Type-Aware Distantly Supervised Relation Extraction with Linked Arguments

Mitchell Koch **John Gilmer** **Stephen Soderland** **Daniel S. Weld**
Department of Computer Science & Engineering
University of Washington
Seattle, WA 98195, USA
{mkoch,jgilme1,soderlan,weld}@cs.washington.edu

Abstract

Distant supervision has become the leading method for training large-scale relation extractors, with nearly universal adoption in recent TAC knowledge-base population competitions. However, there are still many questions about the best way to learn such extractors. In this paper we investigate four orthogonal improvements: integrating named entity linking (NEL) and coreference resolution into argument identification for training and extraction, enforcing type constraints of linked arguments, and partitioning the model by relation type signature.

We evaluate sentential extraction performance on two datasets: the popular set of NY Times articles partially annotated by Hoffmann et al. (2011) and a new dataset, called GoReCo, that is comprehensively annotated for 48 common relations. We find that using NEL for argument identification boosts performance over the traditional approach (named entity recognition with string match), and there is further improvement from using argument types. Our best system boosts precision by 44% and recall by 70%.

1 Introduction

Relation extractors are commonly trained by distant supervision (also known as knowledge-based weak supervision (Hoffmann et al., 2011)), an autonomous technique that creates a labeled training set by heuristically matching the contents of a knowledge base (KB) to *mentions* (substrings) in a textual corpus. For example, if a KB contained the ground tuple BornIn(Albert Einstein, Ulm) then a distant supervision system might label the sentence "While [Einstein]$_1$ was born in [Ulm]$_2$, he moved to Munich at an early age." as a positive training instance of the BornIn relation. Although distant supervision is a simple idea and often creates data with false positives, it has become ubiquitous; for example, all top-performing systems in recent TAC-KBP slot filling competitions used the method.

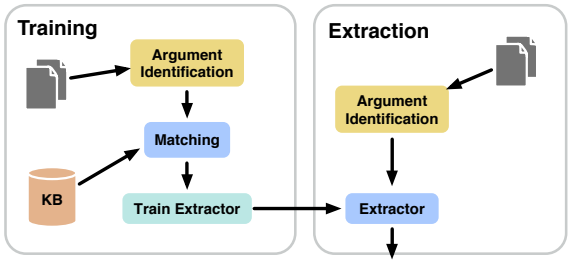


Figure 1: Distantly supervised extraction pipeline.

Surprisingly, however, many aspects of distant supervision are poorly studied. In response we perform an extensive search of ways to improve distant supervision and the extraction process, including using named entity linking (NEL) and coreference to identify arguments for distant supervision and extraction, as well as using type constraints and partitioning the trained model by relation type signatures.

The first step in the distant supervision process is *argument identification* (Figure 1) — finding textual mentions referring to entities that might be in some relation. Next comes *matching*, where KB facts, e.g. tuples such as $R(e_1, e_2)$, are associated with sentences mentioning entities e_1 and e_2 in the assumption that many of these sentences describe the relation R. Most previous systems perform these steps by first using named entity recognition (NER) to identify possible arguments and then using a simple string match, but this crude

approach misses many possible instances. Since the separately-studied task of named entity linking (NEL) is precisely what is needed to perform distant supervision, it is interesting to see if today's optimized linkers lead to improved performance when used to train extractors.

Coreference, the task of clustering mentions that describe the same entity, may also be useful for increasing the number of candidate arguments. Consider the following variant of our previous example: "While $[\text{he}]_1$ was born in $[\text{Ulm}]_2$, $[\text{Einstein}]_3$ moved to Munich at an early age." Since mentions 1 and 3 corefer, one could consider using either the pair $\langle 1, 2\rangle$ or $\langle 3, 2\rangle$ (or both) for training. Intuitively, it seems that $\langle 1, 2\rangle$ is more representative of BornIn and might generalize better, so we consider the use of coreference at both training and extraction time.

Semantic relations often have selectional preferences (also known as type signatures); for example, BornIn holds between people and locations. Therefore, it seems promising to include entity types, whether coarse or fine grained in the distantly supervised relation extraction process. We consider two ways of adding this information. By using NEL to get linked entities, we can impose type constraints on the relation extraction system to only allow relations over appropriately typed mentions. We also investigate using coarse types from NER to learn separate models for different relation type signatures in order to make the models more effective.

