Extracting Relations between Non-Standard Entities using Distant Supervision and Imitation Learning

Isabelle Augenstein* **Andreas Vlachos#** **Diana Maynard***
* Department of Computer Science, University of Sheffield
i.augenstein@sheffield.ac.uk, d.maynard@sheffield.ac.uk
Computer Science Department, University College London
a.vlachos@cs.ucl.ac.uk

Abstract

Distantly supervised approaches have become popular in recent years as they allow training relation extractors without text-bound annotation, using instead known relations from a knowledge base and a large textual corpus from an appropriate domain. While state of the art distant supervision approaches use off-the-shelf named entity recognition and classification (*NERC*) systems to identify relation arguments, discrepancies in domain or genre between the data used for NERC training and the intended domain for the relation extractor can lead to low performance. This is particularly problematic for "non-standard" named entities such as album which would fall into the MISC category. We propose to ameliorate this issue by jointly training the named entity classifier and the relation extractor using imitation learning which reduces structured prediction learning to classification learning. We further experiment with Web features different features and compare against using two off-the-shelf supervised NERC systems, Stanford NER and FIGER, for named entity classification. Our experiments show that imitation learning improves average precision by 4 points over an one-stage classification model, while removing Web features results in a 6 points reduction. Compared to using FIGER and Stanford NER, average precision is 10 points and 19 points higher with our imitation learning approach.

1 Introduction

Factual answers to queries such as "What albums did The Beatles release?" are commonly stored in knowledge bases and can then be accessed by an information retrieval system, a commercial example for this being Google's knowledge vault (Xin et al., 2014). In order to keep knowledge bases up to date should new facts emerge, and to quickly adapt to new domains, there is a need for flexible and accurate information extraction (*IE*) approaches which do not require manual effort to be developed for new domains. A popular approach for creating IE methods to extract such relations is distant supervision (Craven and Kumlien, 1999; Mintz et al., 2009) which is a method for learning relation extractors using relations stored in a knowledge base combined with raw text to automatically generate training data.

An important first step in distant supervision is to identify named entities (*NEs*) and their types to determine if a pair of NEs is a suitable candidate for the relation. As an example, the album relation has a Musical Artist and an Album as arguments. Existing works use supervised named entity recognisers and classifiers (*NERC*) with either a small set of types such as the Stanford NER system (Manning et al., 2014), or fine-grained NE types (Ling and Weld, 2012; Liu et al., 2014). However, supervised NERCs typically focus on recognising persons, locations and organisations and perform poorly for other types of NEs, e.g. we find that Stanford NER only recognises 43% of all MISC NEs in our corpus. In addition, they do not always perform well if they are trained on a different type of text or for a different domain (Derczynski et al., 2015). This issue becomes more important as focus is shifting from using curated text collections such as Wikipedia to texts collected from the Web via search queries (Web-based distant supervision) which can provide better coverage (West et al., 2014).

In order to ameliorate this issue, we propose to recognise NEs with simple heuristics, then use the imitation learning algorithm DAGGER (Ross et al.,

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 747–757,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

2011) to learn the NEC component jointly with relation extraction (*RE*), without requiring explicitly labeled data for NERC. Instead, training signal is obtained by assessing the predictions of the relation extraction component. In this paper we make the following contributions:

1. We learn jointly training a named entity classifier and a relation extractor for Web-based distant supervision. Our method does not rely on hand-labeled training data and is applicable to any domain, which is shown in our evaluation on 18 different relations.
2. We compare different methods for this purpose: (1) we use imitation learning to train separate classifiers for NEC and RE jointly; (2) we aggregate NEC features and RE features and train a one-stage classification model; (3) we train a one-stage classification model with only RE features; (4) we classify NEs with two supervised off-the-shelf NEC systems (Stanford NER and FIGER) and use the NE types as features in RE to achieve a soft NE type constraint.
3. We explore the effects of using different NEC and RE features, including Web features such as links and lists on Web pages, and show that Web-based features improve average precision by 7 points. We further find that high-precision, but low-frequency features perform better than low-precision and high-frequency features.
4. Our experiments show that joint learning of NEC and RE with imitation learning outperforms one-stage classification models by 4 points in average precision, and models based on Stanford NER and FIGER by 19 and 10 points respectively.

2 Distant Supervision

Distantly supervised RE 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 (Mintz et al., 2009):

> 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.

