

# Seed Selection for Distantly Supervised Web-Based Relation Extraction

Isabelle Augenstein

Department of Computer Science  
The University of Sheffield  
United Kingdom  
i.augenstein@dcs.shef.ac.uk

## Abstract

In this paper we consider the problem of distant supervision to extract relations (e.g. origin(musical artist, location)) for entities (e.g. ‘The Beatles’) of certain classes (e.g. musical artist) from Web pages by using background information from the Linking Open Data cloud to automatically label Web documents which are then used as training data for relation classifiers. Distant supervision approaches typically suffer from the problem of ambiguity when automatically labelling text, as well as the problem of incompleteness of background data to judge whether a mention is a true relation mention. This paper explores the hypothesis that simple statistical methods based on background data can help to filter unreliable training data and thus improve the precision of relation extractors. Experiments on a Web corpus show that an error reduction of 35% can be achieved by strategically selecting seed data.

## 1 Introduction

One important aspect to every relation extraction approach is how to annotate training and test data for learning classifiers. In the past, four different types of approaches for this have been proposed.

For *supervised* approaches, training and test data is annotated manually by one or several annotators. While this approach results in a high-quality corpus, it is very expensive and time-consuming. As a consequence, the corpora used tend to be small and biased towards a certain domain or type of text.

*Unsupervised* approaches do not need annotated data for training; they instead cluster similar word sequences and generalise them to relations. Although unsupervised approaches can process very large amounts of data, resulting relations are hard to map to particular schemas. In addition, Fader et al. (2011) observe that these approaches often produce uninformative or incoherent extractions.

*Semi-supervised* methods are methods that only require a small number of seed instances. Hand-crafted seeds are used to extract patterns from a corpus, which are then used to extract more instances and those again to extract new patterns in an iterative way. However, since many iterations are needed, these methods are prone to semantic drift, i.e. an unwanted shift of meaning. As a consequence these methods require a certain amount of human effort - to create seeds initially and also to help keep systems ‘on track’.

A fourth group of approaches, *distant supervision* or *self-supervised* approaches, exploit big knowledge bases such as Freebase (2008) to automatically label entities in text and use the annotated text to extract features and train a classifier (Wu and Weld, 2007; Mintz et al., 2009). Unlike supervised systems, they do not require manual effort to label data and can be applied to large corpora. Since they extract relations which are defined by schemas, these approaches also do not produce uninformative or incoherent relations. Distant supervision approaches are based on the following assumption (Mintz et al., 2009):

“If two entities participate in a relation, any sentence that contains those two entities might express that relation.” In practice, if the information that two entities participate in a relation is contained in the knowledge base, whenever they appear in the same sentence, that sentence is used as positive training data for that relation. This heuristic causes problems if different entities have the same surface form. Consider

the following example:

“*Let It Be* is the twelfth album by *The Beatles* which contains their hit single ‘*Let It Be*’.”

In that sentence, the first mention of *Let It Be* is an example of the *album* relation, whereas the second mention is an example of the *track* relation. If both mentions are used as positive training examples for both relations, this impairs the learning of weights of the relation classifiers. We therefore argue for the careful selection of training data for distant supervision by using measures to discard highly ambiguous training examples. One further aspect that can be problematic when automatically creating negative training data is incompleteness. What Riedel et al. (2010) point out, and our observations also confirm, is that about 20%, or even more, of all true relation mentions in a corpus are not contained in Freebase, although it is a very big knowledge base.

The main contributions of this paper are: to propose and evaluate several measures for detecting and discarding unreliable seeds; and to document a distant supervision system for fine-grained class-based relation extraction on noisy data from the Web.

## 2 Distantly Supervised Relation Extraction

Distant supervision is defined as the automatic labelling of a corpus with properties,  $P$  and entities,  $E$  from a knowledge base,  $KB$  to train a classifier to learn to predict relations. Following previous distant supervision approaches, we only consider binary relations of the form  $(s, p, o)$ , consisting of a subject, a predicate and an object (Mintz et al., 2009). We use the established Semantic Web formalisation, rather than unary and binary first order predicates, to reflect our special and consistent treatment of subjects versus objects. Each subject and object entity  $e \in E$  has a set of lexicalisations,  $L_e \subset L$ . Furthermore, we consider only those subjects which have a particular Freebase class  $C$ .

