Semantic Web 7 (2016) 335–349
DOI 10.3233/SW-150180
IOS Press

Distantly supervised Web relation extraction for knowledge base population

Editor(s): Krzysztof Janowicz, STKO Lab, University of California, Santa Barbara, USA; Stefan Schlobach, Vrije Universiteit Amsterdam, The Netherlands
Solicited review(s): Ulrich Reimer, University of Applied Sciences St. Gallen, Switzerland; Antske Fokkens, Vrije Universiteit Amsterdam, The Netherlands; one anonymous reviewer

Isabelle Augenstein [*], Diana Maynard and Fabio Ciravegna
Department of Computer Science, The University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, United Kingdom
E-mails: i.augenstein@sheffield.ac.uk, d.maynard@sheffield.ac.uk, f.ciravegna@sheffield.ac.uk

Abstract. Extracting information from Web pages for populating large, cross-domain knowledge bases requires methods which are suitable across domains, do not require manual effort to adapt to new domains, are able to deal with noise, and integrate information extracted from different Web pages. Recent approaches have used existing knowledge bases to learn to extract information with promising results, one of those approaches being distant supervision. Distant supervision is an unsupervised method which uses background information from the Linking Open Data cloud to automatically label sentences with relations to create training data for relation classifiers. In this paper we propose the use of distant supervision for relation extraction from the Web. Although the method is promising, existing approaches are still not suitable for Web extraction as they suffer from three main issues: data sparsity, noise and lexical ambiguity. Our approach reduces the impact of data sparsity by making entity recognition tools more robust across domains and extracting relations across sentence boundaries using unsupervised co-reference resolution methods. We reduce the noise caused by lexical ambiguity by employing statistical methods to strategically select training data. To combine information extracted from multiple sources for populating knowledge bases we present and evaluate several information integration strategies and show that those benefit immensely from additional relation mentions extracted using co-reference resolution, increasing precision by 8%. We further show that strategically selecting training data can increase precision by a further 3%.

Keywords: Knowledge base population, distant supervision, relation extraction, Web-based methods, Linked Open Data, Freebase, unsupervised learning, natural language processing

1. Introduction

In the past years, several cross-domain knowledge bases such as Freebase [7], DBpedia and Wikidata [37] have been constructed by Web companies and research communities for purposes such as search and question answering. Even the largest knowledge bases are far from complete, since new knowledge is emerging rapidly. Most of the missing knowledge is available on Web pages in the form of free text. To access that knowledge, information extraction (*IE*) and information integration methods are necessary. In this paper, we focus on the task of relation extraction (*RE*), that is to extract individual mentions of relations from text, and also present how those individual mentions can be integrated and redundancy of information across Web documents can be exploited to extract facts for knowledge base population. One important aspect to every relation extraction approach is how to annotate training and test data for learning classifiers. In the past, four groups of approaches have been proposed (see also Section 2).

[*] Corresponding author. E-mail: i.augenstein@sheffield.ac.uk.

1570-0844/16/$35.00 © 2016 – IOS Press and the authors.

Supervised approaches use manually labelled training and test data. Those approaches are often specific for, or biased towards a certain domain or type of text. This is because IE approaches tend to have a higher performance if training and test data is restricted to the same narrow domain. In addition, developing supervised approaches for different domains requires even more manual effort.

Unsupervised approaches do not need any annotated data for training and instead extract words between entity mentions, then cluster similar word sequences and generalise them to relations. Although unsupervised approaches can process very large amounts of data, the resulting relations are hard to map to ontologies. In addition, it has been documented that these approaches often produce uninformative as well as incoherent extractions [13].

Semi-supervised methods only require a small number of seed instances. The hand-crafted seeds are used to extract patterns from a large corpus, which are then used to extract more instances and those again to extract new patterns in an iterative way. The selection of initial seeds is very challenging – if they do not accurately reflect the knowledge contained in the corpus, the quality of extractions might be low. In addition, since many iterations are needed, these methods are prone to semantic drift, i.e. an unwanted shift of meaning. This means these methods require a certain amount of human effort – to create seeds initially and also to help keep systems "on track" to prevent them from semantic drift.

A fourth group of approaches are *distant supervision* or *self-supervised* learning approaches [30]. The idea is to exploit large knowledge bases (such as Freebase [7]) to automatically label entities in text and use the annotated text to extract features and train a classifier. Unlike supervised systems, these approaches do not require manual effort to label data and can be applied to large corpora. Since they extract relations which are defined by vocabularies, these approaches are less likely to produce uninformative or incoherent relations.

Although promising, distant supervision approaches have several limitations with respect to Web IE that require further research. This work improves on existing distant supervision approaches by addressing four challenges, illustrated with the following example:

> "*Let It Be* is the twelfth and final album by *The Beatles* which contains their hit single '*Let it Be*'. They broke up in 1974."

Unrecognised Entities: Distant supervision approaches tend to use named entity classifiers that recognise entities that were trained for the news domain. Those typically label entities as either persons, locations, organisations or mixed. When applying those approaches to heterogenous Web pages, types of entities which fall into the "mixed" category and also subclasses of person, location and organisation are often not recognised. Two of those types used for the experiments described in this paper are *MusicalArtist:track* and *MusicalArtist:album*.

Restrictive assumption: Existing distant supervision systems [30] only learn to extract relations which do not cross sentences boundaries, i.e. sentences which contain an explicit mention of the name of both the subject and the object of a relation. This results in data sparsity. In the example above, the second sentence does not contain two named entities, but rather a pronoun representing an entity and an NE. While existing co-reference resolution tools could be applied to detect the NE the pronoun refers to, this is only possible if those named entities are detected in the first place.

Ambiguity: In the first sentence, the first mention of *Let It Be* is an example for the *MusicalArtist:album* relation, whereas the second mention is an example of the *MusicalArtist:track* relation. If both mentions are used as positive training data for both relations, this impairs the learning of weights of the relation classifiers. This aspect has already been partly researched by existing distant supervision approaches [30].

Setting: Existing distant supervision approaches generally assume that every text might contain information about any possible property. Making this assumption means that the classifier has to learn to distinguish between all possible properties, which is unfeasible with a large domain and a big corpus.

The contributions of this paper to research on distant supervision for Web information extraction are: (1) recognising named entities across domains on heterogeneous Web pages by using Web-based heuristics; (2) reporting results for extracting relations across sentence boundaries by relaxing the distant supervision assumption and using heuristic co-reference resolution methods; (3) proposing statistical measures for increasing the precision of distantly supervised systems by filtering ambiguous training data, (4) documenting an entity-centric approach for Web relation extraction using distant supervision; and (5) evaluating distant supervision as a knowledge base population approach and evaluating the impact of our different methods on information integration.

2. Related work

There are have been several different approaches for IE from text for populating knowledge bases which try to minimise manual effort in the recent past.