In summary, this paper represents the following contributions:

- We explore several dimensions for improving distantly supervised relation extraction, including better argument identification during training and extraction using both NEL and coreference, partitioning the model by relation type signatures, and enforcing type constraints of linked arguments as a post-processing step. While some of these ideas may seem straightforward, to our knowledge they have not been systematically studied. And, as we show, they lead to dramatic improvements.

- Since previous datasets are incapable of measuring an extractor's true recall, we introduce GoReCo, a new exhaustively-labeled dataset with gold annotations for sentential instances of 48 relations across 128 newswire documents from the ACE 2004 corpus (Doddington et al., 2004).

- We demonstrate that NEL argument identification boosts both precision and recall, and using type constraints with linked arguments further boosts precision, yielding a 43% increase in precision and 27% boost to recall. Using coreference during training argument identification gives an additional 7% improvement to precision and further boosts recall by 9%. Partitioning the model by relation type signature offers further benefits, so our best system yields a total boost of 44% to precision and 70% to recall.

2 Distantly Supervised Extraction

At a sentence-level, the goal for relation extraction is to determine for each sentence, what facts are expressed. We describe these as *relation annotations* of the form $s{\rightarrow}R(m_1, m_2)$, where s is a sentence, $R \in \mathcal{R}$ is a relation name, $\mathcal{R}$ is our finite set of target relations, and m_1 and m_2 are *grounded entity mentions* of the form (s, t_1, t_2, e), where t_1 and t_2 delimit a text span in the sentence, and e is a grounded entity.

2.1 Training

During training, the contents of the KB are heuristically matched to the training corpus according to the *distant supervision hypothesis*: if a relation holds between two entities, any sentence containing those two entities is likely to express that relation.

The training KB Δ contains *fact tuples* of form $R(e_1, e_2)$, where $R \in \mathcal{R}$ is a relation name, $\mathcal{R}$ is our finite set of target relations, and e_1 and e_2 are ground entities. The training text corpus Σ contains documents, which contain sentences. *Argument identification* is performed over the text corpus to get grounded mentions m. Then during sentential instance generation, *sentential instances* of the form (s, m_1, m_2) are generated representing a sentence with two grounded mentions. At this point, these sentential instances can be matched to the seed KB, yielding *candidate relation annotations* of the form $s{\rightarrow}R(m_1, m_2)$ by finding all relations that hold over the entities in a sentential instance. These candidate relation annotations are all positive instances to use for training. Negative instance generation is also performed, generating

negative examples of the form $s \rightarrow NA(m_1, m_2)$ indicating that no relation holds between m_1 and m_2. There are several heuristics for generating negative instances, and the number of negative examples and how they are treated can greatly affect performance (Min et al., 2013).

Because the distant supervision hypothesis often does not hold, this training data is noisy. That a fact is in the KB does not imply that the sentence in question is expressing the relation. There has been much work in combating noise in distant supervision training data, but one of the most successful ideas is to train a multi-instance classifier which assumes at-least-one relation holds for positive bags. We use Hoffmann et al. (2011)'s MULTIR system, which uses a probabilistic graphical model to jointly reason at the corpus-level and sentence-level, handles overlapping relations in the KB so that multiple relations can hold over an entity pair, and scales to large datasets.

2.2 Extraction

The trained relation extractor can assign a most likely relation and a confidence score to a sentential instance (s, m_1, m_2). To get these sentential instances, argument identification and sentential instance generation are applied to new documents. Then the relation extractor potentially yields a relation annotation of the form $s \rightarrow R(m_1, m_2)$, or potentially no relation. At extraction time a mention m might have a NIL link if a corresponding ground entity was not found during argument identification (meaning the entity is not in the KB). The relation annotations have associated confidence scores, so a threshold can be chosen to only use high-confidence relation annotations.

3 Argument Identification

An important piece of relation extraction is determining what can be an argument, and how to form a semantic representation of it. We define an argument identification function $\textit{ArgIdent}_{\Delta}(D)$, which takes a document D, finds potential arguments, and links them to entities in Δ if possible, yielding $\mathbf{m}$, a set of grounded mentions in D. Previous relation extraction systems have based this on NER. We evaluate NER-based argument identification against argument identification based on NEL, as well as NEL with coreference.

3.1 Named Entity Recognition

Named entity recognition (NER) tags spans of tokens with basic types such as PERSON, ORGANIZATION, LOCATION, and MISC. This is a high accuracy tagging task often performed using a sequence classifier (Finkel et al., 2005). Relation extraction systems can base their argument identification on NER, by using NER to identify text spans indicating entities and then find corresponding entities in the KB through exact string match (Riedel et al., 2010). Some downsides of using NER with exact string match for relation extraction is that it does not allow for overlapping mentions, it can only capture arguments with full names, and it can only capture arguments with types of the NER system, e.g., "politician" might not be captured.

