Combining Distant and Partial Supervision for Relation Extraction

Gabor Angeli, Julie Tibshirani, Jean Y. Wu, Christopher D. Manning
Stanford University
Stanford, CA 94305
{angeli, jtibs, jeaneis, manning}@stanford.edu

Abstract

Broad-coverage relation extraction either requires expensive supervised training data, or suffers from drawbacks inherent to distant supervision. We present an approach for providing partial supervision to a distantly supervised relation extractor using a small number of carefully selected examples. We compare against established active learning criteria and propose a novel criterion to sample examples which are both uncertain and representative. In this way, we combine the benefits of fine-grained supervision for difficult examples with the coverage of a large distantly supervised corpus. Our approach gives a substantial increase of 3.9% end-to-end F_1 on the 2013 KBP Slot Filling evaluation, yielding a net F_1 of 37.7%.

1 Introduction

Fully supervised relation extractors are limited to relatively small training sets. While able to make use of much more data, *distantly supervised* approaches either make dubious assumptions in order to simulate fully supervised data, or make use of latent-variable methods which get stuck in local optima easily. We hope to combine the benefits of supervised and distantly supervised methods by annotating a small subset of the available data using selection criteria inspired by active learning.

To illustrate, our training corpus contains 1 208 524 relation mentions; annotating all of these mentions for a fully supervised classifier, at an average of \$0.13 per annotation, would cost approximately \$160 000. Distant supervision allows us to make use of this large corpus without requiring costly annotation. The traditional approach is based on the assumption that every mention of an entity pair (e.g., *Obama* and *USA*) participates in the known relation between the two (i.e., *born in*). However, this introduces noise, as not every mention expresses the relation we are assigning to it.

We show that by providing annotations for only 10 000 informative examples, combined with a large corpus of distantly labeled data, we can yield notable improvements in performance over the distantly supervised data alone. We report results on three criteria for selecting examples to annotate: a baseline of sampling examples uniformly at random, an established active learning criterion, and a new metric incorporating both the *uncertainty* and the *representativeness* of an example. We show that the choice of metric is important – yielding as much as a 3% F_1 difference – and that our new proposed criterion outperforms the standard method in many cases. Lastly, we train a supervised classifier on these collected examples, and report performance comparable to distantly supervised methods. Furthermore, we notice that initializing the distantly supervised model using this supervised classifier is critical for obtaining performance improvements.

This work makes a number of concrete contributions. We propose a novel application of active learning techniques to distantly supervised relation extraction. To the best of the authors knowledge, we are the first to apply active learning to the class of latent-variable distantly supervised models presented in this paper. We show that annotating a proportionally small number of examples yields improvements in end-to-end accuracy. We compare various selection criteria, and show that this decision has a notable impact on the gain in performance. In many ways this reconciles our results with the negative results of Zhang et al. (2012), who show limited gains from naïvely annotating examples. Lastly, we make our annotations available to the research community.[1]

[1] http://nlp.stanford.edu/software/mimlre.shtml

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1556–1567,
October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

2 Background

2.1 Relation Extraction

We are interested in extracting a set of relations $y_1 \ldots y_k$ from a fixed set of possible relations $\mathcal{R}$, given two entities e_1 and e_2. For example, we would like to extract that Barack Obama was born in Hawaii. The task is decomposed into two steps: First, sentences containing mentions of both e_1 and e_2 are collected. The set of these sentences x, marked with the *entity mentions* for e_1 and e_2, becomes the input to the relation extractor, which then produces a set of relations which hold between the mentions. We are predominantly interested in the second step – classifying a set of pairs of entity mentions into the relations they express. Figure 1 gives the general setting for relation extraction, with entity pairs *Barack Obama* and *Hawaii*, and *Barack Obama* and *president*.

Traditionally, relation extraction has fallen into one of four broad approaches: supervised classification, as in the ACE task (Doddington et al., 2004; GuoDong et al., 2005; Surdeanu and Ciaramita, 2007), distant supervision (Craven and Kumlien, 1999; Wu and Weld, 2007; Mintz et al., 2009; Sun et al., 2011; Roth and Klakow, 2013) deterministic rule-based systems (Soderland, 1997; Grishman and Min, 2010; Chen et al., 2010), and translation from open domain information extraction schema (Riedel et al., 2013). We focus on the first two of these approaches.

2.2 Supervised Relation Extraction

Relation extraction can be naturally cast as a supervised classification problem. A corpus of relation mentions is collected, and each mention x is annotated with the relation y, if any, it expresses. The classifier's output is then aggregated to decide the relations between the two entities.

However, annotating supervised training data is generally expensive to perform at large scale. Although resources such as Freebase or the TAC KBP knowledge base have on the order of millions of training tuples over entities it is not feasible to manually annotate the corresponding mentions in the text. This has led to the rise of *distantly supervised* methods, which make use of this indirect supervision, but do not necessitate mention-level supervision.

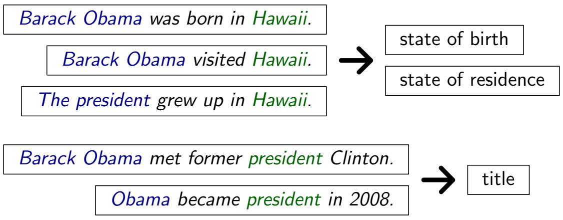

Barack Obama was born in Hawaii.
Barack Obama visited Hawaii.
The president grew up in Hawaii.
state of birth
state of residence
Barack Obama met former president Clinton.
Obama became president in 2008.
title

Figure 1: The relation extraction setup. For a pair of entities, we collect sentences which mention both entities. These sentences are then used to predict one or more relations between those entities. For instance, the sentences containing both *Barack Obama* and *Hawaii* should support the *state of birth* and *state of residence* relation.