Semi-supervised bootstrapping approaches such as KnowItAll [12], NELL [9], PROSPERA [23] and BOA [17] start with a set of seed natural language patterns, then employ an iterative approach to both extract information for those patterns and learn new patterns. For KnowItAll, NELL and PROSPERA, the patterns and underlying schema are created manually, whereas they are created automatically for BOA by using knowledge contained in DBpedia.

Ontology-based question answering systems often use patterns learned by semi-supervised information extraction approaches as part of their approach. Unger et al. [35], for instance, use patterns produced by BOA.

Open information extraction (Open IE) approaches such as TextRunner [43], Kylin [39], StatSnowball [44], Reverb [13], WOE [40], OLLIE [20] and ClausIE [11] are unsupervised approaches, which discover relation-independent extraction patterns from text. Although they can process very large amounts of data, the resulting relations are hard to map to desired ontologies or user needs, and can often produce uninformative or incoherent extractions, as mentioned in Section 1.

Bootstrapping and Open IE approaches differ from our approach in the respect that they learn extraction rules or patterns, not weights for features for a machine learning model. The difference between them is that statistical approaches take more different factors into account to make 'soft' judgements, whereas rule- and pattern-based approaches merge observed contexts to patterns, then only keep the most prominent patterns and make hard judgments based on those. Because information is lost in the pattern merging and selection process, statistical methods are generally more robust to unseen information, i.e. if the training and test data are drawn from different domains, or if unseen words or sentence constructions occur. We opt for a statistical approach, since we aim at extracting information from heterogenous Web pages.

Automatic ontology learning and population approaches such as FRED [25,26] and LODifier [5] extract an ontology schema from text, map it to existing schemas and extract information for that schema. Unlike bootstrapping approaches, they do not employ an iterative approach. However, they rely on several existing natural language processing tools trained on newswire and are thus not robust enough for Web IE.

Finally, *distantly supervised or self-supervised approaches* aim at exploiting background knowledge for RE, most of them for extracting relations from Wikipedia. Mintz et al. [22] aim 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 distant 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. [24]'s approach is very similar to that of Mintz et al., except that they use a different knowledge base, YAGO [32]. They use a Wikipedia-based named entity recogniser and classifier (*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 base derived from Wikipedia. In addition to Wikipedia, distant supervision has also been used to extract relations from newswire [27,28], to extract relations for the biomedical domain [10,29] and the architecture domain [36]. Bunescu and Mooney [8] document a minimal supervision approach for extracting relations from Web pages, but only apply it to the two relations *company-bought-company* and *person-bornIn-place*. Distant supervision has also been used as a pre-processing step for learning patterns for bootstrapping and Open IE approaches, e.g. Kylin, WOE and BOA annotate text with DBpedia relations to learn patterns.

A few strategies for seed selection for distant supervision have already been investigated: at-least-one models [18,21,27,33,42], hierarchical topic models [1,31], pattern correlations [34], and an information retrieval approach [41]. At-least-one models 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" [27]. While positive results have been reported for those models, Riedel et al. [27] argue that it is challenging to train those models because they are quite complex. Hierarchical topic models [1,31] assume that the context of a relation is either specific for the pair of entities, the relation, or neither. Min et al. [21] further propose a 4-layer hierarchical model to only learn from positive examples to address the problem of incomplete negative training data. Pattern correlations [34] 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. [41] 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. Preliminary results for this have already been reported in Augenstein et al. [3,4]. The benefit of this approach compared with other approaches is that it does not result in an increase of run-time during testing and is thus more suited towards Web-scale extraction than approaches which aim at resolving ambiguity during both training and testing. To the best of our knowledge, our approach is the first distant supervision approach to address the issue of adapting distant supervision to relation extraction from heterogeneous Web pages and to address the issue of data sparsity by relaxing the distant supervision assumption.

3. Distantly supervised relation extraction

Distantly supervised relation extraction is defined as automatically labelling a corpus with properties, P and resources, R, where resources stand for entities from a knowledge base, *KB*, to train a classifier to learn to predict binary relations. The distant supervision paradigm is defined as follows [22]:

> If two entities participate in a relation, any sentence that contains those two entities might express that relation.

In general relations are of the form $(s, p, o) \in R \times P \times R$, consisting of a subject, a predicate and an object; during training, we only consider statements which are contained in a knowledge base, i.e. $(s, p, o) \in KB \subset R \times P \times R$. In any single extraction we consider only those subjects in a particular class $C \subset R$, i.e. $(s, p, o) \in KB \cap C \times P \times R$. Each resource $r \in R$ has a set of lexicalisations, $L_r \subset L$. Lexicalisations are retrieved from the *KB*, where they are represented as the name or alias, i.e. less frequent name of a resource.

In the remainder of this paper, several adjustments to this approach are presented, method names are indicated in bold font.

3.1. Seed selection

Before using the automatically labelled corpus to train a classifier, we detect and discard examples containing highly ambiguous lexicalisations. 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. We measure the number of senses by the number of unique resources representing a lexicalisation.

Ambiguity within an entity

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*.

Unam: For a given subject s, if we discover a lexicalisation for a related entity o, 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: $A_l^s = |\{r \mid l \in L_o \cap L_r \wedge (s, p, o) \in KB \wedge (s, q, r) \in KB \wedge r \neq o\}|$.

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 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. [19], which was originally created for the purpose of information retrieval and contains 571 highly frequent words.

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 with other lexicalisations of objects of the same relation. If its ambiguity is comparatively low, we consider it a reliable seed, otherwise we 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 $A_o = |\{o \mid l \in L_o\}|$.

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.

StatRes: Since Stat is mainly aimed at n-ary relations, for which many seeds are available, we want to restrict the impact of Stat for relations with only few object lexicalisations per relation. We compute the number of object lexicalisations per property and view this as a frequency distribution with min, max, median, lower and upper quartile. If the number of object lexicalisations at the upper quartile for a relation is 2 or smaller, we do not discard any seeds for that relation. We apply this method for all variants of StatRes.

3.2. *Relaxed setting*

In addition to increasing the precision of distantly supervised systems by filtering seed data, we also experiment with increasing recall by changing the method for creating test data. Instead of testing, for every sentence, if the sentence contains a lexicalisation of the subject and one additional entity, we relax the former restriction. We make the assumption that the subject of the sentence is mostly consistent within one paragraph as the use of paragraphs usually implies a unit of meaning, i.e. that sentences in one paragraph often have the same subject. In practice this means that we first train classifiers using the original assumption and then, for testing, instead of only extracting information from sentences which contain a lexicalisation of the subject, we also extract information from sentences which are in the same paragraph as a sentence which contains a lexicalisation of the subject.

Our new relaxed distant supervision assumption is then:

> If two entities participate in a relation, any *paragraph* that contains those two entities might express that relation, even if not in the same sentence, provided that another sentence in the paragraph in itself contains a relationship for the same subject.