3.2 Named Entity Linking

Named entity linking (NEL) is the task of grounding textual mentions to entities in a KB, such as Wikipedia. Thus "named entity" here, has a somewhat broader definition than in NER — these are any entities in the KB, not just those expressed with proper names. Hachey et al. (2013) define three stages that NEL systems take to perform this task: extraction (mention detection), search (generating candidate KB entities for a mention), and disambiguation (selecting the best entity for a mention). There has been much work on the task of NEL in recent years (Milne and Witten, 2008; Kulkarni et al., 2009; Ratinov et al., 2011; Cheng and Roth, 2013).

Our definition of a function $\textit{ArgIdent}(D)$ is completely served by an NEL system. It can find any entity in the KB, and those entities are grounded. Additionally, NEL can have overlapping mentions as well as support for abbreviated mentions like "Obama", or acronyms like "US". NEL does not seek to capture anaphoric mentions, however.

3.3 Coreference Resolution

Coreference resolution is the task of clustering mentions of entities together, typically within a single document. Using coreference, we can find even more mentions than NEL, since it can find pronouns and anaphoric mentions. We seek to use coreference to add additional arguments to those found by NEL, and we refer to this combined argument identification method as *NEL+Coref*. Tak-

ing in arguments from NEL argument identification and coreference clusters, we ground the clusters by picking the most common grounded entity from NEL mentions that occur in a coreference cluster. A difficulty is that mentions from NEL and coreference can have small differences in text spans, such as whether determiners are included. We try to assign each NEL argument to a coreference cluster, first looking for an exact span match, then by looking for the shortest coreference mention that contains it. If the coreference cluster already has matched an NEL argument through exact span match that is different from the one found by looking for the shortest containing coreference mention, the new NEL argument is not added. This gives for each coreference cluster a possible grounding to an entity in the KB. What is provided as final arguments for NEL+Coref argument identification are, in order, grounded NEL arguments, grounded coreference arguments that do not overlap with previous arguments, NIL arguments from NEL that do not overlap with previous arguments, and NIL arguments from coreference that do not overlap with previous arguments.

4 Type-Awareness

Relations have expected types for each argument. Entity types, whether coarse-grained, such as from NER, or fine-grained, such as from Freebase entities, are an important source of information that can be useful for making decisions in relation extraction. We bring type-awareness into the system through partitioning the model, as well as by enforcing type constraints on output relation annotations.

Model Partitioning Instead of building a single relation extractor that can generate sentential instances and then relation annotations with arguments of any type, we can instead build separate relation extractors for each possible coarse type signature, e.g., (PERSON, PERSON), (PERSON, LOCATION), etc., and combine the extractions from the extractor for each type signature. This modification allows each trained model to only handle instances of specific types, and thus relations of that type signature, allowing each to do a better job of choosing relations within the type signature.

Type Constraints We can additionally reject relation annotations where the types of the arguments do not agree with the expected types of the relation. That is, we only accept a relation annotation $s \rightarrow R(m_1, m_2)$ when *EntityTypes*$(e_1) \cap \tau_1 \neq \emptyset$ and *EntityTypes*$(e_2) \cap \tau_2 \neq \emptyset$, where m_1 is linked to e_1, m_2 is linked to e_2, *EntityTypes* provides the set of valid types for an entity, τ_1 is the set of allowed types for the first argument of target relation r, and τ_2 for the second argument.

5 Evaluation Setups

Relation extraction is often evaluated from a macro-reading perspective (Mitchell et al., 2009), in which the extracted facts, $R(e_1, e_2)$, are judged true or false independent of any supporting sentence. For these experiments, however, we take a micro-reading approach in order to strictly evaluate whether a relation extractor is able to extract every fact expressed by a sentence $s \rightarrow R(m_1, m_2)$. Micro-reading is more difficult, but it provides fully semantic information at the sentence and document level allowing detailed justifications, and, for our purposes, allows us to better understand the effects of our modifications. In order to fairly evaluate different systems, even those using different methods of argument identification, we want to use gold evaluation data allowing for varying mention types. We additionally use Hoffmann et al. (2011)'s sentential evaluation as-is in order to better compare with prior work. For our training corpus, we use the TAC-KBP 2009 (McNamee and Dang, 2009) English newswire corpus containing one million documents with 27 million sentences.

