negative examples.

Regarding the the 200 training abstracts of the first setup a total of 7868 sentence containing positive instance pairs and 14,4315 sentence containing negative instance pairs were identified and extracted. Although 310 different positive and 857 different negative CUI pairs were extracted from the 200 abstracts (see above), only 290 different positive and 441 different negative CUI pair combinations were detected within the portion of MBR used for this experiment. It is also interesting to note that only 13 positive CUI pairs occur more than 100 times within the 7868 positive examples. The most frequent positive CUI pairs are listed in table 1. 213 of the positive CUI pairs occur fewer than 10 times.

The automatically generated data has a strong bias. To generate an automatically labelled training data with a similar bias as the test set we reduce the amount of negative examples to the same ratio as the manually labelled examples.

5 Relation Extraction

We use the Java Simple Relation Extraction[5] (jSRE) (Giuliano et al., 2006) which is based on LibSVM (Chang and Lin, 2011). jSRE includes an implementation of the shallow linguistic kernel which provides reliable classification results and has been used also for other experiments on the ADE data set (Gurulingappa et al., 2012a; Kang et al., 2014).

The shallow linguistic kernel is a combination of the *global context kernel* and the *local context kernel*. The global context kernel considers n-grams of the words (and other information such as stemmed words and part of speech tags) between

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

[4] MetaMap annotations use UMLS release 2011AB, http://mbr.nlm.nih.gov/Download/MetaMapped_Medline/2012/

[5] https://hlt.fbk.eu/technologies/jsre

frequency	drug	condition
#1352	C0019134='Heparin'	C0272285='Heparin-induced thrombocytopenia'
#1199	C0026549='Morphine'	C0030193='Pain'
#980	C0023175='Lead'	C0020538='Hypertensive disease'
#396	C0031507='Phenytoin'	C0036572='Seizure'

Table 1: Most frequent positive CUI pairs found in the automatically labelled data set

the two entities. The local context kernel considers only a limited amount of information around each entity.

Sentences from the training and test data are parsed using the Charniak-Johnson Parser (Charniak and Johnson, 2005) to generate part of speech tags. Next, words are reduced to their stem using the Porter Stemmer (Porter, 1997).

We use three different methods within the experiments: supervised relation extraction, distantly supervised relation extraction and a relation extraction using a mixture-model. The supervised model uses a set of abstracts (1-200) from the training data as input. The distantly supervised model takes the automatically generated data based on the MetaMap annotated Medline Baseline Repository as input. The mixture-model merges the automatically generated and manually labelled training data to form a combined training set.

6 Experiment

In this experiment we examine different sizes of manually labelled training data. Starting with a single abstract for training we slowly increase the number of seed abstracts to 200. In parallel we generate for each training set a different distantly labelled data set using the given ADE seed facts of the training data. The more information the manually labelled data contains, the more different seeds can be extracted which increases the size of the distantly labelled data. Thereafter we combine in each step both data sets to a mixture-model.

In order to provide reliable results we repeat this experiment five times (five evaluation rounds) with a different selection of abstracts for training and test. In each evaluation round the abstracts utilised for training are chosen randomly. The remaining abstracts are used for evaluation. During a specific evaluation round (increasing training data) the test set remains unchanged. The results of the experiments are presented in table 2 and figure 4. The results represent the mean of all five different evaluation rounds.

The results show that the performance for all models improves as the amount of data increases. Performance of the supervised classifier increases sharply as the number of abstracts is increased from 1 to 10 abstracts. Increasing the size of the training data to 50 abstracts produces a further improvement of approximately 30%. These results demonstrate that even small amounts of training data are sufficient to provide reasonable results on the ADE data set.