This means, however, that we have to resolve the subject in a different way, e.g. by performing co-reference resolution and searching for a pronoun which is coreferent with the subject mention in a different sentence. We test four different methods for our relaxed setting, one of which does not attempt to resolve the subject of sentences, one based on an existing co-reference resolution tool, and two based on gazetteers of Web co-occurrence counts for number and gender of noun phrases.

NoSub: Instead of trying to perform co-reference resolution, our first approach does not attempt to find the subject of the sentence at all. We instead disregard all features which require the position of the subject mention to be known. Features used in both the NoSub setting and the normal setting are documented in Section 4.6.

CorefS: To test how useful existing co-reference resolution tools are for a variety of different classes and properties, we perform co-reference using the Stanford NLP co-reference resolution tool. For every sentence in a paragraph that contains at least one sentence with the subject entity, if any of the sentences contain a pronoun or noun phrase that is coreferent with the subject entity, we treat it as if it were a lexicalisation of the subject entity and extract all features we also extract for the normal setting.

CorefN and **CorefP**: Since the Stanford NLP co-reference resolution tool is a supervised approach trained on the news domain, it might not be able to resolve co-references for some of the classes we are using. Since we do not have training data for all of our domains, we use a heuristic based on Web co-occurrence counts using the gazetteers collected by Bergsma and Dekang [6].

The first step in co-reference resolution is usually to group all mentions in a text, i.e. all noun phrases and pronouns by gender and number. If two mentions disagree in number or gender, they cannot be coreferent. As an example, should we find "The Beatles" and "he" in a sentence, then "The Beatles" and "he" could not be coreferent, because "The Beatles" is a plural neutral noun phrase, whereas "he" is a singular male pronoun. Since we do not have any a-priori information on what number and gender the subject entity is, we instead make those judgments based on the number and gender of the class of the subject, e.g. The Beatles is a Musical Artist, which can be a band (plural) or a female singer or a male singer. Bergsma and Dekang have collected such a resource automatically, which also includes statistics to assess how likely it is for a noun phrase to be a certain number or gender. In particular, they collected co-occurrence counts of different noun phrases with *male*, *female*, *neutral* and *plural* pronouns using Web search. Our heuristic co-reference approach consists of three steps. First, we collect noun phrases which express general concepts related to the subject entity, which we refer to as *synonym gazetteer*. We start with the lexicalisation of the class of the entity (e.g. "Book"), then retrieve synonyms, hypernyms and hyponyms using Wikipedia redirection pages and WordNet [14]. Second, we determine the gender of each class by looking up co-occurrence counts for each general concept in the noun phrase, gender and number gazetteer. We aggregate the co-reference counts for each class and gender or number (i.e. male, female, neutral, plural). If the aggregated count for each number or gender is at least 10% of the total count for all genders and numbers, we consider that gender or number to *agree with* the class. For each class, we then create a *pronoun gazetteer* containing all male, female, neutral or plural personal pronouns including possessives, e.g. for "Book", that gazetteer would contain "it, its, itself". Lastly, we use those gazetteers to resolve co-reference. For every sentence in a paragraph that contains at least one sentence with the subject entity, if any of the following sentences contain a pronoun or noun phrase that is part of the synonym or pronoun gazetteer for that class and it appears in the sentence before the object lexicalisation, we consider that noun phrase or pronoun coreferent with the subject. The reason to only consider noun phrases or pronouns to be coreferent with the subject entity if they appear after the object entity is to improve precision, since anaphora (expressions referring back to the subject) are far more common than cataphora (expressions referring to the subject appearing later in the sentence).

We test two different methods. CorefN only uses the synonym gazetteer, whereas CorefP uses both the synonym and the pronoun gazetteer. If a sentence contains both a synonym and a pronoun, the synonym is selected as coreferent for the subject. We then, as for CorefS, treat those noun phrases and pronouns as lexicalisations of the subject and extract all features also used for the normal setting.

3.3. Information integration

After features are extracted, a classifier is trained and used to predict relation mentions, those predictions can be used for the purpose of knowledge base population by aggregating relation mentions to relations. Since the same relations might be found in different documents, but some contexts might be inconclusive or ambiguous, it is useful to integrate information taken from multiple predictions to increase the chances of predicting the correct relation. We test several different methods to achieve this.

Comb: Instead of integrating extractions, feature vectors for the same relation tuples are aggregated for training and testing.

Aggr: For every Freebase class, we get all relation mentions from the corpus and the classifier's confidence values for 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. 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 $E_{p_2}^l = \{0.1, 0.5\}$. We then only select the relation p with the highest single confidence value $E > 0.5$. In order to form an aggregated confidence for each relation with respect to the lexicalisation, g_l^p, we calculate the mean average for each such set and normalise across relations, as follows: $g^l_p = \overline{E^l_p} \cdot \frac{|E^l_p|}{\sum_{q \in P} |E^l_q|}$. For each lexicalisation l, we select the relation p with the highest confidence g^l_p.

Limit: One of the shortcomings of Aggr is that it returns all possible aggregated predictions for each relation, which sometimes means too many predictions are returned. To address this, we compute the number of object lexicalisations per property and view it as a frequency distribution, compute the maximum and upper quartile of that distribution, then sort all predictions by confidence value in descending order. We then select highest ranked n predictions to return, starting with the one with the highest confidence value. For LimitMax n is the maximum of the object lexicalisation per property frequency distribution, whereas for Limit75 it is the upper quartile.

Multilab: Another shortcoming of Aggr is that it only allows to predict one label per aggregated prediction, i.e. *Let it Be* will either be predicted to be MusicalArtist:album or MusicalArtist:track, but not both. While it is possible to train a multi-label classifier with noisy, ambiguous examples [33], another option, which we are pursuing, is to discard those examples for training, and to integrate them for testing post hoc. To find out which relations have any object lexicalisations overlapping with other relations, this information about *mutual labels* is collected from the part of Freebase used for training. After predictions are aggregated using Aggr, instead of only returning the label with highest confidence, all possible labels are sorted by confidence value. If the label with highest confidence and the one with second highest confidence are mutual labels, both of them are returned, afterwards, if the label with highest confidence and the one with third highest confidence are mutual labels, the label with third highest confidence is also returned.[1]

4. System

4.1. Corpus

Table 1
Freebase classes and properties used

Person	
Musical Artist : album	Politician : birthdate
Musical Artist : active (start)	Politician : birthplace
Musical Artist : active (end)	Politician : educational institution
Musical Artist : genre	Politician : nationality
Musical Artist : record label	Politician : party
Musical Artist : origin	Politician : religion
Musical Artist : track	Politician : spouses
Organisation	
Business : industry	Education : school type
Business : employees	Education : mascot
Business : city	Education : colors
Business : country	Education : city
Business : date founded	Education : country
Business : founders	Education : date founded
Mixed	
Film : release date	Book : author
Film : director	Book : characters
Film : producer	Book : publication date
Film : language	Book : genre
Film : genre	Book : original language
Film : actor	
Film : character	
Location	
River : origin	
River : mouth	
River : length	
River : basin countries	
River : contained by	

To create a corpus for Web relation extraction using background knowledge from Linked Data, seven Freebase classes and their five to seven most prominent properties are selected, as shown in Table 1. The selected classes are subclasses of either "Person" (Musical Artist, Politician), "Location" (River), "Organisation" (Business (Operation)), Education(al Institution)) or "Mixed" (Film, Book). To avoid noisy training data, we only use entities which have values for all of those properties and retrieve them using the Freebase API. This resulted in 1800 to 2200 entities per class. For each entity, at most 10 Web pages were retrieved via the Google Search API using the search pattern "'*subject_entity*" *class_name relation_name*', e.g. "'The Beatles" Musical Artist Origin'. By adding the class name, we expect the retrieved Web pages to be more relevant to our extraction task. Although subject entities can have multiple lexicalisations, 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. In total, the corpus consists of around one

[1] There is only one instance of three mutual labels for our evaluation set, namely *River:origin*, *River:countries* and *River:contained by*.