2.3 Distant Supervision

Traditional distant supervision makes the assumption that for every triple (e_1, y, e_2) in a knowledge base, every sentence containing mentions for e_1 and e_2 express the relation y. For instance, taking Figure 1, we would create a datum for each of the three sentences containing BARACK OBAMA and HAWAII labeled with *state of birth*, and likewise with *state of residence*, creating 6 training examples overall. Similarly, both sentences involving *Barack Obama* and *president* would be marked as expressing the *title* relation.

While this allows us to leverage a large database effectively, it nonetheless makes a number of naïve assumptions. First – explicit in the formulation of the approach – it assumes that every mention expresses some relation, and furthermore expresses the known relation(s). For instance, the sentence *Obama visited Hawaii* would be erroneously treated as a positive example of the *born in* relation. Second, it implicitly assumes that our knowledge base is complete: entity mentions with no known relation are treated as negative examples.

The first of these assumptions is addressed by multi-instance multi-label (MIML) learning, described in Section 2.4. Min et al. (2013) address the second assumption by extending the MIML model with additional latent variables, while Xu et al. (2013) allow feedback from a coarse relation extractor to augment labels from the knowledge base. These latter two approaches are compatible with but are not implemented in this work.

2.4 Multi-Instance Multi-Label Learning

The multi-instance multi-label (MIML-RE) model of Surdeanu et al. (2012), which builds upon work

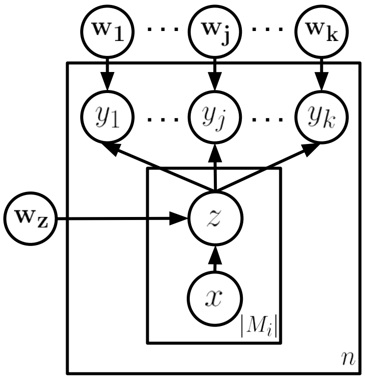


Figure 2: The MIML-RE model, as shown in Surdeanu et al. (2012). The outer plate corresponds to each of the n entity pairs in our knowledge base. Each entity pair has a set of mention pairs M_i, and a corresponding plate in the diagram for each mention pair in M_i. The variable x represents the input mention pair, whereas y represents the positive and negative relations for the given pair of entities. The latent variable z denotes a mention-level prediction for each input. The weight vector for the multinomial z classifier is given by $\mathbf{w}_z$, and there is a weight vector $\mathbf{w}_j$ for each binary y classifier.

by Hoffmann et al. (2011) and Riedel et al. (2010), addresses the assumptions of distantly supervised relations extractors in a principled way by positing a latent mention-level annotation.

The model groups mentions according to their entity pair – for instance, every mention pair with *Obama* and *Hawaii* would be grouped together. A latent variable z_i is created for every mention i, where $z_i \in \mathcal{R} \cup \{\text{None}\}$ takes a single relation label, or a *no relation* marker. We create $|\mathcal{R}|$ binary variables y representing the known positive and negative relations for the entity pair. A set of binary classifiers (log-linear factors in the graphical model) links the latent predictions $z_1 \ldots z_{|M_i|}$ and each y_j. These classifiers include two classes of features: first, a binary feature which fires if *at least one* of the mentions expresses a known relation between the entity pair, and second, a feature for each co-occurrence of relations for a given entity pair. Figure 2 describes the model.

2.5 Background on Active Learning

We describe preliminaries and prior work on active learning; we use this framework to propose two sampling schemes in Section 3 which we use to annotate mention-level labels for MIML-RE.

One way of expressing the generalization error of a hypothesis $\hat{h}$ is through its mean-squared error with the true hypothesis h:

$$\begin{aligned} &E[(h(x) - \hat{h}(x))^2] \\ &\quad = E[E[(h(x) - \hat{h}(x))^2|x]] \\ &\quad = \int_x E[(h(x) - \hat{h}(x))^2|x]p(x)\mathrm{d}x. \end{aligned}$$

The integrand can be further broken into bias and variance terms:

$$\begin{aligned} E[(h(x) - \hat{h}(x))^2] = {} & (E[\hat{h}(x)] - h(x))^2 \\ & + E[(\hat{h}(x) - E[\hat{h}(x)])^2] \end{aligned}$$

where for simplicity we've dropped the conditioning on x.

Many traditional sampling strategies, such as query-by-committee (QBC) (Freund et al., 1992; Freund et al., 1997) and uncertainty sampling (Lewis and Gale, 1994), work by decreasing the variance of the learned model. In QBC, we first create a 'committee' of classifiers by randomly sampling their parameters from a distribution based on the training data. These classifiers then make predictions on the unlabeled examples, and the examples on which there is the most disagreement are selected for labeling. This strategy can be seen as an attempt to decrease the *version space* – the set of classifiers that are consistent with the labeled data. Decreasing the version space should lower variance, since variance is inversely related to the size of the hypothesis space.

In most scenarios, active learning does not concern itself with the bias term. If a model is fundamentally misspecified, then no amount of additional training data can lower its bias. However, our paradigm differs from the traditional setting, in that we are annotating latent variables in a model with a non-convex objective. These annotations may help increase the convexity of our objective, leading us to a more accurate optimum and thereby lowering bias.