3 Approach Overview

The input to the approach is a KB which contains entities and is partly populated with relations, the task is to complete the knowledge base. As an example, consider a KB about musical artists and their albums, which contains names of musical artists, and albums for some of them. The task is then to find albums for the remaining musical artists. Queries are automatically formulated containing C, s and o, e.g. "Musical Artist album 'The Beatles"' and we obtain Web pages using a search engine. For each sentence on the Web pages retrieved which contains s, all candidates for C are identified using NER heuristics (Section 4.2). Next, the distant supervision assumption is applied to all such sentences containing s (e.g. "Michael Jackson") and a candidate for that relation (e.g. "Music & Me"). If the candidate is an example of a relation according to the KB, it is used as a positive example, and if not, as a negative example. The examples are then used to train a model to recognise if the candidate is of the right type for the relation (NEC) and if it is of the correct relation (RE). The model is applied to the sentences of all the incomplete entries in the KB. Since different sentences could predict different answers to the query, all predictions are combined for the final answer.

4 Named Entity Recognition and Relation Extraction

The input to the learning task is a collection of training examples for a specific relation. The examples are sentences containing the subject of the relation and one further NE identified using simple heuristics. The examples are labeled as true (relation is contained in knowledge base) or as false (relation is not contained in the knowledge base).

We model the task in two binary classifica-

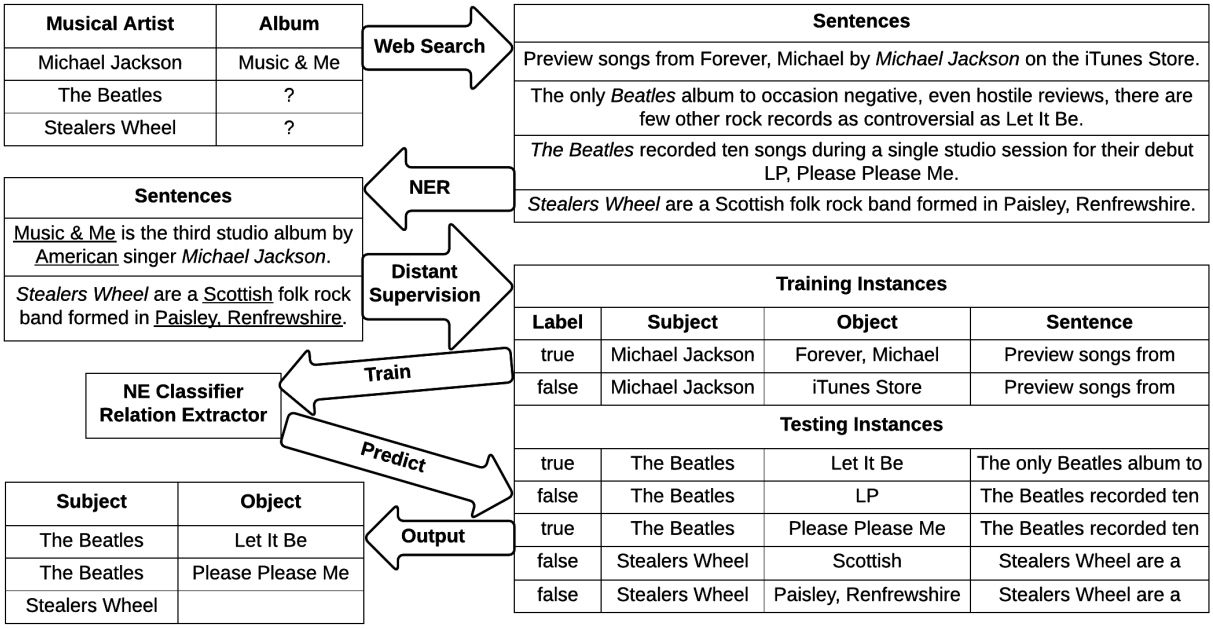

Figure 1: Overview of approach

tion stages: named entity classification (NEC) and relation extraction (RE). Existing approaches assume that named entity recognition and classification is done as part of the pre-processing. However, this is not possible domains for which NE classifiers are not readily available. To ameliorate this issue, existing approaches — e.g Mintz et al. (2009) — perform NEC to provide additional features for relation extraction. We use two such baselines with off-the-shelf NECs and add the NE labels to the relation features. The first baseline (**Stanf**) is with the Stanford NER 7-class (Time, Location, Organization, Person, Money, Percent and Date) model, the second (**FIGER**) is with the fine-grained FIGER (Ling and Weld, 2012).

An alternative approach is to simply add NEC features to relation extraction features, which we call **one-stage model (OS)** here. NEC features are typically morphological features extracted from the NE mention and features to model its context, whereas relation features typically model the path between the subject and object of the relation. While NEC features may be useful to determine if the NE has the correct type for the relation, such features are usually less sparse and also not directly related to the relation extraction task. Consider the following sentence, containing an example of the relation director:

"One of director <o>Steven Spielberg</o>'s greatest heroes was <o>Alfred Hitchcock</o>, the mastermind behind <s>Psycho</s>.

This sentence contains two relation candidates, "Steven Spielberg" and "Alfred Hitchcock", between which the decision for the final prediction has to be made. Both of the candidates are directors, but only one of them is the director of "Psycho". Because the context around "Steven Spielberg" is stronger (preceded by "director"), NEC features alone are more likely to indicate that as the correct candidate and also likely overpower relation features for the final prediction, as the latter tend to be sparser.