Table 2

Distribution of websites per class in the Web corpus sorted by frequency

Musical Artist		Politician	
21	en.wikipedia.org	17	en.wikipedia.org
6	itunes.apple.com	4	www.huffingtonpost.com
5	www.allmusic.com	3	votesmart.org
4	www.last.fm	3	www.washingtonpost.com
3	www.amazon.com	2	www.nndb.com
2	www.debate.org	2	www.evi.com
2	www.reverbnation.com	2	www.answers.com
57	Others	67	Others
Business		**Education**	
13	en.wikipedia.org	23	en.wikipedia.org
6	www.linkedin.com	8	www.linkedin.com
2	www.indeed.com	4	colleges.usnews.rankingsandreviews.com
2	www.glassdoor.co.uk	1	www.forbes.com
1	connect.data.com	1	www.facebook.com
1	www.answers.com	1	www.greatschools.org
1	www.forbes.com	1	www.trulia.com
74	Others	61	Others
Film		**Book**	
15	en.wikipedia.org	20	en.wikipedia.org
15	www.imdb.com	15	www.goodreads.com
3	www.amazon.com	12	www.amazon.com
3	www.rottentomatoes.com	9	www.amazon.co.uk
1	www.amazon.co.uk	4	www.barnesandnoble.com
1	www.tcm.com	3	www.abebooks.co.uk
1	www.nytimes.com	2	www.abebooks.com
61	Others	28	Others
River			
24	en.wikipedia.org		
2	www.britannica.com		
1	www.researchgate.net		
1	www.facebook.com		
1	www.gaiagps.com		
1	www.tripadvisor.co.uk		
1	www.encyclo.co.uk		
69	Other		

million pages drawn from 76,000 different websites. An overview of the distribution of websites per class is given in Table 2.

4.2. NLP pipeline

Text content is extracted from HTML pages using the Jsoup API,[2] which strips text from each element recursively. Each paragraph is then processed with Stanford CoreNLP[3] to split the text into sentences, tokenise it, annotate it with part of speech (POS) tags and normalise time expressions. Named entities are classified using the 7 class (time, location, organisation, person, money, percent, date) named entity model. For the relaxed setting (Section 3.2), co-references are resolved using Stanford coref.

4.3. Relation candidate identification

Some of the relations we want to extract values for cannot be categorised according to the 7 classes detected by the Stanford NERC and are therefore not recognised. An example for this is *MusicalArtist:album*, *MusicalArtist:track* or *MusicalArtist: genre*. Therefore, as well as recognising named entities with Stanford NERC as relation candidates, we also implement our own 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. 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. For greedy matching, we consider whole noun phrases and sequences of capitalised words. For non-greedy matching, we consider all subsequences starting with the first word of the those phrases as well as single tokens, i.e. for ‘science fiction book’, we would consider ‘science fiction book’, ‘science fiction’, ‘science’, ‘fiction’ and ‘book’ as candidates. 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. Annotating sentences

The next step is to identify which sentences express relations. We only use sentences from Web pages which were retrieved using a query which contains the subject of the relation. To annotate sentences, we retrieve all lexicalisations L_s, L_o for subjects and objects related under properties P for the subject’s class

[2] http://jsoup.org

[3] http://nlp.stanford.edu/software/corenlp.shtml

C from Freebase. We then check, for each sentence, if it contains at least two entities recognised using either the Stanford NERC or our own entity recogniser (Section 4.3), one of which having a lexicalisation of a subject and the other a lexicalisation of an object of a relation. If it does, we use this sentence as training data for that property. All sentences which contain a subject lexicalisation and one other entity that is not a lexicalisation of an object of any property of that subject are used as negative training data for the classifier. Mintz et al. [22] only use 1% of their negative training data, but we choose to deviate from this setting because we have less training data overall and have observed that using more negative training data increases precision and recall of the system. For testing we use all sentences that contain at least two entities recognised by either entity recogniser, one of which must be a lexicalisation of the subject. For our relaxed setting (Section 3.2) only the paragraph the sentence is in must contain a lexicalisation of the subject.

4.5. *Seed selection*

After training data is retrieved by automatically annotating sentences, we select seeds from it, or rather discard some of the training data, according to the different methods outlined in Section 3.1. Our Baseline models do not discard any training seeds.

4.6. *Features*

Given a relation candidate as described in Section 4.3, our system then extracts the following lexical features and named entity features, some of them also used by Mintz et al. [22]. Features marked with (*) are only used in the normal setting, but not for the NoSub setting (Section 3.2).

- 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 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 named entities in a 3-word window to the left of the occurrence
- Any nouns, verbs, adjectives, adverbs or named entities in a 3-word window to the right of the occurrence

In comparison with Mintz et al. [22] we use richer feature set, specifically more bag of words features, patterns, a numerical feature and a different, more fine-grained named entity classifier.

We experiment both with predicting properties for relations, as in Mintz et al. [22], and with predicting properties for relation mentions. Predicting relations means that feature vectors are aggregated for relation tuples, i.e. for tuples with the same subject and object, for training a classifier. In contrast, predicting relation mentions means that feature vectors are not aggregated for relation tuples. While predicting relations is sufficient if the goal is only to retrieve a list of values for a certain property, and not to annotate text with relations, combining feature vectors for distant supervision approaches can introduce additional noise for ambiguous subject and object occurrences.

4.7. *Models*

Our models differ with respect to how sentences are annotated for training, how positive training data is selected, how negative training data is selected, which features are used, how sentences are selected for testing and how information is integrated.

Mintz: This model follows the setting of the model which only uses lexical features described in Mintz et al. [22]. Sentences are annotated using the Stanford NERC [15] to recognise subjects and objects of relations, 1% of unrelated entities are used as negative training data and a basic set of lexical features is used. If the same relation tuple is found in several sentences, feature vectors extracted for those tuples are aggregated. For testing, all sentences containing two entities recognised by the Stanford NERC are used.