The other component to consider is $\int_x \cdots p(x)\mathrm{d}x$. This suggests that it is important to choose examples that are representative of the underlying distribution $p(x)$, as we want to label points that will improve the classifier's predictions on as many and as high-probability examples as possible. Incorporating a representativeness metric has been shown to provide a significant improvement over plain QBC or uncertainty sampling (McCallum and Nigam, 1998; Settles, 2010).

2.6 Active Learning for Relation Extraction

Several papers have explored active learning for relation extraction. Fu and Grishman (2013) employ active learning to create a classifier quickly for new relations, simulated from the ACE corpus. Finn and Kushmerick (2003) compare a number of selection criteria – including QBC – for a supervised classifier. To the best of our knowledge, we are the first to apply active learning to distantly supervised relation extraction. Furthermore, we evaluate our selection criteria live in a real-world setting, collecting new sentences and evaluating on an end-to-end task.

For latent variable models, McCallum and Nigam (1998) apply active learning to semi-supervised document classification. We take inspiration from their use of QBC and the choice of metric for classifier disagreement. However their model assumes a fully Bayesian set-up, whereas ours does not require strong assumptions about the parameter distributions.

Settles et al. (2008) use active learning to improve a multiple-instance classifier. Their model is simpler in that it does not allow for unobserved variables or multiple labels, and the authors only evaluate on image retrieval and synthetic text classification datasets.

3 Example Selection

We describe three criteria for selection examples to annotate. The first – sampling uniformly – is a baseline for our hypothesis that intelligently selecting examples is important. For this criterion, we select mentions uniformly at random from the training set to annotate. This is the approach used in Zhang et al. (2012). The other two criteria rely on a metric for disagreement provided by QBC; we describe our adaptation of QBC for MIML-RE as a preliminary to introducing these criteria.

3.1 QBC For MIML-RE

We use a version of QBC based on bootstrapping (Saar-Tsechansky and Provost, 2004). To create the committee of classifiers, we re-sample the training set with replacement 7 times and train a model over each sampled dataset. We measure disagreement on z-labels among the classifiers using a generalized Jensen-Shannon divergence (McCallum and Nigam, 1998), taking the average KL divergence of all classifier judgments.

We first calculate the mention-level confidences. Note that $z_i^{(m)} \in M_i$ denotes the latent variable in entity pair i with index m; $\mathbf{z}_i^{(-m)}$ denotes the set of all latent variables except $z_i^{(m)}$:

$$\begin{aligned} p(z_i^{(m)}|\mathbf{y_i},\mathbf{x_i}) &= \frac{p(\mathbf{y_i}, z_i^{(m)}|\mathbf{x_i})}{p(\mathbf{y_i}|\mathbf{x_i})} \\ &= \frac{\sum_{\mathbf{z_i^{(-m)}}} p(\mathbf{y_i},\mathbf{z_i}|\mathbf{x_i})}{\sum_{z_i^{(m)}} p(\mathbf{y_i}, z_i^{(m)}|\mathbf{x_i})}. \end{aligned}$$

Notice that the denominator just serves to normalize the probability within a sentence group. We can rewrite the numerator as follows:

$$\begin{aligned} &\sum_{\mathbf{z_i^{(-m)}}} p(\mathbf{y_i},\mathbf{z_i}|\mathbf{x_i}) \\ &\quad = \sum_{\mathbf{z_i^{(-m)}}} p(\mathbf{y_i}|\mathbf{z_i})p(\mathbf{z_i}|\mathbf{x_i}) \\ &\quad = p(z_i^{(m)}|\mathbf{x_i}) \sum_{\mathbf{z_i^{(-m)}}} p(\mathbf{y_i}|\mathbf{z_i})p(\mathbf{z_i^{(-m)}}|\mathbf{x_i}). \end{aligned}$$

For computational efficiency, we approximate $p(\mathbf{z_i^{(-m)}}|\mathbf{x_i})$ with a point mass at its maximum. Next, we calculate the Jensen-Shannon (JS) divergence from the k bootstrapped classifiers:

$$\frac{1}{k}\sum_{c=1}^{k} \mathrm{KL}(p_c(z_i^{(m)}|\mathbf{y_i},\mathbf{x_i})||p_{\text{mean}}(z_i^{(m)}|\mathbf{y_i},\mathbf{x_i})) \quad (1)$$

where p_c is the probability assigned by each of the k classifiers to the latent $z_i^{(m)}$, and p_{mean} is the average of these probabilities. We use this metric to capture the *disagreement* of our model with respect to a particular latent variable. This is then used to inform our selection criteria.

We note that QBC may be especially useful in our situation as our objective is highly nonconvex. If two committee members disagree on a latent variable, it is likely because they converged to different local optima; annotating that example could help bring the classifiers into agreement.

The second selection criterion we consider is the most straightforward application of QBC – selecting the examples with the highest JS disagreement. This allows us to compare our criterion, described next, against an established criterion from the active learning literature.

3.2 Sample by JS Disagreement

We propose a novel active learning sampling criterion that incorporates not only disagreement but also *representativeness* in selecting examples to annotate. Prior work has taken a weighted combination of an example's disagreement and a score corresponding to whether the example is drawn from a dense portion of the feature space (e.g., McCallum and Nigam (1998)). However, this requires both selecting a criterion for defining density (e.g., distance metric in feature space), and tuning a parameter for the relative weight of disagreement versus representativeness.