5.1 Hoffmann et al. Sentential Evaluation

Hoffmann et al. (2011) generated their gold data by taking the union of sentential instances where some system being evaluated extracted a relation as well as the sentential instances matching arguments in the KB. They took a random sample of these sentential instances and manually labeled them with either a single relation or *NA*. Although this process provides good coverage, since is is sampled from extractions over a large corpus, it does not allow one to measure true recall. Indeed, Hoffmann's method *significantly overestimates recall*, since the random sample is only over sentential instances where a program detected an extraction or a KB match was found. Furthermore, this test set only contains sentential instances in which arguments are marked using NER, which makes it impossible to determine if the use of NEL or

coreference confers any benefit.

Finally, it does not allow for the possibility that there may be multiple relations that should be extracted for a pair of arguments. For example, a CeoOf relation, and an EmployedBy relation might both be present for (Larry Page, Google). To address these issues, we manually annotate a full set of documents with relation annotations. Because we are evaluating changing various aspects of the distant supervision process, we cannot use Riedel et al. (2010)'s distant supervision data as-is as others did on the Hoffmann et al. (2011) sentential evaluation. Instead, we use the TAC-KBP data described above.

5.2 GoReCo Evaluation

In order to allow for variations on mentions (NER, NEL, and coreference each has its own definition of what a mention boundary should be), we want gold relation annotations over coreference clusters broadly defined to allow mentions obtained from NER and NEL, as well as gold coreference mentions. So as long as a relation extraction system extracts a relation annotation $s{\rightarrow}R(m_1, m_2)$ where m_1 and m_2 are allowed options (based on text spans), it will get credit for extracting the relation annotation. We introduce the GoReCo (gold relations and coreference) evaluation to satisfy these constraints.

We start with an existing gold coreference dataset, ACE 2004 (Doddington et al., 2004) newswire, consisting of 128 documents. To get relation annotations over coreference clusters, we define two human annotation tasks and use the BRAT (Stenetorp et al., 2012) tool for visualization and relation and coreference annotations.

Relation Annotation The annotator is presented with a document with gold mentions indicated and asked to determine for each sentence, what facts involving target relations are expressed by the sentence. They draw an arrow for each fact and label it with the relation. They also have the ability to add mentions not present (ReAnn mentions).

Supplemental Coreference Mentions from NER and NEL are displayed along with ACE and ReAnn mentions from the previous task. The annotator draws coreference links from NER or NEL mentions to an ACE or ReAnn mention if they are coreferent.

We randomly shuffle the 128 ACE 2004 newswire documents and use 64 as a development set and 64 as a test set. To complete annotations of these documents, we only used one original human annotator (hired using the oDesk crowdsourcing platform) and found mistakes by having others check the work, as well as checking false positives of relation extractors on the development set to find patterns of annotation mistakes. On average, there are 7 relation annotations per document.

For the GoReCo evaluation, we define our train/test split (with the separate TAC-KBP corpus used for training) such that each has a different set of documents and entities, in order to evaluate how well the system performs on unseen entities. To do this, we remove entities found in the gold evaluation set from the training KB. (We do not remove entities for the Hoffmann et al. (2011) evaluation, since they do not.) We choose the threshold confidence score for each system using the development set to optimize for F1 and report results on the test set.

5.2.1 Target Relations

Since we use a different evaluation, we also seek to choose a more comprehensive and interesting set of relations than prior work. Riedel et al. (2010), whose train and test data is also used by Hoffmann et al. (2011) and Surdeanu et al. (2012), use Freebase properties under domains /people, /business, and /location. Since /location relations such as /location/location/contains dominate the results (and are relatively uninteresting in that they rarely change), we do not use any /location relations, and instead use the domains /people, /business, and /organization (Google, 2012).

Since many Freebase properties are between an entity and a table instead of another entity, we also use joined relations, such as /people/person/employment_history $\bowtie$ /business/employment_tenure/company , in this case representing employment. We bring in an additional 20 relations of this form, also under /person, /business, and /organization. Additionally we use NELL (Carlson et al., 2010a) relations mapped to Freebase by Zhang et al. (2012).

We only include a relation in our set of target relations if both of its entity arguments are contained in the set of entities found via NER with exact string match or NEL over the training corpus. We also remove inverse relations, since they represent needless duplication. This gives us a set

$\mathcal{R}$ of 105 target relations based on joins and unions of Freebase properties. Of the 105 target relations, 48 were used at least once in the GoReCo data.