Baseline: This group of models follows the setting described in Section 4. It uses sentences annotated with both Stanford NERC and our NER (Section 4.3). All negative training data is used. For testing, all sentences containing two entities recognised by both Stanford NERC and our NER are used.

Comb: This group of models uses the same settings as Baseline models except that feature vectors for the same relation tuples are aggregated.

Aggr, Limit, MultiLab: These models use the same strategy for named entity recognition and selecting negative training data as the Comb group of models. However, feature vectors are not aggregated. Instead, labels are predicted for relation mentions and relations are predicted using the different information integration methods describe in Section 3.3.

Unam, Stop, Stat, StatRes: Those models select seed data according to the different strategies outlined in Section 3.1.

NoSub: This group of models uses the relaxed setting described in Section 3.2 which does not require sentences to explicitly contain subjects and only uses a restricted set of features for testing which do not require the position of the subject entity to be known.

CorefS: This is a variant of the relaxed setting, also described in Section 3.2 which uses Stanford coref to resolve co-references. The full set of features is extracted for testing.

CorefN, CorefP: Co-references are resolved for those variants of the relaxed setting using gender and number gazetteers. As for CorefS, the full set of features is extracted for testing.

4.8. *Predicting relations*

In order to be able to compare our results, we choose the same classifier as in Mintz et al. [22], a multi-class logistic regression classifier. We train one classifier per class and model. The models are used to classify each relation mention candidate into one of the relations of the class or NONE (no relation). Relation mention predictions are then aggregated to predict relations using the different information integration methods described in Section 3.3.

5. Evaluation

The goal of our evaluation is to measure how the different distant supervision models described in Section 4.7 perform for the task of knowledge base population, i.e. to measure how accurate the information extraction methods are at replicating the test part of the knowledge base. To this end we carry out a hold-out evaluation, for which 50% of the knowledge base is used for training and 50% for testing. We annotate the whole corpus with relations already present in Freebase, as described in Section 4 and use 50% of it for training and 50% for testing.

The following metrics are computed: precision, recall and a top line. Precision is defined as the number of correctly labelled relations divided by the number of correctly labelled plus the number of incorrectly labelled relations. Recall is defined as the number of correctly labelled relations divided by the number of all relation tuples in the knowledge base. The number of all relation tuples includes all different lexicalisations of objects contained in the knowledge base. To achieve a perfect recall of 1, all relation tuples in the knowledge base have to be identified as relation candidates in the corpus first. However, not all relation tuples also have a textual representation in the corpus. To provide insight into how many of them do, we compute a top line for recall. The top line would usually be computed by dividing the number of all relation tuples appearing in the corpus by the number of relation tuples in the knowledge base, as e.g. in [16]. The top line we provide is only an estimate, since the corpus is too big to examine each sentence manually. We instead compute the top line by dividing the number of relation tuples identified using the most inclusive relation candidate identification strategy, those used by theNoSub models, by the number of relation tuples in our test knowledge base.

Results for different seed selection models detailed in Section 4.7 averaged over all properties of each class are listed in Table 3. Model settings are incremental, i.e. the row Baseline lists results for the model Baseline, the row after that, + Stop lists results for the model Baseline using the seed selection method Stop, the row after that lists results for the seed selection methods Stop and Unam, and so forth. Results for different information integration models are listed in Table 4 and results for different co-reference resolution methods per class are listed in Table 5. Finally, Table 6 shows results for the best performing normal and the best performing model for the relaxed setting per Freebase class.

5.1. *Results*

From our evaluation results we can observe that there is a significant difference in terms of performance between the different model groups.

Table 3

Seed selection results: micro average of precision (P) and recall (R) over all relations, using the Multilab+Limit75 integration strategy and different seed selection models. The top line for recall is 0.0917

Model	P	R
Mintz	0.264	0.0359
Baseline	0.770	**0.0401**
+ Stop + Unam	0.773	0.0395
+ Stat75	**0.801**	0.0243
+ Stat50	0.801	0.0171
+ Stat25	0.767	0.00128
Baseline + Stop + Unam + StatRes75	0.784	0.0353
+ StatRes50	0.787	0.0341
+ StatRes25	0.78	0.0366
NoSub	0.645	0.0536
CorefS	0.834	0.0504
CorefN	0.835	0.0492
CorefP	0.830	**0.0509**
CorefN + Stop + Unam + Stat75	**0.857**	0.0289

Table 4

Information integration results: micro average of precision (P) and recall (R) over all relations, using the CorefN+Stop+Unam+Stat75 model and different information integration methods

Model	P	R
Comb	0.742	0.0328
Aggr	0.813	**0.0341**
LimitMax	0.827	0.0267
MultiLab	0.837	0.0307
Limit75 + MultiLab	**0.857**	0.0289

Table 5

Co-reference resolution results: micro average of precision (P) and recall (R) over all relations, using the CorefN+Stop+Unam+Stat75 model and different co-reference resolution methods

Class	CorefS		CorefN		CorefP	
	P	R	P	R	P	R
Musical Artist	0.736	0.0112	0.744	0.0112	**0.7473**	**0.01121**
Politician	**0.796**	**0.0577**	0.788	0.0498	0.788	0.0567
River	**0.890**	0.0902	0.889	0.902	0.873	**0.0932**
Business	0.849	0.1232	**0.861**	0.1352	0.856	**0.1593**
Education	0.927	**0.09**	**0.928**	0.0893	0.926	0.898
Book	**0.814**	0.0465	0.804	0.0461	0.808	**0.0484**
Film	0.8	0.0405	**0.0803**	0.0411	0.795	**0.0415**

Table 6

Best overall results: micro average of precision (P), recall (R) and top line recall (top line) over all relations. The best normal method is the Stop+Unam+Stat75 seed selection strategy and the MultiLab+Limit75 integration strategy, the best extended method uses the same strategies for seed selection and information integration and CorefN for co-reference resolution

Class	Best normal		Best extended		
	P	R	P	R	top line
Musical Artist	0.671	0.006	0.7443	0.0112	0.0354
Politician	0.76	0.0316	0.7876	0.0498	0.1777
River	0.875	0.0234	0.889	0.0902	0.14
Business Operation	0.851	0.071	0.8611	0.1352	**0.232**
Educational Institution	**0.931**	**0.0795**	0.9283	0.0893	0.1343
Book	0.773	0.0326	0.8044	0.0461	0.105
Film	0.819	0.0258	0.8026	0.0411	0.1804

The **Mintz** baseline model we re-implemented has the lowest precision out of all models. This is partly because the amount of available training data for those models is much smaller than for other models. For candidate identification, only entities recognised by Stanford NERC are used and in addition the approach by Mintz et al. only uses 1% of available negative training data. For other models we also use our own NER, which does not assign a NE label to instances. As a result, the NE class feature for the relation extractor is missing for all those NEs only detected by our own NER, which makes it much more difficult to predict a label. In the original paper this is solved by using more training data and only training a classifier for relations which have at least 7000 training examples. Unfortunately we cannot compare our results to Mintz et al.'s approach directly as they use a Wikipedia as a corpus and use a different evaluation setup.