Instead, we account for choosing representative examples by sampling without replacement proportional to the example's disagreement. Formally, we define the probability of selecting an example $z_i^{(m)}$ to be proportional to the Jensen-Shannon divergence in (1). Since the training set is an approximation to the prior distribution over examples, sampling uniformly over the training set is an approximation to sampling from the prior probability of seeing an input x. We can view our criterion as an approximation to sampling proportional to the product of two densities: a prior over examples x, and the JS divergence mentioned above.

4 Incorporating Sentence-Level Annotations

Following Surdeanu et al. (2012), MIML-RE is trained through hard discriminative Expectation Maximization, inferring the latent z values in the E-step and updating the weights for both the z and y classifiers in the M-step. During the E-step, we constrain the latent z to match our sentence-level annotations when available.

It is worth noting that even in the hard-EM regime, we can in principle incorporate annotator uncertainty elegantly into the model. At each E step, each z_i is set according to

$$z_i{}^{(m)*} \approx \arg\max_{z \in \mathcal{R}} \Big[p(z \mid x_i^{(m)}, \mathbf{w}_z) \times \prod_r p(y_i^{(r)} \mid \mathbf{z}_i', \mathbf{w}_y^{(r)}) \Big]$$

where $\mathbf{z}_i'$ contains the inferred labels from the previous iteration, but with its mth component replaced by $z_i^{(m)}$.

By setting the distribution $p(z \mid x_i^{(m)}, \mathbf{w}_z)$ to reflect uncertainty among annotators, we can leave open the possibility for the model to choose a relation which annotators deemed unlikely, but the model nonetheless prefers. For simplicity, however, we treat our annotations as a hard assignment.

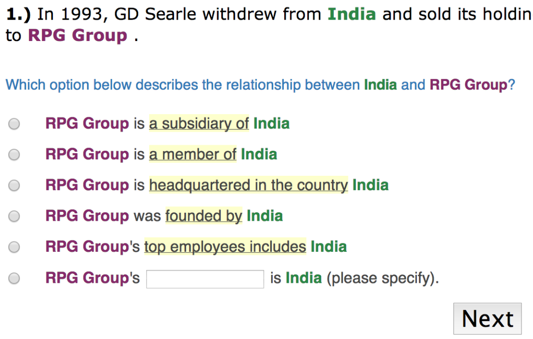
In addition to incorporating annotations during training, we can also use this data to intelligently initialize the model. Since the MIML-RE objective is non-convex, the initialization of the classifier weights w_y and w_z is important. The y classifiers are initialized with the "at-least-once" assumption of Hoffmann et al. (2011); w_z can be initialized either using traditional distant supervision or from a supervised classifier trained on the annotated sentences. If initialized with a supervised classifier, the model can be viewed as augmenting this supervised model with a large distantly labeled corpus, providing both additional entity pairs to train from, and additional mentions for an annotated entity pair.

5 Crowdsourced Example Annotation

Most prior work on active learning is done by simulation on a fully labeled dataset; such a dataset doesn't exist for our case. Furthermore, a key aim of this paper is to practically improve state-of-the-art performance in relation extraction in addition to evaluating active learning criteria. Therefore, we develop and execute an annotation task for collecting labels for our selected examples.

We utilize Amazon Mechanical Turk to crowdsource annotations. For each task, the annotator (Turker) is presented with the task description, followed by 15 questions, 2 of which are randomly placed controls. For each question, we present Turkers with a relation mention and the top 5 relation predictions from our classifier. The Turker also has an option to freely specify a relation not presented in the first five options, or mark that there is no relation. We attempt to heuristically match common free-form answers to official relations.

To maintain the quality of the results, we discard all submissions in which both controls were answered incorrectly, and additionally discard all submissions from Turkers who failed the controls on more than $\frac{1}{3}$ of their submissions. Rejected tasks were republished for other workers to complete. We collect 5 annotations for each example, and use the most commonly agreed answer as the ground truth. Ties are broken arbitrarily, except in



1.) In 1993, GD Searle withdrew from **India** and sold its holdings to **RPG Group** .

Which option below describes the relationship between **India** and **RPG Group**?

○ **RPG Group** is a subsidiary of **India**

○ **RPG Group** is a member of **India**

○ **RPG Group** is headquartered in the country **India**

○ **RPG Group** was founded by **India**

○ **RPG Group**'s top employees includes **India**

○ **RPG Group**'s ______ is **India** (please specify).

Figure 3: The task shown to Amazon Mechanical Turk workers. A sentence along with the top 5 relation predictions from our classifier are shown to Turkers, as well as an option to specify a custom relation or manually enter "no relation." The correct response for this example should be either no relation or a custom relation.

the case of deciding between a relation and no relation, in which case the relation was always chosen.

A total of 23 725 examples were annotated, covering 10 000 examples for each of the three selection criteria. Note that there is overlap between the examples selected for the three criteria. In addition, 10 023 examples were annotated during development; these are included in the set of all annotated examples, but excluded from any of the three criteria. The compensation per task was 23 cents; the total cost of annotating examples was $3156, in addition to $204 spent on developing the task. Informally, Turkers achieved an accuracy of around 75%, as evaluated by a paper author, performing disproportionately well on identifying the *no relation* label.