Ideally, we would like to train two models, one for NEC and one for RE, which would be applied in sequence. If the NEC stage concludes that the candidate is of the correct type for the relation, the RE stage determines whether the relation between the two entities is expressed. If the NEC stage concludes that the entity is not of the correct type, then the RE stage is not reached. However, distant supervision only provides positive labels for NEC, since if a sentence is labeled as false we do not know if it is due to the candidate not being of the correct type, or the relation not being true for the two entities. To overcome this, we learn models for the two stages, NEC and RE, jointly using the imitation learning algorithm DAGGER (Ross et al., 2011), as described in the next section.

4.1 Imitation Learning

Imitation learning[1] algorithms such as SEARN (Daumé III et al., 2009) and DAGGER (Ross et al., 2011) have been applied successfully to a variety of structured prediction tasks due to their flexibility in incorporating features and their ability to learn with non-decomposable loss functions. Sample applications include biomedical event extraction (Vlachos and Craven, 2011), dynamic feature selection (He et al., 2013), and machine translation (Grissom II et al., 2014).

Imitation learning algorithms for structured prediction decompose the prediction task into a sequence of actions; these actions are predicted by classifiers which are trained to take into account the effect of their predictions on the whole sequence by assessing their effect using a (possibly non-decomposable) loss function on the complete structure predicted. The dependencies between the actions are learnt via appropriate generation of training examples.

The ability to learn by assessing only the final prediction and not the intermediate steps is very useful in the face of missing labels, such as in the case of the labels for the NEC stage. Recall that the action sequence in our case consists of one NEC action and possibly one RE action, dependent on whether the NEC action is true, i.e. the entity is of the appropriate type for the relation. Following Vlachos and Clark (2014), for each training instance, we obtain supervision for the NEC stage by taking both options for this stage, true or false, obtaining the prediction from the RE stage in the former case and then comparing the outcomes against the label obtained from distant supervision. Thus the NEC stage is learned so that it enhances the performance of RE. In parallel, the RE stage is learned using only instances that actually reach this stage. The process is iterated so that the models learned adjust to each other. For more details on this we refer the reader to Vlachos and Clark (2014).

4.2 Relation Candidate Identification

To extract relations among NEs, the latter have to be detected first. Most distantly supervised approaches use supervised NER systems for this, which, especially for relations involving MISC NEs, achieve a low recall. High recall for NE identification is more important than high precision, since precision errors can be dealt with by the NEC stage. For a relation candidate identification stage with higher recall we instead rely on POS-based heuristics for detecting NEs[2] and HTML markup. We use the following POS heuristics:

- **Noun phrases:** Sequences of *N* tags
- **Capitalised phrases:** Those can be distinct from noun phrases, e.g. some album titles are capitalised verb phrases.

We further consider as relation candidates words which contain the following HTML markup:

- **Phrases from HTML markup:** All sequences of words marked as: `<ahref>` (links), `<li>` (list elements), `<h1>` or `<h2>` or `<h3>` (headers and subheaders, i.e. titles), `<strong>` or `<b>` (bold), `<em>` (emphasised), `<i>` (italics)

Different relation candidate identification strategies are then applied depending on the coarse NE types of objects of relation as defined in the KB (Table 2).

- **PER**: All capitalised noun phrases. We allow for a maximum of two characters to be surrounded by quotes to capture alternative first names, e.g. "Jerome David 'J. D.' Salinger".
- **LOC**: All capitalised noun phrases.
- **ORG**: All capitalised phrases and phrases from HTML markup. The latter is to capture ORG names which are not capitalised, e.g. the school "Woodrow Wilson School of Public and International Affairs" or the record label "Sympathy for the Record Industry".
- **MISC**: As for ORG, we use all capitalised phrases and phrases from HTML markup. MISC NEs are even more varied than ORG NEs and it is difficult to find the right balance between recognising most of them and generating unecessary candidates.

To assess how useful these strategies are, we randomly sample 30 instances of each Freebase class per coarse NE type of the object and manually examine all sentences which contain the subject of the relation. We used precision, i.e. how many of the relation candidates are appropriate

[1] Also referred to as search-based structured prediction or learning to search.