## 3 Seed Selection

Before using the automatically labelled corpus to train a classifier, we include a seed selection step, which consist of several measures to discard unreliable seeds.

### Ambiguity Within An Entity

**Unam:** Our first approach is to discard lexicalisations of objects if they are ambiguous for the subject entity, i.e. if a subject is related to two different objects which have the same lexicalisation, and express two different relations. To illustrate this, let us consider the problem outlined in the introduction again: *Let It Be* can be both an *album* and a *track* of the subject entity *The Beatles*, therefore we would like to discard *Let It Be* as a seed for the class *Musical Artist*. We measure the degree to which a lexicalisation  $l \in L_o$  of an object  $o$  is ambiguous by the number of senses the lexicalisation has. For a given subject  $s$ , if we discover a lexicalisation for a related entity, i.e.  $(s, p, o) \in KB$  and  $l \in L_o$ , then, since it may be the case that  $l \in L_r$  for some  $R \ni r \neq o$ , where also  $(s, q, r) \in KB$  for some  $q \in P$ , we say in this case that  $l$  has a “sense”  $o$  and  $r$ , giving rise to ambiguity. We then define  $A_l^s$ , the ambiguity of a lexicalisation with respect to the subject as follows:  $A_l^s = |\{e \mid l \in L_e \cap L_w \wedge (s, p, e) \in KB \wedge (s, v, w) \in KB \wedge w \neq e\}|$ .

### Ambiguity Across Classes

In addition to being ambiguous for a subject of a specific class, lexicalisations of objects can be ambiguous across classes. Our assumption is that the more senses an object lexicalisation has, the more likely it is that that object occurrence is confused with an object lexicalisation of a different property of any class. An example for this are common names of book authors or common genres as in the sentence “*Jack* mentioned that he read *On the Road*”, in which *Jack* is falsely recognised as the author Jack Kerouac.

**Stop:** One type of very ambiguous words with many senses are stop words. Since some objects of relations in our training set might have lexicalisations which are stop words, we discard those lexicalisations if they appear in a stop word list (we use the one described in Lewis et al. (2004)).

**Stat:** For other highly ambiguous lexicalisations of object entities our approach is to estimate cross-class ambiguity, i.e. to estimate how ambiguous a lexicalisation of an object is compared to other lexicalisations of objects of the same relation. If its ambiguity is comparatively low, we consider it a reliable seed,

otherwise we want to discard it. For the set of classes under consideration, we know the set of properties that apply,  $D \subset P$  and can retrieve the set  $\{o \mid (s, p, o) \in KB \wedge p \in D\}$ , and retrieve the set of lexicalisations for each member,  $L_o$ . We then compute  $A_o$ , the number of senses for every lexicalisation of an object  $L_o$ , where  $A_o = |\{o \mid \in L_o\}|$ .

We view the number of senses of each lexicalisation of an object per relation as a frequency distribution. We then compute min, max, median ( $Q2$ ), the lower ( $Q1$ ) and the upper quartile ( $Q3$ ) of those frequency distributions and compare it to the number of senses of each lexicalisation of an object. If  $A_l > Q$ , where  $Q$  is either  $Q1$ ,  $Q2$  or  $Q3$  depending on the model, we discard the lexicalisation of the object.

### Incompleteness

One further aspect of knowledge bases that can be problematic when automatically creating negative training data is incompleteness. Our method for creating negative training data is to assume that all entities which appear in a sentence with the subject  $s$ , but are not in a relation with it according to the knowledge base, can be used as negative training data. Other distant supervision approaches (Mintz et al., 2009) follow a similar approach, but only use a random sample of unrelated entities pairs.

**Incomp:** Our approach is to discard negative training examples which are likely to be true relation mentions, but missing from the knowledge base. If we find a lexicalisation  $l$  where  $\nexists o, p \cdot l \in L_o \wedge (s, p, o) \in KB$ , then before we consider this a negative example we check if  $\exists t \in C \cdot (t, q, r) \in KB$  and  $l \in L_r$ , i.e. if any of the properties of the class we examine has an object lexicalisation  $l$ .