6 Experiments

We evaluate the three high-level research contributions of this work: we show that we improve the accuracy of MIML-RE, we validate the effectiveness of our selection criteria, and we provide a corpus of annotated examples, evaluating a supervised classifier trained on this corpus. The training and testing methodology for evaluating these contributions is given in Sections 6.1 and 6.2; experiments are given in Section 6.3.

6.1 Training Setup

We adopt the setup of Surdeanu et al. (2012) for training the MIML-RE model, with minor modifications. We use both the 2010 and 2013 KBP official document collections, as well as a July 2013 dump of Wikipedia as our text corpus. We subsample negatives such that $\frac{1}{3}$ of our dataset consists of entity pairs with no known relations. In all experiments, MIML-RE is trained for 7 iterations of EM; for efficiency, the z classifier is optimized using stochastic gradient descent;[2] the y classifiers are optimized using L-BFGS.

Similarly to Surdeanu et al. (2011), we assign negative relations which are either incompatible with the known positive relations (e.g., relations whose co-occurrence would violate type constraints); or, are actually functional relations in which another entity already participates. For example, if we know that Obama was born in the United States, we could add *born in* as a negative relation to the pair Obama and Kenya.

Our dataset consists of 325 891 entity pairs with at least one positive relation, and 158 091 entity pairs with no positive relations. Pairs with at least one known relation have an average of 4.56 mentions per group; groups with no known relations have an average of 1.55 mentions per group. In total, 1 208 524 distinct mentions are considered; the annotated examples are selected from this pool.

6.2 Testing Methodology

We compare against the original MIML-RE model using the same dataset and evaluation methodology as Surdeanu et al. (2012). This allows for an evaluation where the only free variable between this and prior work is the predictions of the relation extractor.

Additionally, we evaluate the relation extractors in the context of Stanford's end-to-end KBP system (Angeli et al., 2014) using the NIST TAC-KBP 2013 English Slotfilling evaluation. In the end-to-end framework, the input to the system is a query entity and a set of articles, and the output is a set of *slot fills* – each slot fill is a candidate triple in the knowledge base, the first element of which is the query entity. This amounts to roughly populating a data structure like Wikipedia infoboxes automatically from a large corpus of text.

Importantly, an end-to-end evaluation in a top-performing full system gives a more accurate idea of the expected real-world gain from each model. Both the information retrieval component providing candidates to the relation extractor, as well as

[2] For the sake of consistency, the supervised classifiers and those in Mintz++ are trained identically to the z classifiers in MIML-RE.

		Active Learning Criterion														
Method	**Init**	Not Used			Uniform			High JS			Sample JS			All Available		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Mintz++	—	41.3	28.2	33.5		—			—			—			—	
MIML-RE	Dist	38.0	30.5	33.8	39.2	30.4	34.2	41.7	28.9	34.1	36.6	31.1	33.6	37.5	30.6	33.7
	Sup	35.1	35.6	35.4	34.4	35.0	34.7	**46.2**	30.8	37.0	39.4	36.2	**37.7**	36.0	**37.1**	36.5
Supervised	—		—		**35.5**	28.9	31.9	31.3	33.2	32.2	33.5	**35.0**	**34.2**	32.9	33.4	33.2

Table 1: A summary of results on the end-to-end KBP 2013 evaluation for various experiments. The first column denotes the algorithm used: either traditional distant supervision (Mintz++), MIML-RE, or a supervised classifier. In the case of MIML-RE, the model may be initialized either using Mintz++, or the corresponding supervised classifier (the "Not Used" column is initialized with the "All" supervised classifier). One of five active learning scenarios are evaluated: no annotated examples provided, the three active learning criteria, and all available examples used. The entry in blue denotes the basic MIML-RE model; entries in gray perform worse than this model. The bold items denote the best performance among selection criteria.

the consistency and inference performed on the classifier output introduce bias in this evaluation's sensitivity to particular types of errors. Mistakes which are easy to filter, or are difficult to retrieve using IR are less important in this evaluation; in contrast, factors such as providing good confidence scores for consistency become more important.

For the end-to-end evaluation, we use the official evaluation script with two changes: First, all systems are evaluated with provenance ignored, so as not to penalize any system for finding a new provenance not validated in the official evaluation key. Second, each system reports its optimal F_1 along its P/R curve, yielding results which are optimistic when compared against other systems entered into the competition. However, this also yields results which are invariant to threshold tuning, and is therefore more appropriate for comparing between systems in this paper.

Development was done on the KBP 2010–2012 queries, and results are reported using the 2013 queries as a simulated test set. Our best system achieves an F_1 of 37.7; the top two teams at KBP 2013 (of 18 entered) achieved F_1 scores of 40.2 and 37.1 respectively, ignoring provenance.

6.3 Results

Table 1 summarizes all results for the end-to-end task; relevant features of the table are copied in subsequent sections to illustrate key trends. Models which perform worse than the original MIML-RE model (MIML-RE, initialized with "Dist," under "Not Used") are denoted in gray. The best performing model improves on the base model by 3.9 F_1 points on the end-to-end task.

System	**P**	**R**	**F_1**
Mintz++	41.3	28.2	33.5
MIML + Dist	38.0	30.5	33.8
MIML + Sup	35.1	35.6	35.4
MIML + Dist + SampleJS	36.6	31.1	33.6
MIML + Sup + SampleJS	**39.4**	**36.2**	**37.7**

Table 2: A summary of improvements to MIML-RE on the end-to-end slotfilling task, copied from Table 1. Mintz++ is the traditional distantly supervised model. The second row corresponds to the unmodified MIML-RE model. The third row corresponds to MIML-RE initialized with a supervised classifier (trained on all examples). The fourth row is MIML-RE with annotated examples incorporated during training (but not initialization). The last row shows the best results obtained by our model.