6 Experiments and Results

We conduct experiments to determine how changing distantly supervised relation extraction along various dimensions affects performance. We examine the choice of argument identification during training and extraction, as well as the effects of model type partitioning, and type constraints. We consider the space of all combinations of these dimensions, but focus on specific combinations where we find improvements.

6.1 Relation Extraction Setup

We use and modify Hoffmann et al. (2011)'s system MultiR to control our experiments and as a baseline. For NER argument identification as well as for the use of NER in the features, we use use Stanford NER (Finkel et al., 2005). For NEL argument identification we use Wikipedia Miner with the default threshold 0.5, and allowing repeated mentions (Milne and Witten, 2008). Since Wikipedia Miner does not support NIL links, we use non-overlapping NER mentions as NIL links. For coreference, we use Stanford's sieve-based deterministic coreference system (Lee et al., 2013). For sentential instance generation, we take all pairs of non-overlapping arguments in a sentence (in either order). If the arguments have KB links, we do not allow sentential instances where both arguments represent the same entity. We use the same lexical and syntactic features as MultiR, based on the features of Mintz et al. (2009). As required for features, we use Stanford CoreNLP's tokenizer, part of speech tagger (Toutanova et al., 2003), and dependency parser (de Marneffe and Manning, 2008), and use the Charniak Johnson constiuent parser (Charniak and Johnson, 2005). For negative training generation, we take a similar approach to Riedel et al. (2010) and define a percentage parameter n of the number of negative instances divided by the number of total instances. Experimenting with $n \in \{0, 20\%, 80\%\}$, we find that $n = 20\%$ works best for our evaluations, optimizing for F1, although using 80% negative training gives high precision at lower recall. We use frequency-based feature selection to eliminate features that appear less than 10 times, which is helpful both for reducing overfitting as well as constraining memory usage. Since the perceptron learning of MultiR is sensitive to instance ordering, we perform 10 random shuffles and average the models.

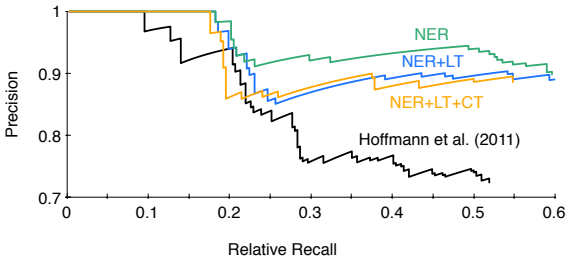


Figure 2: Methods evaluated in the context of Hoffmann et al. (2011)'s sentential extraction evaluation. NER: our NER baseline used for training and extraction, LT: use NEL for training only, CT: use coreference for training only. (NER+LT+CT means we use NER for extraction, and NEL+Coref for training.)

For model type partitioning, when training with NER, we ensure that the NER types match the coarse relation type signatures. For NEL, we attempt to use NER for coarse types of arguments, but if an NER type is not present, we map the Freebase type to its FIGER type (Ling and Weld, 2012) to its coarse type. For type constraints, we use Freebase's expected input and output types for relations. For NIL links, we use the NER type of PERSON, ORGANIZATION, or LOCATION, if available, mapping it to appropriate Freebase types.

6.2 NER Baseline

As a result of a larger training set, as well as model averaging, our baseline, which is otherwise equivalent to the methods of Hoffmann et al. (2011) and uses their MultiR system, has slightly higher precision as shown in Figure 2, curve NER. It is also higher than that of Xu et al. (2013), who achieved higher performance than Hoffmann et al. (2011); our baseline gets 89.9% precision and 59.6% relative recall, while Xu et al. (2013)'s system gets 84.6% precision and 56.1% relative recall. See Figure 3 and Table 1 for results on GoReCo.

6.3 NEL and Type Constraints

On GoReCo, using NEL argument identification increases recall and gives higher precision over the entire curve. We further find that filtering results using type constraints gives a large boost in pre-

cision at a small cost to recall. Note the increase in performance from NER to NEL to NEL+TC in Figure 3a, as well as in Table 1. Using NEL gives more recall, since it is able to capture arguments that NER cannot, such as professions like "paleontologist". The decrease in recall from type constraints comes from false positives in the type constraints process including from non-ideal links, e.g., "paleontologist" might get linked to the entity Paleontology, so will not have the type required for the Profession relation.