## 4 System

### 4.1 Corpus

To create a corpus for Web relation extraction using Linked Data, three Freebase classes and their six to seven most prominent properties (see Table 1) are selected and their values retrieved using the Freebase API. To avoid noisy training data, entities which only have values for some of those relations were not used. This resulted in 1800 to 2200 entities per class which were split equally for training and test. For each entity, at most 10 Web pages were retrieved via the Google Search API using the search pattern ““*subject\_entity*” *class property\_name*”, e.g. ““The Beatles” Musical Artist Origin” resulting in a total of 450,000 pages<sup>1</sup>. By adding the class, we expect the retrieved Web pages to be more relevant to our extraction task. For entities, Freebase distinguishes between the most prominent lexicalisation (the entity name) and other lexicalisations (entity aliases). We use the entity name for all of the search patterns.

Class	Property	Class	Property	Class	Property
Book	author	Musical Artist	album	Politician	birthdate
	characters		active (start)		birthplace
	publication date		active (end)		educational institution
	genre		genre		nationality
	ISBN		record label		party
	original language		origin		religion
			track		spouses

Table 1: Freebase classes and properties we use for our evaluation

### 4.2 NLP Pipeline

Text content is extracted from HTML pages using the jsoup API<sup>2</sup>, which strips text from each element recursively. Each paragraph is then processed with Stanford CoreNLP<sup>3</sup> to split the text into sentences,

<sup>1</sup>URLs of those Web pages are available via <http://staffwww.dcs.shef.ac.uk/people/I.Augenstein/SWAIE2014/>

<sup>2</sup><http://jsoup.org/>

<sup>3</sup><http://nlp.stanford.edu/software/corenlp.shtml>

tokenise, POS tag it and normalise time expressions. Named entities are classified using the 7 class (time, location, organisation, person, money, percent, date) named entity model.

### 4.3 Relation candidate identification

Some of the objects of relations cannot be categorised according to the 7 named entity (NE) classes detected by the Stanford named entity classifier (NERC) and are therefore not recognised, for example *MusicalArtist:album* or *Book:genre*. Therefore, in addition to recognising entities with Stanford NERC, we also implement our own named entity recogniser (NER), which only recognises entity boundaries, but does not classify them. To detect entity boundaries, we recognise sequences of nouns and sequences of capitalised words and apply both greedy and non-greedy matching. For greedy matching, we consider whole noun phrases and for non-greedy matching all subsequences starting with the first word of the those phrases, i.e. for ‘science fiction book’, we would consider ‘science fiction book’, ‘science fiction’ and ‘book’ as candidates. The reason to do greedy as well as non-greedy matching is because the lexicalisation of an object does not always span a whole noun phrase, e.g. while ‘science fiction’ is a lexicalisation of an object of *Book:genre*, ‘science fiction book’ is not. However, for *MusicalArtist:genre*, ‘pop music’ would be a valid lexicalisation of an object. We also recognise short sequences of words in quotes. This is because lexicalisation of objects of *MusicalArtist:track* and *MusicalArtist:album* often appear in quotes, but are not necessarily noun phrases.

### 4.4 Identifying Relation Candidates and Selecting Seeds

The next step is to identify which sentences potentially express relations. We only use sentences from Web pages which were retrieved using a query which contains the subject of the relation. We then select, or rather discard seeds for training according to the different methods outlined in Section 3. Our *baseline* model does not discard any training seeds.

### 4.5 Features

Our system uses some of the features described in Mintz et al. (2009), and other standard lexical features and named entity features:

- The object occurrence
- The bag of words of the occurrence
- The number of words of the occurrence
- The named entity class of the occurrence assigned by the 7-class Stanford NERC
- A flag indicating if the object or the subject entity came first in the sentence
- The sequence of part of speech (POS) tags of the words between the subject and the occurrence
- The bag of words between the subject and the occurrence
- The pattern of words between the subject entity and the occurrence (all words except for nouns, verbs, adjectives and adverbs are replaced with their POS tag, nouns are replaced with their named entity class if a named entity class is available)
- Any nouns, verbs, adjectives, adverbs or NEs in a 3-word window to the left of the occurrence
- Any nouns, verbs, adjectives, adverbs or NEs in a 3-word window to the right of the occurrence

### 4.6 Classifier and Models

As a classifier, we choose a first-order conditional random field model (Lafferty et al., 2001). We use the software CRFSuite<sup>4</sup> and L-BFGS (Nocedal, 1980) for training our classifiers. We train one classifier per Freebase class and model. Our models only differ in the way training data is selected (see Section 3). The models are then used to classify each object candidate into one of the relations of the Freebase class or NONE (no relation).