We evaluate each of the individual contributions of the paper: improving the accuracy of the MIML-RE relation extractor, evaluating our example selection criteria, and demonstrating the annotated examples' effectiveness for a fully-supervised relation extractor.

Improve MIML-RE Accuracy A key goal of this work is to improve the accuracy of the MIML-RE model; we show that we improve the model both on the end-to-end slotfilling task (Table 2) as well as on a standard evaluation (Figure 5). Similar to our work, recent work by Pershina et al.

System	P	R	F_1
MIML + Sup	35.1	35.6	35.4
MIML + Sup + Uniform	34.4	35.0	34.7
MIML + Sup + HighJS	**46.2**	30.8	37.0
MIML + Sup + SampleJS	39.4	36.2	**37.7**
MIML + Sup + All	36.0	**37.1**	36.5

Table 3: A summary of the performance of each example selection criterion. In each case, the model was initialized with a supervised classifier. The first row corresponds to the MIML-RE model initialized with a supervised classifier. The middle three rows show performance for the three selection criteria, used both for initialization and during training. The last row shows results if all available annotations are used, independent of their source.

System	P	R	F_1
Mintz++	**41.3**	28.2	33.5
MIML + Dist	38.0	30.5	33.8
Supervised + SampleJS	33.5	**35.0**	**34.2**
MIML + Sup	35.1	35.6	35.5
MIML + Sup + SampleJS	39.4	36.2	37.7

Table 4: A comparison of the best performing supervised classifier with other systems. The top section compares the supervised classifier with prior work. The lower section highlights the improvements gained from initializing MIML-RE with a supervised classifier.

(2014) incorporates labeled data to guide MIML-RE during training. They make use of labeled data to extract *training guidelines*, which are intended to generalize across many examples. We show that we can match or outperform their improvements with our best criterion.

A few interesting trends emerge from the end-to-end results in Table 2. Using annotated sentences during training alone did not improve performance consistently, even hurting performance when the SampleJS criterion was used. This supports an intuition that the initialization of the model is important, and that it is relatively difficult to coax the model out of a local optimum if it is initialized poorly. This is further supported by the improvement in performance when the model is initialized with a supervised classifier, even when no examples are used during training. Similar trends are reported in prior work, e.g., Smith and Eisner (2007) Section 4.4.6.

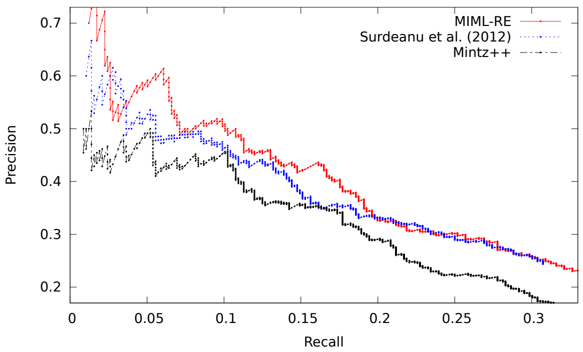


Figure 4: MIML-RE and Mintz++ evaluated according to Surdeanu et al. (2012). The original model from the paper is plotted for comparison, as our training methodology is somewhat different.

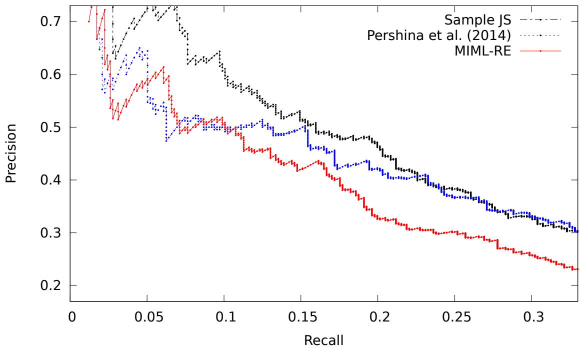


Figure 5: Our best active learning criterion evaluated against our version of MIML-RE, alongside the best system of Pershina et al. (2014).

Also interesting is the relatively small gain MIML-RE provides over traditional distant supervision (Mintz++) in this setting. We conjecture that the mistakes made by Mintz++ are often relatively easily filtered by the downstream consistency component. This is supported by Figure 4; we evaluate our trained MIML-RE model against Mintz++ and the results reported in Surdeanu et al. (2012). We show that our model performs as well or better than the original implementation, and consistently outperforms Mintz++.

Evaluate Selection Criteria A key objective of this work is to evaluate how much of an impact careful selection of annotated examples has on the overall performance of the system. We evaluate the three selection criteria from Section 3.2, showing the results for MIML-RE in Table 3; results for the supervised classifier are given in Table 1. In both cases, we show that the sampled JS criterion performs comparably to or better than the other criteria.

At least two interesting trends can be noted from these results: First, the uniformly sampled criterion performed worse than MIML-RE initialized with a supervised classifier. This may be due to noise in the annotation: a small number of annotation errors on entity pairs with only a single corresponding mention could introduce dangerous noise into training. These singleton mentions will rarely have disagreement between the committee of classifiers, and therefore will generally only be selected in the uniform criterion.

Second, adding in the full set of examples did not improve performance – in fact, performance generally dropped in this scenario. We conjecture that this is due to the inclusion of the uniformly sampled examples, with performance dropping for the same reasons as above.