[2] The Stanford POS tagger uses Penn Treebank POS tags, see `http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html` for a list of tags

NE type	Model	R	P	F1
PER	heuristic	0.976	0.1287	0.227
PER	Stanford	0.774	0.1781	0.29
LOC	heuristic	0.963	0.1176	0.21
LOC	Stanford	0.889	0.1611	0.272727273
ORG	heuristic	0.95	0.0265	0.0516
ORG	Stanford	0.8	0.0505	0.095
MISC	heuristic	0.854	0.0496	0.0938
MISC	Stanford	0.427	0.053	0.0943

Table 1: Results for POS-based candidate identification strategies compared to Stanford NER

for the relation, and recall to compare the relation candidate identification strategies described above against the identification of candidates by Stanford NER (ignoring the NE label). As shown in Table 1, while supervised identification of NE labels achieves a higher precision for all NE types, the recall is higher for all NE types using POS-based heuristics. The simple heuristics are especially helpful for MISC NEs, for which recall is twice as high compared to Stanford NER and precision only marginally higher. If we were to use the NE label to enforce hard constraints, recall would be reduced even further: 88% of all PER entities are correctly identified as persons, compared to 58% for locations and 87% for organisations. MISC NE are identified as PER (45%), LOC (40%) or ORG (15%). Overall, precision is not as important for candidate identification as recall, since choosing correct entities among the candidates can be dealt with in a NEC stage.

4.3 NEC features

For the one-stage and imitation learning model, we use the following **Web features** based on HTML markup, both as local features if the entity mention contains the markup, and as global features if a mention somewhere else in the document with the same lexicalisation contains that markup: is link, is list element, is header or subheader, is bold, is emphasised, is italics, is title, is in title.

In addition, the following NEC features are extracted, based on Nadeau et al. (2007) and Hoffmann et al. (2011):
Word features (mentfeats):

- Object occurrence
- Sequence and BOW of occurrence
- Sequence and bag of POS of occurrence
- Number of words, characters and digits of object
- Ends with period, is roman number, contains apostrophe, hyphen, ampersand, possessive
- Digit and capitalisation pattern

Context features, as 1-grams (1cont) and 2-grams, 2 words to left and right of occurrence (2cont): BOW, sequence, bag of POS, POS sequence.

4.4 RE Features

The following features are used for RE, based on Hoffman et al (2011) and Mintz et al. (2009):

- 1cont and 2cont features
- Flag indicating which entity came first in sentence
- Sequence of POS tags and bag of words (BOW) between the subject and the object occurrence

Parsing features as full sequences (parse):

- Dependency path between subject and object, POS tags of words on that path
- Lemmas on dependency path, same with NNP and CD tokens substituted by POS tags

4.5 Supervised NEC Features for RE

For the baselines with off-the-shelf NECs, sentences are preprocessed with the two NEC systems Stanford NER and FIGER. NE labels are then used in addition to the RE features listed in Section 4.4. For the Stanf baseline, Stanford NER 7-class labels are added as RE features. Those are: Time, Location, Organization, Person, Money, Percent, Date. FIGER classifies NEs according to 112 types, most of which are subtypes of Person, Organization, Location, Product, Art, Event and Building. Some of the types are relation types we evaluate (see Table 2 for relation types): educational institution, city, director, actor and author. Since FIGER performs multi-label classification, it annotates some of the relation candidates with more than one NE label. In that case, we add all NE labels returned as features, though more experiments on how best to integrate multiple NE labels as features could be performed, as shown by Liu et al. (2014).

Musical Artist		Politician	
Relation type	**NE type**	**Relation type**	**NE type**
album	MISC	birthplace	LOC
record label	ORG	educational institution	ORG
track	MISC	spouse	PER
Business		**Educational Institution**	
Relation type	**NE type**	**Relation type**	**NE type**
employees	PER	mascot	MISC
founders	PER	city	LOC
Film		**Book**	
Relation type	**NE type**	**Relation type**	**NE type**
director	PER	author	PER
producer	PER	characters	MISC
actor	PER		
character	MISC		
River			
Relation type	**NE type**		
origin	LOC		
mouth	LOC		

Table 2: Relation types and corresponding coarse NE types

5 Evaluation

5.1 Corpus

To create a corpus[3] for Web RE, seven Freebase classes and two to four of their relations are selected (Table 2). The selected classes are subclasses of PER (Musical Artist, Politician), LOC (River), ORG (Business (Operation)), Education(al Institution)) or MISC (Film, Book). To avoid noisy training data, we only use entities which have values for all of those properties, which resulted in 1800 to 2200 entities per class. For each entity, 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'. In total, the corpus consists of around one million pages drawn from 76,000 different websites. Text content is extracted from HTML pages using the Jsoup API[4] and processed with Stanford CoreNLP[5].

5.2 Models and Metrics

We evaluate the following models: imitation learning (**IL**) as described in Section 4.1, a one-stage model (**OS**), a one-stage model with relation features only (**RelOnly**), and using Stanford (**Stanf**) and FIGER (**FIGER**) NE labels as features (Section 4). For all models we use linear classifiers learned with passive-aggressive updates (Crammer et al., 2006). For imitation learning, we use the learning algorithm DAGGER (Ross et al., 2011), which requires two parameters: the learning rate, i.e. how quickly the learning algorithm moves away from the optimal policy, and the number of iterations. We found empirically that the best learning rate for our prediction task is 0.25 and that the best number of iterations is 12.