---

<sup>4</sup><http://www.chokkan.org/software/crfsuite/>

## 4.7 Merging and Ranking Results

We understand relation extraction as the task of predicting the relations which can be found in a corpus. While some approaches aim at correctly predicting every single mention of a relation separately, we instead choose to aggregate predictions of relation mentions. For every Freebase class, we get all relation mentions from the corpus and the classifier’s confidence values for Freebase classes assigned to object occurrences. There are usually several different predictions, e.g. the same occurrence could be predicted to be *MusicalArtist:album*, *MusicalArtist:origin* and *MusicalArtist:NONE*. By aggregating relation mentions across documents we have increased chances of choosing the right relation, since some contexts of occurrences are inconclusive or ambiguous and thus the classifier chooses the wrong property wfor those. For a given lexicalisation  $l$ , representing an object to which the subject is related, the classifier gives each object occurrence a prediction which is the combination of a predicted relation and a confidence. We collect these across the chosen documents to form a set of confidence values, for each predicted relation, per lexicalisation  $E_p^l$ . For instance if the lexicalisation  $l$  occurs three times across the documents and is predicted to represent an object to relation  $p_1$  once with confidence 0.2, and in other cases to represent the object to relation  $p_2$  with confidence 0.1 and 0.5 respectively, then  $E_{p_1}^l = 0.2$  and  $E_{p_2}^l = \{0.1, 0.5\}$ . In order to form an aggregated confidence for each relation with respect to the lexicalisation,  $g_p^l$ , we calculate the mean average for each such set and normalise across relations, as follows:  $g_p^l = \overline{E_p^l} \cdot \frac{|E_p^l|}{\sum_{q \in P} |E_q^l|}$

## 5 Evaluation

### 5.1 Corpus

Although we automatically annotate the training and test part of the Web corpus with properties, we hand-annotate a portion of the test corpus. The portion of the corpus we manually annotate is the one which has *NONE* predictions for object occurrences, i.e. for which occurrences do not have a representation in Freebase. They could either get *NONE* predictions because they are not relation mentions, or because they are missing from Freebase. We find that on average 45% of the occurrences which are predicted by our models are true relation mentions, but missing from Freebase. Note that some of the automatically evaluated positive predictions could in fact be false positives.

### 5.2 Results

Figures 1 and 2 show the precision with respect to confidence and precision@K for our self-supervised relation extraction models which only differ in the way training data is selected, as described in Section 3.

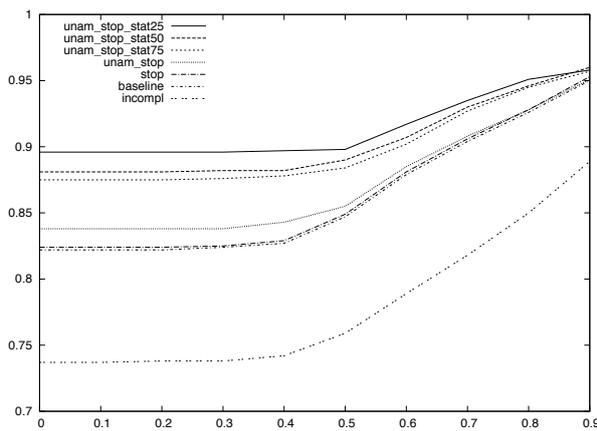


Figure 1: Precision / confidence graph

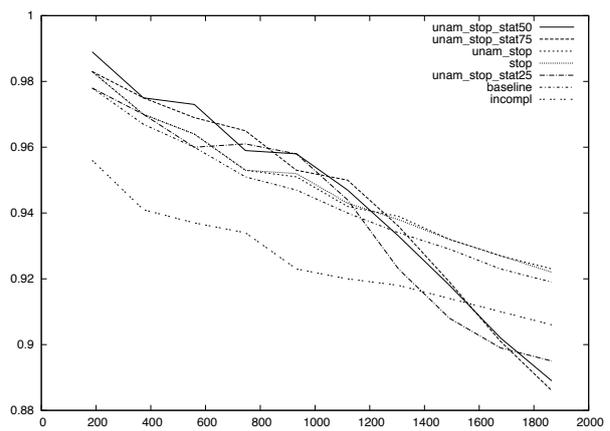


Figure 2: Precision@K

Figure 1 shows the precision of our models on the y-axis with respect to a cutoff at a minimum confidence displayed on the x-axis. The precision at a minimum confidence of 0 is equivalent to the precision over the whole test corpus. For all of the models, the precision rises with increasing confidence, which means our confidence measure succeeds at ranking results by precision. With respect to the baseline