Both of these results can be reconciled with the results of Zhang et al. (2012); like this work, they annotated examples to analyze the trade-off between adding more data to a distantly supervised system, and adding more direct supervision. They conclude that annotations provide only a relatively small improvement in performance. However, their examples were uniformly selected from the training corpus, and did not make use of the structure provided by MIML-RE. Our results agree in that neither the uniform selection criterion nor the supervised classifier significantly outperformed the unmodified MIML-RE model; nonetheless, we show that if care is taken in selecting these labeled examples we can achieve noticeable improvements in accuracy.

We also evaluate our selection criteria on the evaluation of Surdeanu et al. (2012), both initialized with Mintz++ (Figure 7) and with the supervised classifier (Figure 6). These results mirror those in the end-to-end evaluation; when initialized with the supervised classifier the high disagreement (High JS) and sampling proportional to disagreement (Sample JS) criteria clearly outperform both the base MIML-RE model as well as the uniformly sampling criterion. Using the annotated examples only during training yielded no perceivable benefit over the base model (Figure 7).

Supervised Relation Extractor The examples collected can be used to directly train a supervised classifier, with results summarized in Table 4. The most salient insight is that the performance of the best supervised classifier is similar to that of the MIML-RE model, despite being trained on nearly two orders of magnitude less training data.

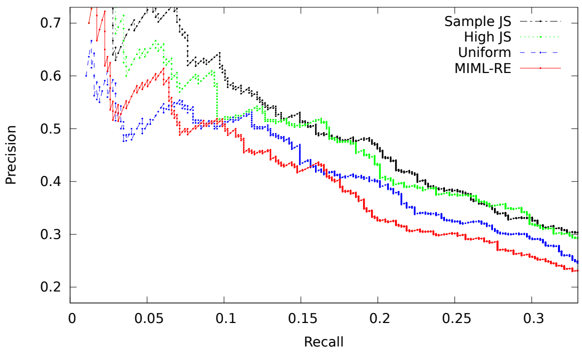


Figure 6: A comparison of models trained with various selection criteria on the evaluation of Surdeanu et al. (2012), all initialized with the corresponding supervised classifier.

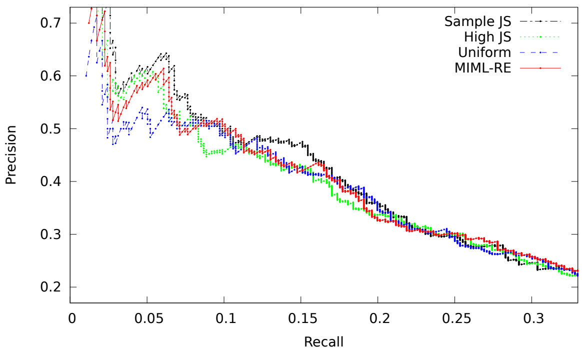


Figure 7: A comparison of models trained with various selection criteria on the evaluation of Surdeanu et al. (2012), all initialized with Mintz++.

More interestingly, however, the supervised classifier provides a noticeably better initialization for MIML-RE than Mintz++, yielding better results even without enforcing the labels during EM. These results suggest that the power gained from the the more sophisticated MIML-RE model is best used in conjunction with a small amount of training data. That is, using MIML-RE as a principled model for combining a large distantly labeled corpus and a small number of careful annotations yields significant improvement over using either of the two data sources alone.

Relation	#	P		R		F_1	
no_relation	3073						
employee_of	1978	29	**32**	33	**46**	31	**38**
countries_of_res.	1061	30	**42**	7	**40**	11	**41**
states_of_residence	427	57	**33**	14	**7**	23	**12**
cities_of_residence	356	31	**52**	9	**30**	14	**38**
(org:)member_of	290	0	0	0	0	0	0
country_of_hq	280	63	**62**	65	**62**	64	**62**
top_members	221	36	**26**	50	**60**	42	**36**
country_of_birth	205	22	**0**	40	**0**	29	**0**
parents	196	10	**26**	31	**54**	15	**35**
city_of_hq	194	46	**52**	57	**61**	51	**56**
(org:)alt_names	184	52	**48**	39	39	45	**43**
founded_by	180	100	**89**	29	**38**	44	**53**
city_of_birth	145	17	**50**	8	**17**	11	**25**
state_of_hq	132	50	**64**	30	**35**	38	**45**
title	121	20	**26**	28	**35**	23	**30**
subsidiaries	105	33	**25**	6	**3**	10	**5**
founded	90	62	**82**	62	**69**	62	**75**
spouse	88	37	**54**	85	85	51	**66**
origin	86	42	**43**	68	**70**	51	**53**
state_of_birth	83	0	**50**	0	**10**	0	**17**
charges	69	54	54	16	16	24	24
cause_of_death	69	93	93	39	39	55	55
(per:)alt_names	69	9	**20**	2	**3**	3	**6**
country_of_death	65	100	100	10	10	18	18
members	54	0	0	0	0	0	0
children	52	53	**62**	14	**18**	22	**27**
parents	50	64	64	28	28	39	39
city_of_death	38	42	**75**	16	**19**	23	**30**
dissolved	38	0	0	0	0	0	0
date_of_death	33	64	64	44	**39**	52	**48**
political_affiliation	23	7	**25**	100	100	13	**40**
state_of_death	19	0	0	0	0	0	0
shareholders	19	0	0	0	0	0	0
siblings	16	50	50	33	33	40	40
schools_attended	14	80	**78**	41	**48**	54	**60**
date_of_birth	11	100	100	85	85	92	92
other_family	9	0	0	0	0	0	0
age	4	94	**97**	94	**90**	94	**93**
#_of_employees	3	0	0	0	0	0	0
religion	2	100	100	29	29	44	44
website	0	25	**0**	3	**0**	6	**0**

Table 5: A summary of relations annotated, and end-to-end slotfilling performance by relation. The first column gives the relation; the second shows the number of examples annotated. The subsequent columns show the performance of the unmodified MIML-RE model and our best performing model (SampleJS). Changes in values are bolded; positive changes are shown in green and negative changes in red. The most frequent 10 relations in the evaluation are likewise bolded.