[3]The resources for experiments documented in this paper are available online via `http://tinyurl.com/o8ykn4y`

[4]http://jsoup.org

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

Model	R-top	P-top	F1-top	R-all	P-all	P-avg
RelOnly	0.1943	0.404	0.255	0.223	0.309	0.373
Stanf	0.233	0.436	0.304	0.268	0.329	0.398
FIGER	0.228	0.497	0.298	0.251	0.413	0.483
OS	**0.269**	0.58	**0.356**	**0.288**	0.486	0.552
IL	0.246	**0.600**	0.329	0.271	**0.521**	**0.588**

Table 3: Results for best model for each relation, macro average over all relations.

The output of the models is a score for each relation example and stage, i.e. for the one-stage model, the output is one score and for the imitation learning model, there is a score each for the NEC stage and the RE stage. The default for deciding whether the relation label should be true or false depends on stage thresholds, which are 0 by default. Instead of using the default thresholds, we automatically pick thresholds for all models on 1/3 of the training set, which we set aside as a development set, then retrain on the whole training set and predict relations based on the learnt thresholds.

We use the metrics first best precision (P-top), first best recall (R-top), first best F1 (F1-top), all precision (P-all), all recall (P-all), and all average precision (P-avg). For top, only the top-ranked answer is considered, whereas for all all answers are returned until either the correct one is found or they are exhausted. Finally, in the all mode we evaluated precision at all recall points by varying the thresholds used in the respective classifiers and we report average precision (P-avg) (Manning et al., 2008). This evaluation measure provides an assessment of how well a system trades precision for recall. The number of all results for computing recall is the number of all relation tuples in the KB.

Relation	RelOnly		Stanf		FIGER		OS		IL	
	F1-top	P-avg	F1-top	P-avg	F1-top	P-avg	F1-top	P-avg	F1-top	P-avg
Musical Artist : album	0.071	0.175	0.079	0.109	0.116	0.203	0.158	0.409	0.115	**0.569**
Musical Artist : record label	0.090	0.182	0.100	0.345	0.179	0.636	0.404	0.758	0.376	**0.926**
Musical Artist : track	0.093	0.109	0.053	0.175	0.104	0.400	0.118	**0.471**	0.114	0.367
Politician : birthplace	0.410	0.594	0.514	0.541	0.496	0.609	0.585	**0.709**	0.516	0.548
Politician : educational institution	0.321	0.387	0.330	0.426	0.366	0.560	0.419	0.719	0.381	**0.831**
Politician : spouse	0.148	0.197	0.152	0.197	0.082	0.309	0.218	**0.319**	0.150	0.181
Business : employees	0.059	0.090	0.097	0.153	0.082	0.325	0.149	0.291	0.133	**0.493**
Business : founders	0.341	0.256	0.462	0.332	0.404	0.542	0.448	0.663	0.429	**0.693**
Education : mascot	0.148	0.362	0.195	0.483	0.226	0.500	0.225	0.506	0.206	**0.585**
Education : city	0.630	0.705	0.711	0.740	0.701	0.770	0.724	0.847	0.690	**0.872**
Film : director	0.383	0.548	0.445	0.603	0.358	0.554	0.439	0.601	0.387	**0.612**
Film : producer	0.149	0.384	0.209	0.395	0.164	0.387	0.198	0.355	0.227	**0.400**
Film : actor	0.246	0.576	0.308	0.633	0.351	0.609	0.342	0.684	0.312	**0.732**
Film : character	0.093	0.123	0.093	0.117	0.180	0.195	0.194	0.298	0.173	**0.319**
Book : author	0.629	0.852	0.703	0.852	0.781	0.878	0.773	0.867	0.781	**0.885**
Book : characters	0.224	0.127	0.193	0.127	0.262	0.328	0.268	0.315	0.231	**0.355**
River : origin	0.175	0.328	0.232	0.493	0.160	0.351	0.256	0.406	0.228	**0.550**
River : mouth	0.336	0.594	0.423	0.564	0.347	0.529	0.488	**0.709**	0.479	0.668

Table 4: Results for best model for each relation, highest P-avg in bold

6 Results and Discussion

6.1 Comparison of Models

Overall results in Table 3 show that both of our models (**IL** and **OS**) outperform the baselines with off-the-shelf supervised NEC (**Stanf**, **FIGER**) for all metrics. Detailed results for different relations (Table 4) show that IL outperforms both OS and Base in terms of average precision. FIGER results fall in between Stanf and OS results. For some relations, there is a dramatic improvement by using fine-grained FIGER NE features over coarse-grained Stanford NE features; occasionally FIGER even outperforms OS, as for the relation author. This is because FIGER has a corresponding NE type (see Section 4.5).