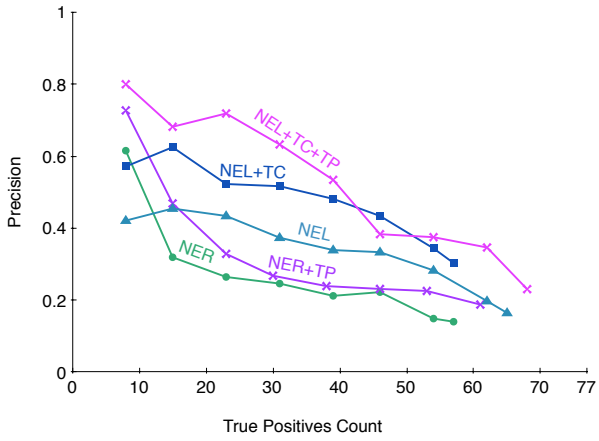
On the Hoffmann et al. (2011) sentential evaluation, we were not able to use NEL argument identification at extraction time, because the instances in the test set are from NER argument identification. We tried using NEL only at training time and found that it got similar performance to using NER (Figure 2, curve NER+LT). Doing the same on GoReCo yielded slightly lower recall, because of the mismatch of features learned from NEL arguments (Figure 3b, curve NER+LT).

6.4 NEL+Coref Argument Identification

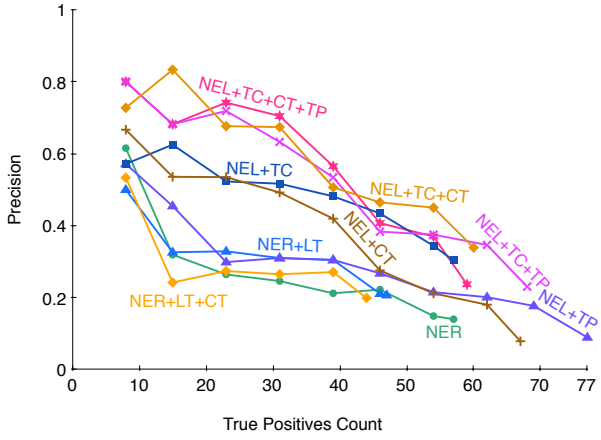
Using NEL+Coref for both training and extraction (see Table 1) introduces noise from arguments not encountered during training time, but using NEL+Coref just for training results in a decrease in recall but similar precision (Figures 2 and 3b).

We found using NEL+Coref at test time unhelpful for this dataset, because there were no examples we could find where coreference could recover arguments that NEL could not. There were three true positives from NEL+Coref involving pronouns in the GoReCo development set, but there were also proper name versions of the arguments nearby in the same sentences, making coreference unnecessary. Additionally, coreference brings in many mentions such as times like "Friday" or "1954" that do not have corresponding KB matches during training time. These sentential instances have similar features to others involving coreference mentions, and there are not negative instances to weigh against these, since these types do not appear in the training data. Better features more suited to coreference mentions could be helpful here.

At both training and extraction time, coreference can cluster together mentions that can be considered to be separate, such as in "Brian Kain, a 33-year-old accountant", "Brian Kain" and "accountant" are coreferent in the gold ACE 2004



(a)



(b)

Figure 3: Precision versus true positives count curves for different versions of the system evaluated on the GoReCo test set, containing 470 gold instances. NER/NEL: argument identification used in training and extraction, LT: use NEL for training only, CT: use coreference for training only, TC: type constraints, TP: model type partitioning.

dataset. This means that type constraints will disregard a Profession annotation between these when it should not, because "Brian Kain" (which would have been a NIL link) gets the link of "accountant". This effect contributes to the decrease in recall.

6.5 Model Type Partitioning

Using type partitioning helps both NER and NEL based models as shown with the +TP curves in Figure 3). Partitioning by type signature results in each model being able to better choose relations for sentential instances of that type signature. In the Partitioned columns of Table 1, removing type partitioning from the best system (NEL training

	Single			Partitioned		
	R	P	F1	R	P	F1
NER training						
NER extraction	7.9	21.8	11.6	11.3	21.0	14.7
NEL extraction	8.5	21.4	12.2	9.8	19.7	13.1
NEL training						
NER extraction	9.6	21.1	13.2	8.9	25.1	13.2
NEL extraction	10.0	30.5	15.1	**15.3**	16.7	16.0
NEL w/TC extraction	11.7	31.1	17.0	13.4	31.3	**18.8**
NEL+Coref training						
NER extraction	9.4	19.2	12.6	6.8	28.3	11.0
NEL extraction	12.1	27.5	16.8	11.1	21.6	14.6
NEL w/TC extraction	**12.8**	**33.3**	**18.5**	12.1	**34.1**	17.9
NEL+Coref extraction	10.6	20.4	14.0	10.0	12.9	11.3
NEL+Coref w/TC extraction	9.4	22.7	13.3	7.9	19.1	11.1

Table 1: Evaluation of different versions of the relation extraction system on the GORECO test set. For nearly all systems, partitioning the model by argument types boosts F1, as does using NEL at either training or extraction time, and using coreference at training time with type constraints (w/TC) raises F1 except with coreference at extraction time and when combined with type partitioning.

and extraction, with type constraints, Partitioned) results in a decrease in F1 from 18.8% to 17.0%. Table 2 shows by-relation performance results for the best system (curve NEL+TC+TP in Figure 3a).