which does not filter seeds, our best-performing model increases the total precision from 0.825 to 0.896, which is a total error reduction of 35%. We achieve the best results in terms of total precision with the model `unam_stop_stat25`, which filters lexicalisations which are ambiguous for a subject, filters lexicalisations which are stop words and filters lexicalisations with an ambiguity value higher than the lower quartile of the distribution for the relation in question. The worst-performing model is, surprisingly, *incompl*, the model we built to discard negative training data which are likely to be true relation mentions, but missing from the knowledge base. We discuss this further in Section 8. Figure 2 shows the precision, sorted by confidence, for the K highest ranked documents. We decided to use the precision@K measure instead of computing recall because it is not feasible to manually annotate 450,000 Web pages with relation mentions. Note, however, that distant supervision is not aimed at high recall anyway - because only sentences which contain both the subject and the object entity explicitly are used, many relation mentions will be missed out on. The highest value on the x-axis is the number of predicted relations of the model with the smallest number of predicted relations. The models which filter seeds improve all above the baseline in terms of precision@K for 0% to about 65% of the maximum K, from 65% to 100%, only stop and unam\_stop improve on the baseline.

## 6 Discussion

Although we cannot directly compare our results to that of other distantly supervised relation extraction models because we use different evaluation data and a different set of relations, our baseline model, which has a total precision of 0.825, as well as our best-performing model, which has a total precision of 0.896 seem to perform as well as, if not better than previous systems. Overall, our seed selection methods seem to perform well at removing unreliable training data to improve precision.

What is still unsuccessful is our *incompl* model. The idea behind it was that relations which, for a given subject, have more than one object (e.g. *Book:genre*) are prone to be “incomplete” - the objects in the knowledge base are often just the most prominent ones and other objects, which could be discovered from text, are missing. When annotating training data for distant supervision, those missing objects would be considered negative training data, which could potentially be harmful for training. However, just assuming that all negative training examples could potentially be false negatives if they match one of the objects does not lead to improved results. One of the reasons for this could be that most of those potential false negatives are instead objects of relations which expect the same kinds of values - and thus crucial for training the models. Some relations for which we observed this are *Book:originalLanguage* and *Book:translations*, as well as *Book:firstPublicationDate* and *Book:dateWritten*. Interestingly, neither *Book:originalLanguage* nor *Book:firstPublicationDate* are n:n relations.

## 7 Related Work

While lots of approaches in the past have focused on supervised, unsupervised (Yates et al., 2007; Fader et al., 2011) or semi-supervised relation extraction (Hearst, 1992; Carlson et al., 2010), there have also been some distantly supervised relation extraction approaches in the past few years, which aim at exploiting background knowledge for relation extraction, most of them for extracting relations from Wikipedia.

Mintz et al. (2009) describe one of the first distant supervision approaches which aims at extracting relations between entities in Wikipedia for the most frequent relations in Freebase. They report precision of about 0.68 for their highest ranked 10% of results depending what features they used. In contrast to our approach, Mintz et al. do not experiment with changing the distance supervision assumption or removing ambiguous training data, they also do not use fine-grained relations and their approach is not class-based. Nguyen et al. (2011)’s approach is very similar to that of Mintz et al. (2009), except that they use a different knowledge base, YAGO (Suchanek et al., 2008). They use a Wikipedia-based NERC, which, like the Stanford NERC classifies entities into persons, relations and organisations. They report a precision of 0.914 for their whole test set, however, those results might be skewed by the fact that YAGO is a knowledge based derived from Wikipedia.

A few strategies for seed selection for distant supervision have already been investigated: At-least-one models (Hoffmann et al., 2011; Surdeanu et al., 2012; Riedel et al., 2010; Yao et al., 2010; Min et al., 2013),