For most relations, including those whose objects are of type MISC, IL shows a significant improvement in terms of F1 or average precision over OS (Table 5). This confirms our hypothesis that separating the NEC and relation extraction stages using imitation learning can achieve a higher precision and recall for non-standard relations than preprocessing sentences with a supervised NEC model. Furthermore, we show that it can also be useful for most standard relations. The main relations for which Stanf, FIGER or OS can have a benefit over IL are those for which entities are easy to classify, specifically LOC NEs, but also PER NEs. This is because, if NEs are easy to classify, a separate NEC is less likely to be useful.

6.2 Imitation Learning vs One-Stage

To give more insight into why **IL** is overall more successful than **OS**, common errors made by OS are shown here, along with an explanation of how those errors are prevented by using IL. One example of IL predicting correctly but OS incorrectly is from the following sentence, expressing the director relation:

"In 2010 he appeared in a leading role in <o>Alicia Duffy</o>'s <s>All Good Children</s>.

In that example, the NEC features extracted for <o>Alicia Duffy</o> are not very strong indicators, since neither the object string itself nor the surrounding context give any direct indication for the director relation. The RE features, which are based on the dependency path, are a stronger indicator. Since in the OS model all features are combined, the NEC features overpower the RE features. The IL model, on the other hand, learns a permissive NEC as a first stage, which filters NEs with respect to if they are generally appropriate for the relation or not, and then leaves the RE to the second stage.

Another example is a sentence for which OS incorrectly predicts the relation author, whereas IL correctly predicts "false":

"<o>Laura</o> and Mary went to school for the first time in Pepin rather than Walnut Grove, which is not included in <s>Little House in the Big Woods</s>."

Relation	NEC Features	Rel Features
Musical Artist : album	2cont + 1cont + mentfeats + web	parse
Musical Artist : record label	2cont + 1cont + mentfeats + web	parse + 2contword
Musical Artist : track	parse + 2cont + 1cont + mentfeats	parse
Politician : birthplace	2cont + 1cont + mentfeats + web	parse
Politician : educational institution	parse + cont + ment	parse
Politician : spouse	parse + 2cont + 1cont + web	parse
Business : employees	2cont + 1cont + mentfeats + web	parse + 2contword
Business : founders	parse + cont + ment	parse
Education : mascot	parse + 2contwordpos	parse + cont
Education : city	parse + cont + ment	parse + 2contwordpos
Film : director	2cont + 1cont + mentfeats + web	parse + 2contword
Film : producer	parse + cont	parse + 2contwordpos
Film : actor	parse + 2cont + web	parse + 2contwordpos
Film : character	parse + cont + ment	parse + 2contword
Book : author	2cont + 1cont + mentfeats + web	parse
Book : characters	parse + cont + ment	parse
River : origin	2cont + 1cont + mentfeats + web	parse + 2contword
River : mouth	2cont + 1cont + mentfeats + web	parse

Table 5: Best feature combination for IL

NEC Features	Rel Features	P-top	R-top	F1-top	P-all	R-all	P-avg
2cont	parse	0.215	0.399	0.28	0.253	0.316	0.381
2cont + 1cont + mentfeats	parse	0.239	0.456	0.313	0.275	0.378	0.441
2cont + 1cont + mentfeats + web	parse	0.248	0.51	0.322	0.276	0.431	0.502
2cont + web	parse	0.204	0.375	0.264	0.244	0.289	0.35
2cont	parse + 2contwordpos	0.236	0.43	0.305	0.275	0.338	0.402
2cont + 1cont + mentfeats	parse + 2contwordpos	0.239	0.456	0.313	0.275	0.378	0.441
2cont + 1cont + mentfeats + web	parse + 2contwordpos	0.248	0.518	0.324	0.275	0.421	0.486
2cont + web	parse + 2contwordpos	0.24	0.402	0.3	0.279	0.305	0.371
2cont	parse + 2contword	0.215	0.394	0.278	0.258	0.309	0.372
2cont + 1cont + mentfeats	parse + 2contword	0.231	0.453	0.295	0.266	0.352	0.43
2cont + 1cont + mentfeats + web	parse + 2contword	0.25	0.54	0.325	0.284	0.433	0.505
2cont + web	parse + 2contword	0.223	0.395	0.285	0.263	0.305	0.373

Table 6: Imitation learning results for different NE and relation features, macro average over all relations.

For this example, OS relation features have small positive weights, which then overall lead to a positive prediction. For IL, the first stage predicts "false", since the one-token string <o>Laura</o> is not a likely candidate for author.

6.3 Comparison of Features

All different feature groups have an overall positive effect on the results (see Table 6). While low precision, high frequency features improve recall (1cont), they do not always improve precision. Both **OS** and **IL** benefit from high precision, low frequency features, e.g. for author and mouth, the best results are achieved with only sparse parsing features for RE.