6.6 Other Dimensions Explored

We also experimented with adding generalized features that replaced lexemes with WordNet classes (Fellbaum, 1998), which had uneven results. We observed a small but consistent improvement on the NER baseline (11.6% F1 to 12.7% F1 on GORECO), but after introducing NEL argument identification and partitioning, we no longer observed the improvement. For some relations, there was a small gain in recall that was offset by a loss in precision, but for others, the gain in recall outweighed the loss of precision.

We experimented with a negative instance feedback loop that ran a trained extractor over the training corpus and tested whether each extraction made was in fact a negative example. Even though the training corpus contains one million documents, this method only yielded a few thousand new negative instances due to the difficulty of being certain an extraction should be negative. A naïve approach would simply ensure that both entities participate in a relation in the KB; this is troublesome, because of KB incompleteness and because of type errors. For example Freebase contains BornIn(Barack Obama, Honolulu), but our extractor extracted BornIn(Barack Obama, Hawaii). To avoid labeling this true extraction as a negative instance we have to be robust about location semantics. We selected new negative instances $NA(e_1, e_2)$ from our initial extractor that had e_1 in the knowledge base, with e_1 participating as the first argument in the extracted relation but without e_2 as the second argument. The results were promising for some relations but overall inconclusive as identifying true negatives is quite difficult.

Relation	#Extractions	#TP	#FP
Nationality	**50**	11	38
Profession	43	**23**	20
EmployedBy	27	17	10
Spouse	22	2	20
LivedIn	6	4	2
OrgInCitytown	4	3	1
AthletePlaysForTeam	2	2	0
OrgType	1	1	0

Table 2: By-relation evaluation of the best system (NEL with type constraints and type partitioning) on the GORECO test set. The true positives (TP) are the number of gold relations over coreference clusters that matched, so multiple extractions can match a single true positive.

7 Related Work

There has been much recent work on distantly supervised relation extraction. Mintz et al. (2009) use Freebase to train relation extractors over Wikipedia without labeled data using multi-class logistic regression and lexical and syntactic features. Hoffmann et al. (2011) use a probabilistic graphical model for multi-instance, multi-label

learning and extract over newswire text using Freebase relations. Surdeanu et al. (2012) take a similar approach and use soft constraints and logistic regression. Riedel et al. (2013) integrate open information extraction with schema-based, proposing a universal schema approach, including using features based on latent types. There has also been recent work on reducing noise in distantly supervised relation extraction (Nguyen and Moschitti, 2011; Takamatsu et al., 2012; Roth et al., 2013; Ritter et al., 2013). Xu et al. (2013) and Min et al. (2013) improve the quality of distant supervision training data by reducing false negative examples.

Distant supervision is related to semi-supervised bootstrap learning work such as Carlson et al. (2010b) and many others. Note that distant supervision can be viewed as a subroutine of bootstrap learning; bootstrap learning can continue the process of distant supervision by taking the new tuples found and then training on those again, and repeating the process.

There has also been work on performing NEL and coreference jointly (Cucerzan, 2007; Hajishirzi et al., 2013), however these systems do not perform relation extraction. Singh et al. (2013) performs joint relation extraction, NER, and coreference in a fully-supervised manner. They get slight improvement by adding coreference, but do not use NEL. Ling and Weld (2013) extend MULTIR to find meronym relations in a biology textbook. They get slight improvement over NER by using coreference to pick the best mention of an entity in the sentence for the meronym relation at training and extraction time.

8 Conclusions and Future Work

Given the growing importance of distant supervision, a comprehensive understanding of its variants is crucial. While some of the optimizations we propose may seem intuitive, they have not previously been systematically explored. Our experiments show that NEL, type constraints, and type partitioning are extremely important in order to best take advantage of the seed KB during training as well as known information at extraction time. Our best system results in a 44% increase in precision, and a 70% increase in recall over our NER baseline using GORECO. While we were not able to evaluate all our methods on Hoffmann et al. (2011)'s sentential evaluation, our baseline performs significantly better than previous methods, especially in precision, and training-only modifications perform similarly in both evaluations.