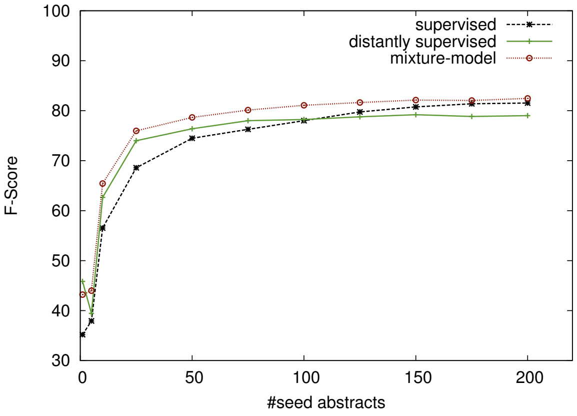

Figure 4: Effect of varying number of seed abstracts

Performance of the distantly supervised classifier shows a similar pattern. Increasing the number of seed abstracts results in a larger distantly labelled training data set which improves classification results. The distantly supervised classifier outperforms the supervised one when there are fewer than 100 seed abstracts. The reason for this is the supervised classifier does not have access to a sufficient volume of training data while the distant supervision is able to generate more. As the number of seed abstracts increases the situation is reversed with the supervised classifier outperforming the distantly supervised one. When more than 100 abstracts are available the supervised classifier has the advantage of having access to enough accurately labelled examples to train a relation ex-

	supervised model			distant supervision			mixture model		
#SA	**prec.**	**rec.**	**F1**	**prec.**	**rec.**	**F1**	**prec.**	**rec.**	**F1**
1	52.75	42.33	35.18	42.20	53.95	**45.84**	43.54	50.40	43.20
5	68.48	32.53	37.92	76.78	37.47	39.45	78.40	38.76	**43.98**
10	66.85	51.17	56.53	71.90	61.33	62.66	73.90	61.97	**65.43**
25	68.01	69.88	68.57	69.01	81.48	73.99	71.88	81.58	**75.96**
50	72.29	76.88	74.48	69.27	86.68	76.39	72.62	86.46	**78.66**
75	73.77	79.18	76.27	68.35	91.10	78.01	73.43	88.30	**80.13**
100	75.41	80.85	78.00	67.79	92.56	78.24	73.80	89.99	**81.09**
125	75.79	84.16	79.75	69.11	91.65	78.77	74.91	89.77	**81.64**
150	76.89	85.06	80.77	70.15	90.99	79.19	75.81	89.65	**82.13**
175	77.14	86.15	81.39	68.50	93.03	78.84	74.45	91.40	**82.04**
200	77.32	86.28	81.54	68.77	92.98	79.02	75.01	91.63	**82.47**

Table 2: Effect of varying size of training data set

traction system. The distantly supervised classifier still has access to more data but it is not as accurate.

	manual lab.		distantly lab.		seeds	
#SA	**pos**	**neg**	**pos**	**neg**	**pos**	**neg**
10	67	121	510	891	15	54
25	180	232	1048	1404	38	103
50	388	485	2026	2580	81	213
75	590	756	2643	3398	123	330
100	804	1024	3818	4851	172	448
150	1200	1447	5663	6863	248	636
200	1632	1900	8289	9607	336	834

Table 3: ADE training data size (mean across five runs)

The mixture model produces the best results of all approaches when 5 or more abstracts are used. This result is interesting since the manually labelled data is simply extended using a simple form of distant supervision that is straightforward to apply. The mixture model tends to achieve higher precision but lower recall than the distantly supervised approach, possibly because the training data used by the mixture model is more accurate and contains fewer ”false positive” examples. On the other hand the precision and recall of the mixture model are often higher than the supervised model. The increase in recall is presumably caused by having access to additional training data and the precision scores suggest that the classifier is not harmed by some of these containing noisy labels.

The difference in performance between the supervised and the mixture-models gets smaller as the number of seed abstracts increases.

Table 3 shows the mean size of the different sets of training data. The amount of distantly labelled data is much larger than the manually labelled data at each classification step. Larger amounts of manually labelled data increase the number of ADE seed instances that can be extracted which leads to more distantly supervised examples.

7 Discussion and Conclusion

This paper introduced a new distantly supervised method for relation extraction that was applied to the identification of ADE relations from biomedical documents. The approach is able to use information from an existing training data set to automatically acquire new training data. Using this data, a relational classifier can be trained to detect and extract similar information in natural language. The classifier is able to provide comparable results to a supervised classifier using a small gold standard as input. Furthermore we presented a mixture model using manually labelled and distantly labelled data which is able to outperform a classifier using only (a small set of) gold standard data. This result is notable since distantly supervised data tends to be much noisier than manually labelled data and therefore produce less accurate classifiers.