Web features improve performance for 10 out of 18 relations. For n-ary relations the is list element feature is very useful because Web pages about musical artist, films or books often contain lists with their attributes, e.g. a Web page about a musical artist typically contains a list with their albums. For relations with persons as objects, is link and is bold is useful because Web pages often highlight persons or provide links to Web pages with more information about them. As an example, for the author relation, the strongest positive Web feature is is in title and the strongest negative feature is is list element. This makes sense since a book is frequently mentioned with its author and one of the most important attributes of a book, whereas lists on Web pages about books mention less important attributes, such as the characters.

6.4 Overall Comparison

Overall, we showed that using an off-the-shelf NEC as a pre-processing step for distant super-

vision as done by existing works often causes errors which can be prevented by instead separating NEC and RE with imitation learning. We also showed that using Web features increases precision for NEC. Finally, it is worth noting that the recall for some of the relations is quite low because they only infrequently occur in text, especially in the same sentence as the subject of the relation. These issues can be overcome by performing coreference resolution (Augenstein et al., 2014; Koch et al., 2014), by retrieving more Web pages or improving the information retrieval component of the approach (West et al., 2014) and by combining extractors operating on sentences with other extractors for semi-structured content on Web pages (Carlson et al., 2010).

7 Related Work

One of the first papers to introduce distant supervision was Mintz et al. (2009), which aims at extracting relations between entities in Wikipedia for the most frequent relations in Freebase. Most distant supervision research focuses on addressing the disadvantages of heuristic labelling, namely reducing false positive training data (Hoffmann et al., 2011; Surdeanu et al., 2012; Riedel et al., 2010; Riedel et al., 2013; Yao et al., 2010; Alfonseca et al., 2012; Roth and Klakow, 2013; Takamatsu et al., 2012; Xu et al., 2013) and dealing with false negatives due to missing entries in the knowledge base (Min et al., 2013), as well as combining distant supervision with active learning (Angeli et al., 2014)

Distant supervision has been researched for different domains, including newswire (Riedel et al., 2010; Riedel et al., 2013), Wikipedia (Mintz et al., 2009; Nguyen and Moschitti, 2011), the biomedical domain (Craven and Kumlien, 1999; Roller and Stevenson, 2014), the architecture domain (Vlachos and Clark, 2014) and the Web (Xin et al., 2014; Augenstein et al., 2014; Augenstein et al., 2015).

To date, there is very little research on improving NERC for distant supervision to extract relations between non-standard entities such as musical artists and albums. Some research has been done on improving distant supervision by using fine-grained named entity classifiers (Ling and Weld, 2012; Liu et al., 2014) and on using named entity linking for distant supervision (Koch et al., 2014). Liu et al. (2014) train a supervised fine-grained NERC on Wikipedia and show that using those types as entity contraints improves precision and recall for a distantly supervised RE on newswire. However, they assume that labeled training data is available, making it unsuitable for applying distant supervision to domains with relations involving non-standard entity types.

Vlachos and Clark (2014) also proposed a distantly supervised approach for joint learning of NEC and RE with imitation learning for the architecture domain. However, they only used two relations in their experiments which involved rather standard entity types and they did not compare against using off-the shelf NEC systems.

8 Conclusion and Future Work

In this paper, we proposed a method for extracting non-standard relations with distant supervision that learns a NEC jointly with relation extraction using imitation learning. Our proposed imitation learning approach outperforms models with supervised NEC for relations involving non-standard entities as well as relations involving persons, locations and organisations. We achieve an increase of 4 points in average precision over a simple one-stage classification model, and an increase in 10 points and 19 points over baselines with FIGER and Stanford NE labels. We further demonstrate that using specialised Web features, such as appearances of entities in lists and links to other Web pages, improves average precision by 7 points, which other Web search-based relation extraction approaches could also benefit from (Xin et al., 2014; Augenstein et al., 2014).

In future work, the proposed approach could be combined with other approaches to solve typical issues arising in the context of distant supervision, such as dealing with overlapping relations (Hoffmann et al., 2011), improving heuristic labelling of sentences (Takamatsu et al., 2012) or dealing with incomplete knowledge bases (Min et al., 2013).

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 ACL*, volume 2, pages 54–59, Jeju, South Korea. ACL.

Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D. Manning. 2014. Combining distant and par-

tial supervision for relation extraction. In *Proceedings of EMNLP*, pages 1556–1567, Doha, Qatar. ACL.

Isabelle Augenstein, Diana Maynard, and Fabio Ciravegna. 2014. Relation Extraction from the Web using Distant Supervision. In *Proceedings of EKAW*, pages 26–41. Springer.