The **Comb** group of models have a much higher precision than the Mintz model. This difference can be explained by the difference in features, but mostly the fact that the Mintz model only uses 1% of available negative training data. The absolute number of correctly recognised property values in the text is about 5 times as high as the Mintz group of features which, again, is due to the fact that Stanford NERC fails to recognise some of the relevant entities in the text.

For our different seed selection methods, **Unam, Stop, Stat and StatRes**, we observe that removing some of the ambiguities helps to improve the precision of models, but always at the expense of recall. However, removing too many positive training instances also hurts precision. The highest overall precision is achieved using the Stop+Unam+Stat75 seed selection method.

Although strategically selecting seeds improves precision, the different **information integration** methods we tested have a much bigger impact on precision. Allowing multiple labels for predictions (MultiLab) amounts to a significant boost in precision, as well as restricting the maximum number of results per relation. Limit75 leads to a higher precision than LimitMax at a small decrease in recall.

Our different models based on the relaxed setting show a surprisingly high precision. They outperform all models in terms of recall, and even increase precision for most classes. The classes they do not increase precision for are "Educational Institution" and "Film", both of which already have a high precision for the normal setting. The NoSub model has the highest recall out of all models based on the relaxed setting, since it is the least restrictive one. However, it also has the lowest precision. The different co-reference resolution models overall achieve very similar precision and recall. There is a difference in performance between different classes though: the gazetteer-based method outperforms the Stanford coref model in terms of precision for the classes "Musical Artist", "Business Operation", "Educational Institution" and "Film", whereas the Stanford coref method outperforms the gazetteer-based method for "Politician", "River" and "Book". This suggests that in the context of Web information extraction for knowledge base population, simple co-reference resolution methods based on synonym gazetteers are equally as effective as supervised co-reference resolution models. The models which perform co-reference resolution have about the same recall as other models, but increase precision by up to 11% depending on the class. The reason those models perform so well is that individual predictions are combined. Even if predicting individual mentions is more challenging using co-reference resolution compared to just using sentences which contain mentions of entities explicitly, some relation mentions can be predicted with a high confidence. This redundancy gained from additional results helps to improve overall precision.

In general, the availability of test data poses a challenge, which is reflected by the top line. The top line is quite low, depending on the class between 0.035 and 0.23. Using a search based method to retrieve Web pages for training and testing is quite widely used, e.g. [36] also use it for gathering a corpus for distant supervision. To increase the top line, one strategy could be to just retrieve more pages per query, as Vlachos do. Another option would be to use more sophisticated method for building search queries, as for instance researched by West et al. [38]. As for different relations and classes, we can observe that there is a sizable difference in precision for them. Overall, we achieve the lowest precision for *Musical Artist* and the highest for *Educational Institution*.

When examining the training set we further observe that there seems to be a strong correlation between the number of training instances and the precision for that property. This is also an explanation as to why removing possibly ambiguous training instances only improves precision up to a certain point: the classifier is better at dealing with noisy training data than too little training data.

We also analyse the test data to try to identify patterns of errors. The two biggest groups of errors are entity boundary recognition and subject identification errors. An example for the first group is the following sentence:

> "<s>The Hunt for Red October</s> remains a masterpiece of military <o>fiction</o>."

Although "fiction" would be correct result in general, the correct property value for this specific sentence would be "military fiction". Our NER suggests both as possible candidates (since we employ both greedy and non-greedy matching), but the classifier should only classify the complete noun phrase as a value of *Book:genre*. There are several reasons for this: "military fiction" is more specific than "fiction", and since Freebase often contains the general category ("fiction") in addition to more fine-grained categories, we have more property values for abstract categories to use as seeds for training than for more specific categories. Second, our Web corpus also contains more mentions for broader categories than for more specific ones. Third, when annotating training data, we do not restrict positive candidates to whole noun phrases, as explained in Section 4.2. As a result, if none of the lexicalisations of the entity match the whole noun phrase, but there is a lexicalisation which matches part of the phrase, we use that for training and the classifier learns wrong entity boundaries. The second big group of errors is that occurrences are classified for the correct relation, but the wrong subject.

> "<s>Anna Karenina</s> is also mentioned in <o>R. L. Stine</o>'s Goosebumps series Don't Go To Sleep."

In that example, "R. L. Stine" is predicted to be a property value for *Book:author* for the entity "Anna Karenina". This happens because, at the moment, we do not take into consideration that two entities can be in *more than one* relation. Therefore, the classifier learns wrong, positive weights for certain contexts.

6. Discussion and future work

In this paper, we have documented and evaluated a distantly supervised class-based approach for relation extraction from the Web which strategically selects seeds for training, extracts relation mentions across sentence boundaries, and integrates relation mentions to predict relations for knowledge base population. Previous distantly supervised approaches have been tailored towards extraction from narrow domains, such as news and Wikipedia, and are therefore not fit for Web relation extraction: they fail to identify named entities correctly, they suffer from data sparsity, and they either do not try to resolve noise caused by ambiguity or do so at a significant increase of runtime. They further assume that every sentence may contain any entity in the knowledge base, which is very costly.

Our research has made a first step towards achieving those goals. We experiment with a simple NER, which we use in addition to a NERC trained for the news domain and find that it can especially improve on the number of extractions for non-standard named entity classes such as *MusicalArtist:track* and *MusicalArtist:album*. At the moment, our NER only recognises, but does not classify NEs. In future work, we aim to research distantly supervised named entity classification methods to assist relation extraction.

To overcome data sparsity and increase the number of extractions, we extract relation mentions across sentence boundaries and integrate them to predict relations. We find that extracting relation mentions across sentence boundaries not only increases recall by up to 25% depending on the model, but also increases precision by 8% on average. Moreover, we find that a gazetteer-based method for co-reference resolution achieves the same performance on our Web corpus as the Stanford CoreNLP co-reference resolution system. To populate knowledge bases, we test different information integration strategies, which differ in performance by 5%. We further show that simple, statistical methods to select seeds for training can help to improve performance of distantly supervised Web relation extractors, increasing precision by 3% on average. The performance of those methods is dependent on the type of relation it is applied to and on how many seeds there are available for training. Removing too many seeds tends to hurt performance rather than improve it.

One potential downside of using distant supervision for knowledge base population is that it either requires a very large corpus, such as the Web, or a big knowledge base for training. As such, distant supervision itself is an unsupervised domain-independent approach, but might not necessarily be useful for scenarios for which only a small corpus of documents or only a very small number of relation tuples is available in the knowledge base. For our experiments, we use a relatively large part of the knowledge base for training, i.e. 1000 seed entities for training per class, and 10 Web documents per entity and relation. In other experimental setups for distant supervision, only 30 seed entities, but 300 Web documents per entity and relation are used [36]. It is not just the quantity of documents retrieved that matters, but also the relevance to the information extraction task. Information retrieval for Web relation extraction, i.e. how to formulate queries to retrieve relevant documents for the relation extraction task is something that has already been researched, but not been exploited for distant supervision yet [38]. In future work, we plan to research how to jointly extract relations from text, lists and tables on Web pages in order to reduce the impact of data sparsity and increase precision for relation mention extraction. A detailed description of future work goals is also documented in Augenstein [2].