Distant supervision is a well explored technique for relation extraction that has proven to be effective. Our proposed methods differs slightly in the way seed instances are generated. Rather than using a knowledge base we directly extract positive and negative seed pairs from an existing data set and use them for distant supervision.

We plan to extend the work described in this paper in various ways. Firstly we would like to experiment with alternative classifiers such as applying dependency features and stacking or merging to combine different kernel models. We would also like to explore different techniques for combining the supervised and the distantly supervised model.

Acknowledgements

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

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, DL '00, pages 85–94, New York, NY, USA. ACM.

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.

Sergey Brin. 1999. Extracting patterns and relations from the world wide web. In *Selected papers from the International Workshop on The World Wide Web and Databases*, WebDB '98, pages 172–183, London, UK, UK. Springer-Verlag.

Razvan Bunescu and Raymond J. Mooney. 2006. Subsequence kernels for relation extraction. In *Submitted to the Ninth Conference on Natural Language Learning (CoNLL-2005)*, Ann Arbor, MI, July.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Tilia Ellendorff, Fabio Rinaldi, and Simon Clematide. 2014. Using large biomedical databases as gold annotations for automatic relation extraction. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Kathleen M. Giacomini, Ronald M. Krauss, Dan M. Roden, Michel Eichelbaum, Michael R. Hayden, and Yusuke Nakamura. 2007. When good drugs go bad. In *Nature, 466(7139)*, pages 975–977.

Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy.

Harsha Gurulingappa, Abdul MateenRajput, and Luca Toldo. 2012a. Extraction of potential adverse drug events from medical case reports. *Journal of Biomedical Semantics*, 3(1):15.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012b. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892. Text Mining and Natural Language Processing in Pharmacogenomics.

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.

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*, ACL '11, pages 541–550.

Ning Kang, Bharat Singh, Chinh Bui, Zubair Afzal, Erik van Mulligen, and Jan Kors. 2014. Knowledge-based extraction of adverse drug events from biomedical text. *BMC Bioinformatics*, 15(1):64.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of bionlp shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1–6, Portland, Oregon, USA, June. Association for Computational Linguistics.

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.

Mengwen Liu, Yuan Ling, Yuan An, Xiaohua Hu, Alan Yagoda, and Rick Misra. 2014. Relation extraction from biomedical literature with minimal supervision and grouping strategy. In *Proceedings of IEEE Conference on Bioinformatics and Biomedicine (BIBM14)*, pages 444–449.

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.

Günter Neumann and Sven Schmeier. 2012. Exploratory search on the mobile web. In *In 4th International Conference on Agents and Artificial Intelligence (ICAART 2012)*, pages 110–119, Vilamoura, Algarve, Portugal.

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.

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)*, pages 732–738, Baltimore, Maryland, June. Association for Computational Linguistics.

M. F. Porter. 1997. Readings in information retrieval. chapter An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

KE Ravikumar, Haibin Liu, Judith Cohn, Michael Wall, and Karin Verspoor. 2012. Literature mining of protein-residue associations with graph rules learned through distant supervision. *Journal of Biomedical Semantics*, 3(Suppl 3):S2.

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)*.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, AAAI '99/IAAI '99, pages 474–479, Menlo Park, CA, USA. American Association for Artificial Intelligence.

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.

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.

Ang Sun, Ralph Grishman, and Satoshi Sekine. 2011. Semi-supervised relation extraction with large-scale word clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 521–529, Portland, Oregon, USA, June. 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.

Feiyu Xu, Hans Uszkoreit, and Hong Li. 2007. A seed-driven bottom-up machine learning framework for extracting relations of various complexity. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 584–591, Prague, Czech Republic, June. Association for Computational Linguistics.

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.

Roman Yangarber. 2003. Counter-training in discovery of semantic patterns. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 343–350, Stroudsburg, PA, USA. Association for Computational Linguistics.