Isabelle Augenstein, Diana Maynard, and Fabio Ciravegna. 2015. Distantly Supervised Web Relation Extraction for Knowledge Base Population. *Semantic Web Journal*. to appear.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, and Tom M. Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. In *Proceedings of AAAI*, Palo Alto, California, USA. AAAI Press.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

M. Craven and J. Kumlien. 1999. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. In *Proceedings of ISMB*, pages 77–86, Palo Alto, California, USA. AAAI Press.

Hal Daumé III, John Langford, and Daniel Marcu. 2009. Search-based Structured Prediction. *Machine Learning*, 75:297–325.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, and Kalina Bontcheva. 2015. Analysis of Named Entity Recognition and Linking for Tweets. *Information Processing and Management*, 51:32–49.

Alvin Grissom II, Jordan Boyd-Graber, He He, John Morgan, and Hal Daumé III. 2014. Don't Until the Final Verb Wait: Reinforcement Learning for Simultaneous Machine Translation. In *Proceedings of EMNLP*, Doha, Qatar. ACL.

He He, Hal Daumé III, and Jason Eisner. 2013. Dynamic Feature Selection for Dependency Parsing. In *Proceedings of EMNLP*, pages 1455–1464, Seattle, Washington, USA. ACL.

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 ACL*, pages 541–550, Portland, Oregon, USA. ACL.

Mitchell Koch, John Gilmer, Stephen Soderland, and Daniel S. Weld. 2014. Type-aware distantly supervised relation extraction with linked arguments. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of EMNLP*, pages 1891–1901, Doha, Qatar, October. Association for Computational Linguistics.

Xiao Ling and Daniel S. Weld. 2012. Fine-Grained Entity Recognition. In *Proceedings of AAAI*, pages 94–100. AAAI Press.

Yang Liu, Kang Liu, Liheng Xu, and Jun Zhao. 2014. Exploring Fine-grained Entity Type Constraints for Distantly Supervised Relation Extraction. In *Proceedigs of COLING*, pages 2107–2116, Dublin, Ireland. ACL.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of ACL: System Demonstrations*, pages 55–60, Baltimore, Maryland. ACL.

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, Atlanta, Georgia, USA. ACL.

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, Suntec, Singapore. ACL.

David Nadeau and Satoshi Sekine. 2007. A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes*, 30(1):3–26.

Truc-Vien T. Nguyen and Alessandro Moschitti. 2011. End-to-End Relation Extraction Using Distant Supervision from External Semantic Repositories. In *Proceedings of ACL-HLT*, pages 277–282, Portland, Oregon, USA. ACL.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling Relations and Their Mentions without Labeled Text. In *Proceedings of ECML-PKDD (3)*, volume 6323, pages 148–163. Springer.

Sebastian Riedel, Limin Yao, Benjamin M. Marlin, and Andrew McCallum. 2013. Relation Extraction with Matrix Factorization and Universal Schemas. In *Proceedings of HLT-NAACL*, pages 74–84, Atlanta, Georgia, USA. ACL.

Roland Roller and Mark Stevenson. 2014. Self-supervised Relation Extraction Using UMLS. In Evangelos Kanoulas, Mihai Lupu, Paul D. Clough, Mark Sanderson, Mark M. Hall, Allan Hanbury, and Elaine G. Toms, editors, *Proceedings of CLEF*, volume 8685 of *Lecture Notes in Computer Science*, pages 116–127. Springer.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. 2011. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In

Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudk, editors, *Proceedings of AISTATS*, volume 15, pages 627–635. JMLR.

Benjamin Roth and Dietrich Klakow. 2013. Combining Generative and Discriminative Model Scores for Distant Supervision. In *Proceedings of EMNLP*, pages 24–29, Seattle, Washington, USA. ACL.

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, Jeju Island, Korea. ACL.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing Wrong Labels in Distant Supervision for Relation Extraction. In *Proceedings of ACL*, pages 721–729, Jeju Island, Korea. ACL.

Andreas Vlachos and Stephen Clark. 2014. Application-Driven Relation Extraction with Limited Distant Supervision. In *Proceedings of Aha!*, pages 1–6, Dublin, Ireland. ACL.

Andreas Vlachos and Mark Craven. 2011. Search-based structured prediction applied to biomedical event extraction. In *Proceedings of CoNLL*, pages 49–57, Portland, Oregon, USA. ACL.

Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. 2014. Knowledge Base Completion via Search-Based Question Answering. In *Proceedings of WWW*, pages 515–526, New York, NY, USA. ACM.

Luna Dong Xin, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion. In *Proceedings of KDD*, pages 601–610, New York, NY, USA. ACM.

Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. 2013. Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction. In *Proceedings of ACL*, pages 665–670, Sofia, Bulgaria. ACL.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2010. Collective Cross-document Relation Extraction Without Labelled Data. In *Proceedings of EMNLP*, pages 1013–1023, Cambridge, MA, USA. ACL.