hierarchical topic models (Alfonseca et al., 2012; Roth and Klakow, 2013), pattern correlations (Takamatsu et al., 2012), and an information retrieval approach (Xu et al., 2013). At-least-one models (Hoffmann et al., 2011; Surdeanu et al., 2012; Riedel et al., 2010; Yao et al., 2010; Min et al., 2013) are based on the idea that “if two entities participate in a relation, at least one sentence that mentions these two entities might express that relation”. While positive results have been reported for those models, Riedel et al. (Riedel et al., 2010) argues that it is challenging to train those models because they are quite complex. Hierarchical topic models (Alfonseca et al., 2012; Roth and Klakow, 2013) assume that the context of a relation is either specific for the pair of entities, the relation, or neither. Min et al. (Min et al., 2013) further propose a 4-layer hierarchical model to only learn from positive examples to address the problem of incomplete negative training data. Pattern correlations (Takamatsu et al., 2012) are also based on the idea of examining the context of pairs of entities, but instead of using a topic model as a pre-processing step for learning extraction patterns, they first learn patterns and then use a probabilistic graphical model to group extraction patterns. Xu et al. (Xu et al., 2013) propose a two-step model based on the idea of pseudo-relevance feedback which first ranks extractions, then only uses the highest ranked ones to re-train their model. Our research is based on a different assumption: Instead of trying to address the problem of noisy training data by using more complicated multi-stage machine learning models, we want to examine how background data can be even further exploited by testing if simple statistical methods based on data already present in the knowledge base can help to filter unreliable training data.

## 8 Future Work

In this paper, we have documented and evaluated an approach to discard unreliable seed data for distantly supervised relation extraction. Our two hypotheses were that discarding highly ambiguous relation mentions and discarding unreliable negative training seeds could help to improve precision of self-supervised relation extraction models. While our evaluation indicates that discarding highly ambiguous relation mentions based on simple statistical methods helps to improve the precision of distantly supervised relation extraction systems, discarding negative training data does not. We have also described our distantly supervised relation extraction system, which, unlike other previous systems learns to extract from Web pages and also learns to extract fine-grained relations for specific classes instead of relations which are applicable to several broad classes.

In future work, we want to work on increasing the number of extractions for distant supervision systems: The distant supervision assumption requires sentences to contain both the subject and the object of a relation. While this ensures high precision and is acceptable for creating training data, most sentences - at least those in Web documents - do not mention the subject of relations explicitly and we thus miss out on a lot of data to extract from. We further want to extend our distant supervision approach to extract information not only from free text, but also from lists and relational tables from Web pages. Finally, we would like to train distantly supervised models for entity classification to assist relation extraction. A more detailed description of future work can also be found in Augenstein (2014).

## Acknowledgements

We thank Barry Norton, Diana Maynard, as well as the anonymous reviewers for their valuable feedback. This research was partly supported by the EPSRC funded project LODIE: Linked Open Data for Information Extraction, EP/J019488/1.

## References

- Enrique Alfonseca, Katja Filippova, Jean-Yves Delort, and Guillermo Garrido. 2012. Pattern Learning for Relation Extraction with a Hierarchical Topic Model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, volume 2, pages 54–59.
- Isabelle Augenstein. 2014. Joint Information Extraction from the Web using Linked Data. *Doctoral Consortium Proceedings of the 13th International Semantic Web Conference*. to appear.

- Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.
- A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr., and T.M. Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. In *Proceedings of the Conference on Artificial Intelligence*, pages 1306–1313.
- Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.
- Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545.
- Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke S. 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*, pages 541–550.
- John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.
- D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5:361–397.
- 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 HLT-NAACL*, pages 777–782.
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP*, volume 2, pages 1003–1011.
- Truc-Vien T. Nguyen and Alessandro Moschitti. 2011. End-to-End Relation Extraction Using Distant Supervision from External Semantic Repositories. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, volume 2, pages 277–282.
- Jorge Nocedal. 1980. Updating Quasi-Newton Matrices with Limited Storage. *Mathematics of Computation*, 35(151):773–782.
- 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*, pages 148–163.
- 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.
- Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. YAGO: A Large Ontology from Wikipedia and WordNet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217.
- Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance Multi-label Learning for Relation Extraction. In *Proceedings of EMNLP-CoNLL*, pages 455–465.
- 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*, pages 721–729.
- Fei Wu and Daniel S. Weld. 2007. Autonomously Semantifying Wikipedia. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, pages 41–50.
- 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*, pages 665–670.
- Limin Yao, Sebastian Riedel, and Andrew McCallum. 2010. Collective Cross-document Relation Extraction Without Labelled Data. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1013–1023.
- Alexander Yates, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen Soderland. 2007. TextRunner: Open Information Extraction on the Web. In *Proceedings of HLT-NAACL: Demonstrations*, pages 25–26.