Future work will explore the use of NEL in distantly supervised relation extraction further, tuning a confidence parameter for the NEL system, and determining whether different confidence parameters should be used for training and extraction. Another possible direction is interleaving NEL with relation extraction by using newly extracted facts to try to improve NEL performance.

We freely distribute GORECO a new gold standard evaluation for relation extraction consisting of exhaustive annotations of the 128 documents from ACE 2004 newswire for 48 relations. The source code of our system, its output, as well as our gold data are available at
`http://cs.uw.edu/homes/mkoch/re`.

Acknowledgements

We thank Raphael Hoffmann, Luke Zettlemoyer, Mausam, Xiao Ling, Congle Zhang, Hannaneh Hajishirzi, Leila Zilles, and the anonymous reviewers for helpful feedback. Additionally, we thank Anand Mohan and Graeme Britz for annotations and revisions of the GORECO dataset. This work was supported by Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181, ONR grant N00014-12-1-0211, a gift from Google, a grant from Vulcan, and the WRF / TJ Cable Professorship. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1256082.

References

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010a. Toward an architecture for never-ending language learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-10)*.

Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka, Jr., and Tom M. Mitchell. 2010b. Coupled semi-supervised learning for information extraction. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 101–110, New York, NY, USA. ACM.

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

Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In *EMNLP*.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, pages 708–716.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, CrossParser '08, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Google. 2012. Freebase data dumps. `https://developers.google.com/freebase/data`.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2013. Evaluating entity linking with wikipedia. *Artif. Intell.*, 194:130–150, January.

Hannaneh Hajishirzi, Leila Zilles, Daniel S. Weld, and Luke S. Zettlemoyer. 2013. Joint coreference resolution and named-entity linking with multi-pass sieves. In *EMNLP*, pages 289–299. ACL.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *ACL-HLT*, pages 541–550.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 457–466, New York, NY, USA. ACM.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Comput. Linguist.*, 39(4):885–916, December.

Xiao Ling and Daniel S. Weld. 2012. Fine-grained entity recognition. In *Proceedings of the 26th Conference on Artificial Intelligence (AAAI)*.

Xiao Ling and Daniel S. Weld. 2013. Extracting meronyms for a biology knowledge base using distant supervision. In *Automated Knowledge Base Construction (AKBC) 2013: The 3rd Workshop on Knowledge Extraction at CIKM*.

Paul McNamee and Hoa Trang Dang. 2009. Overview of the tac 2009 knowledge base population track. In *Text Analysis Conference (TAC)*, volume 17, pages 111–113.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 509–518, New York, NY, USA. ACM.

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 NAACL-HLT*, pages 777–782.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL-2009)*, pages 1003–1011.

Tom M. Mitchell, Justin Betteridge, Andrew Carlson, Estevam Hruschka, and Richard Wang. 2009. Populating the semantic web by macro-reading internet text. In *The Semantic Web-ISWC 2009*, pages 998–1002. Springer.

Truc-Vien T. Nguyen and Alessandro Moschitti. 2011. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 277–282, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1375–1384, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *ECML/PKDD (3)*, pages 148–163.

Sebastian Riedel, Limin Yao, Benjamin M. Marlin, and Andrew McCallum. 2013. Relation extraction with matrix factorization and universal schemas. In *Joint Human Language Technology Conference/Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL '13)*, June.

Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. 2013. Modeling missing data in distant supervision for information extraction. *TACL*, 1:367–378.

Benjamin Roth, Tassilo Barth, Michael Wiegand, and Dietrich Klakow. 2013. A survey of noise reduction methods for distant supervision. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 73–78, New York, NY, USA. ACM.

Sameer Singh, Sebastian Riedel, Brian Martin, Jiaping Zheng, and Andrew McCallum. 2013. Joint inference of entities, relations, and coreference. In *CIKM Workshop on Automated Knowledge Base Construction (AKBC)*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, April. Association for Computational Linguistics.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465. Association for Computational Linguistics.

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

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wei Xu, Zhao Le, Raphael Hoffmann, and Ralph Grishman. 2013. Filling knowledge base gaps for distant supervision of relation extraction. In *Proceedings of the 2013 Conference of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria, July. Association for Computational Linguistics.

Congle Zhang, Raphael Hoffmann, and Daniel S. Weld. 2012. Ontological smoothing for relation extraction with minimal supervision. In *AAAI*.