Acknowledgements

We thank the EKAW and SWJ 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

[1] E. Alfonseca, K. Filippova, J.-Y. Delort and G. Garrido, Pattern learning for relation extraction with a hierarchical topic model, in: *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers – Volume 2, ACL'12*, H. Li, C.-Y. Lin, M. Osborne, G.G. Lee and J.C. Park, eds, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 54–59.

[2] I. Augenstein, Joint information extraction from the Web using Linked Data, in: *International Semantic Web Conference (2)*, P. Mika, T. Tudorache, A. Bernstein, C. Welty, C.A. Knoblock, D. Vrandecic, P.T. Groth, N.F. Noy, K. Janowicz and C.A. Goble, eds, Lecture Notes in Computer Science, Vol. 8797, Springer, Heidelberg, Germany, 2014, pp. 505–512.

[3] I. Augenstein, Seed selection for distantly supervised Web-based relation extraction, in: *Proc. of the Third Workshop on Semantic Web and Information Extraction*, Dublin, Ireland, D. Maynard, M. van Erp and B. Davis, eds, 2014, Association for Computational Linguistics and Dublin City University, pp. 17–24.

[4] I. Augenstein, D. Maynard and F. Ciravegna, Relation extraction from the Web using distant supervision, in: *EKAW*, K. Janowicz, S. Schlobach, P. Lambrix and E. Hyvönen, eds, Lecture Notes in Computer Science, Vol. 8876, Springer, Heidelberg, Germany, 2014, pp. 26–41.

[5] I. Augenstein, S. Padó and S. Rudolph, LODifier: Generating Linked Data from unstructured text, in: *ESWC*, E. Simperl, P. Cimiano, A. Polleres, Ó. Corcho and V. Presutti, eds, Lecture Notes in Computer Science, Vol. 7295, Springer, Heidelberg, Germany, 2012, pp. 210–224.

[6] S. Bergsma and D. Lin, Bootstrapping path-based pronoun resolution, in: *Proc. of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association of Computational Linguistics*, Jeju Island, Korea, N. Calzolari, C. Cardie and P. Isabelle, eds, The Association for Computer Linguistics, 2006.

[7] K. Bollacker, C. Evans, P. Paritosh, T. Sturge and J. Taylor, Freebase: A collaboratively created graph database for structuring human knowledge, in: *Proc. of the 2008 ACM SIGMOD International Conference on Management of Data*, ACM, New York, NY, USA, 2008, pp. 1247–1250.

[8] R. Bunescu and R. Mooney, Learning to extract relations from the web using minimal supervision, in: *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, A. Zaenen and A. van den Bosch, eds, 2007, Association for Computational Linguistics, pp. 576–583.

[9] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr. and T.M. Mitchell, Toward an architecture for never-ending language learning, in: *Proc. of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, M. Fox and D. Poole, eds, AAAI Press, Palo Alto, California, USA, 2010.

[10] M. Craven, J. Kumlien et al., Constructing biological knowledge bases by extracting information from text sources, in: *Proc. of the International Conference on Intelligent Systems for Molecular Biology*, T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.-W. Mewes and R. Zimmer, eds, Vol. 1999, AAAI Press, Palo Alto, California, USA, 1999, pp. 77–86.

[11] L. Del Corro and R. Gemulla, ClausIE: Clause-based open information extraction, in: *Proc. of the 23rd International Conference on World Wide Web*, Rio de Janeiro, Brazil, D. Schwabe, V.A.F. Almeida, H. Glaser, R.A. Baeza-Yates and S.B. Moon, eds, ACM, 2013, pp. 355–366.

[12] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D.S. Weld and A. Yates, Web-scale information extraction in KnowItAll, in: *Proc. of the 13th International Conference on World Wide Web*, Rio de Janeiro, Brazil, S. Feldman, M. Uretsky, M. Najork and C. Wills, eds, ACM, 2004.

[13] A. Fader, S. Soderland and O. Etzioni, Identifying relations for open information extraction, in: *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu and S. Bethard, eds, Association for Computational Linguistics, Seattle, Washington, USA, 2011, pp. 1535–1545.

[14] C. Fellbaum (ed.), *Wordnet, an Electronic Lexical Database*, Language, Speech, and Communication, MIT Press, Cambridge, Massachusetts, USA, 1998.

[15] J.R. Finkel, T. Grenager and C.D. Manning, Incorporating non-local information into information extraction systems by Gibbs sampling, in: *Proc. of the 43nd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, K. Knight, H. Tou Ng and K. Oflazer, eds, 2005, Association for Computational Linguistics, pp. 363–370.

[16] A.L. Gentile, Z. Zhang, I. Augenstein and F. Ciravegna, Unsupervised wrapper induction using Linked Data, in: *Proc. of the 7th International Conference on Knowledge Capture*, V.R. Benjamins, M. d'Aquin and A. Gordon, eds, ACM, New York, NY, USA, 2013, pp. 41–48.

[17] D. Gerber and A.-C.N. Ngomo, Extracting multilingual natural-language patterns for RDF predicates, in: *Knowledge Engineering and Knowledge Management*, A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Aquin, A. Nikolov, N. Aussenac-Gilles and N. Hernandez, eds, Lecture Notes in Computer Science, Vol. 7603, Springer, Heidelberg, Germany, 2012, pp. 87–96.

[18] R. Hoffmann, C. Zhang, X. Ling, L.S. Zettlemoyer and D.S. Weld, Knowledge-based weak supervision for information extraction of overlapping relations, in: *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, Y. Matsumoto and R. Mihalcea, eds, The Association for Computer Linguistics, 2011, pp. 541–550.

[19] D.D. Lewis, Y. Yang, T.G. Rose and F. Li, RCV1: A new benchmark collection for text categorization research, *Journal of Machine Learning Research* **5** (2004), 361–397.

[20] Mausam, M. Schmitz, S. Soderland, R. Bart and O. Etzioni, Open language learning for information extraction, in: *Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, J. Tsujii, J. Henderson and M. Paşca, eds, Association for Computational Linguistics, 2012, pp. 523–534.

[21] B. Min, R. Grishman, L. Wan, C. Wang and D. Gondek, Distant supervision for relation extraction with an incomplete knowledge base, in: *Proc. of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, L. Vanderwende, H. Daumé III and K. Kirchhoff, eds, 2013, The Association for Computational Linguistics, pp. 777–782.

[22] M. Mintz, S. Bills, R. Snow and D. Jurafsky, Distant supervision for relation extraction without labeled data, in: *Proc. 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*, Suntec, Singapore, K.-Y. Su, J. Su, J. Wiebe and H. Li, eds, Association for Computational Linguistics, 2009, pp. 1003–1011.

[23] N. Nakashole, M. Theobald and G. Weikum, Scalable knowledge harvesting with high precision and high recall, in: *Proc. of the 4th ACM International Conference on Web Search and Data Mining*, I. King, W. Nejdl and H. Li, eds, ACM, New York, NY, USA, 2011, pp. 227–236.

[24] T.V.T. Nguyen and A. Moschitti, End-to-end relation extraction using distant supervision from external semantic repositories, in: *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, Y. Matsumoto and R. Mihalcea, eds, Association for Computational Linguistics, 2011, pp. 277–282.

[25] V. Presutti, S. Consoli, A.G. Nuzzolese, D.R. Recupero, A. Gangemi, I. Bannour and H. Zargayouna, Uncovering the

semantics of Wikipedia pagelinks, in: *EKAW*, K. Janowicz, S. Schlobach, P. Lambrix and E. Hyvönen, eds, Lecture Notes in Computer Science, Vol. 8876, Springer, Heidelberg, Germany, 2014, pp. 413–428.

[26] V. Presutti, F. Draicchio and A. Gangemi, Knowledge extraction based on discourse representation theory and linguistic frames, in: *EKAW* A. Ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. D'Aquin, A. Nikolov, N. Aussenac-Gilles and N. Hernandez, eds, Lecture Notes in Computer Science, Vol. 7603, Springer, Heidelberg, Germany, 2012, pp. 114–129.

[27] S. Riedel, L. Yao and A. McCallum, Modeling relations and their mentions without labeled text, in: *Proc. of the 2010 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (3)*, J.L. Balcázar, F. Bonchi, A. Gionis and M. Sebag, eds, Lecture Notes in Computer Science, Vol. 6323, Springer, Heidelberg, Germany, 2010, pp. 148–163.

[28] S. Riedel, L. Yao, A. McCallum and B.M. Marlin, Relation extraction with matrix factorization and universal schemas, in: *Proc. of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, L. Vanderwende, H. Daumé III and K. Kirchhoff, eds, Association for Computational Linguistics, 2013, pp. 74–84.

[29] R. Roller and M. Stevenson, Self-supervised relation extraction using UMLS, in: *Proc. of the 5th International Conference of the CLEF Initiative*, E. Kanoulas, M. Lupu, P.D. Clough, M. Sanderson, M.M. Hall, A. Hanbury and E.G. Toms, eds, Lecture Notes in Computer Science, Vol. 8685, Springer, Heidelberg, Germany, 2014, pp. 116–127.

[30] B. Roth, T. Barth, M. Wiegand and D. Klakow, A survey of noise reduction methods for distant supervision, in: *Proc. of the 2013 Workshop on Automated Knowledge Base Construction*, F. Suchanek, S. Riedel, S. Singh and P.P. Talukdar, eds, ACM, New York, NY, USA, 2013, pp. 73–78.

[31] B. Roth and D. Klakow, Combining generative and discriminative model scores for distant supervision, in: *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu and S. Bethard, eds, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 24–29.

[32] F.M. Suchanek, G. Kasneci and G. Weikum, YAGO: A large ontology from Wikipedia and WordNet, *Web Semantics: Science, Services and Agents on the World Wide Web* **6**(3) (2008), 203–217.

[33] M. Surdeanu, J. Tibshirani, R. Nallapati and C.D. Manning, Multi-instance multi-label learning for relation extraction, in: *Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, J. Tsujii, J. Henderson and M. Pasça, eds, Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 455–465.

[34] S. Takamatsu, I. Sato and H. Nakagawa, Reducing wrong labels in distant supervision for relation extraction, in: *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jeju Island, Korea, H. Li, C.-Y. Lin, M. Osborne, G.G. Lee and J.C. Park, eds, Association for Computational Linguistics, 2012, pp. 721–729.

[35] C. Unger, L. Bühmann, J. Lehmann, A.-C. Ngonga Ngomo, D. Gerber and P. Cimiano, Template-based question answering over RDF data, in: *Proc. of the 21st International Conference on World Wide Web*, A. Mille, F. Gandon, J. Misselis, M. Rabinovich and S. Staab, eds, ACM, New York, NY, USA, 2012, pp. 639–648.

[36] A. Vlachos and S. Clark, Application-driven relation extraction with limited distant supervision, in: *Proc. of the First AHA!-Workshop on Information Discovery in Text*, Dublin, Ireland, A. Akbik and L. Visengeriyeva, eds, 2014, Association for Computational Linguistics and Dublin City University, pp. 1–6.

[37] D. Vrandečić and M. Krötzsch, Wikidata: A free collaborative knowledgebase, *Communications of the ACM* **57**(10) (2014), 78–85.

[38] R. West, E. Gabrilovich, K. Murphy, S. Sun, R. Gupta and D. Lin, Knowledge base completion via search-based question answering, in: *Proc. of the 23rd International Conference on World Wide Web*, C.-W. Chung, A.Z. Broder, K. Shim and T. Suel, eds, ACM, New York, NY, USA, 2014, pp. 515–526.

[39] F. Wu and D.S. Weld, Autonomously semantifying Wikipedia in: *Proc. of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, M.J. Silva, A.O. Falcão, A.H.F. Laender, R. Baeza-Yates, B. Olstad and Ø.H. Olsen, eds, ACM, New York, NY, USA, 2007, pp. 41–50.

[40] F. Wu and D.S. Weld, Open information extraction using Wikipedia, in: *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, J. Hajič, S. Carberry, S. Clark and J. Nivre, eds, Association for Computational Linguistics, 2010, pp. 118–127.

[41] W. Xu, R. Hoffmann, l. Zhao and R. Grishman, Filling knowledge base gaps for distant supervision of relation extraction, in: *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, P. Fung and M. Poesio, eds, Association for Computational Linguistics, 2013, pp. 665–670.

[42] L. Yao, S. Riedel and A. McCallum, Collective cross-document relation extraction without labelled data, in: *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, H. Li and L. M'arquez, eds, Association for Computational Linguistics, 2010, pp. 1013–1023.

[43] A. Yates, M. Banko, M. Broadhead, M. Cafarella, O. Etzioni and S. Soderland, TextRunner: Open information extraction on the Web, in: *Proc. of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York, USA, B. Carpenter, A. Stent and J.D. Williams, eds, Association for Computational Linguistics, 2007, pp. 25–26.

[44] J. Zhu, Z. Nie, X. Liu, B. Zhang and J.-R. Wen, StatSnowball: A statistical approach to extracting entity relationships, in: *Proc. of the 18th International Conference on World Wide Web*, J. Quemada, G. León, Y. Maarek and W. Nejdl, eds, ACM, New York, NY, USA, 2009, pp. 101–110.