6.4 Analysis By Relation

In this section, we explore which of the KBP relations were shown to Turkers, and whether the improvements in accuracy correspond to these relations. We compare only the unmodified MIML-RE model, and our best model (MIML-RE initialized with the supervised classifier, under the SampleJS criterion). Results are shown in Table 5.

A few interesting trends emerge from this analysis. We note that annotating even 80+ examples for a relation seems to provide a consistent boost in accuracy, whereas relations with fewer annotated examples tended to show little or no change. However, the gains of our model are not universal across relation types, even dropping noticeably on some – for instance, F_1 drops on both *state of residence* and *country of birth*. This could suggest systematic noise from Turker judgments; e.g., for foreign geography (*state of residence*) or ambiguous relations (*top members*).

An additional insight from the table is the mismatch between examples chosen to be annotated, and the most popular relations in the KBP evaluation. For instance, by far the most popular KBP relation (*title*) had only 121 examples annotated.

7 Conclusion

We have shown that providing a relatively small number of mention-level annotations can improve the accuracy of MIML-RE, yielding an end-to-end improvement of 3.9 F_1 on the KBP task. Furthermore, we have introduced a new active learning criterion, and shown both that the choice of criterion is important, and that our new criterion performs well. Lastly, we make available a dataset of mention-level annotations for constructing a traditional supervised relation extractor.

Acknowledgements

We thank the anonymous reviewers for their thoughtful comments, and Dan Weld for his feedback on additional experiments and analysis. We gratefully acknowledge the support of the Defense Advanced Research Projects Agency (DARPA) Deep Exploration and Filtering of Text (DEFT) Program under Air Force Research Laboratory (AFRL) contract no. FA8750-13-2-0040. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA, AFRL, or the US government.

References

Gabor Angeli, Arun Chaganty, Angel Chang, Kevin Reschke, Julie Tibshirani, Jean Y. Wu, Osbert Bastani, Keith Siilats, and Christopher D. Manning. 2014. Stanford's 2013 KBP system. In *TAC-KBP*.

Zheng Chen, Suzanne Tamang, Adam Lee, Xiang Li, Wen-Pin Lin, Matthew Snover, Javier Artiles, Marissa Passantino, and Heng Ji. 2010. CUNY-BLENDER. In *TAC-KBP*.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *AAAI*.

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*.

Aidan Finn and Nicolas Kushmerick. 2003. Active learning selection strategies for information extraction. In *International Workshop on Adaptive Text Extraction and Mining*.

Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. 1992. Information, prediction, and query by committee. In *NIPS*.

Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. 1997. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2-3):133–168.

Lisheng Fu and Ralph Grishman. 2013. An efficient active learning framework for new relation types. In *IJCNLP*.

Ralph Grishman and Bonan Min. 2010. New York University KBP 2010 slot-filling system. In *Proc. TAC 2010 Workshop*.

Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. 2005. Exploring various knowledge in relation extraction. In *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*.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR*.

Andrew McCallum and Kamal Nigam. 1998. Employing EM and pool-based active learning for text classification. In *ICML*.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *NAACL-HLT*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL*.

Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. 2014. Infusion of labeled data into distant supervision for relation extraction. In *ACL*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*. Springer.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *NAACL-HLT*.

Benjamin Roth and Dietrich Klakow. 2013. Feature-based models for improving the quality of noisy training data for relation extraction. In *CIKM*.

Maytal Saar-Tsechansky and Foster Provost. 2004. Active sampling for class probability estimation and ranking. *Machine Learning*, 54(2):153–178.

Burr Settles, Mark Craven, and Soumya Ray. 2008. Multiple-instance active learning. In *Advances in neural information processing systems*, pages 1289–1296.

Burr Settles. 2010. Active learning literature survey. *University of Wisconsin Madison Technical Report 1648*.

Noah Smith and Jason Eisner. 2007. *Novel estimation methods for unsupervised discovery of latent structure in natural language text*. Ph.D. thesis, Johns Hopkins.

Stephen G Soderland. 1997. *Learning text analysis rules for domain-specific natural language processing*. Ph.D. thesis, University of Massachusetts.

Ang Sun, Ralph Grishman, Wei Xu, and Bonan Min. 2011. New York University 2011 system for KBP slot filling. In *Proceedings of the Text Analytics Conference*.

Mihai Surdeanu and Massimiliano Ciaramita. 2007. Robust information extraction with perceptrons. In *ACE07 Proceedings*.

Mihai Surdeanu, Sonal Gupta, John Bauer, David McClosky, Angel X Chang, Valentin I Spitkovsky, and Christopher D Manning. 2011. Stanfords distantly-supervised slot-filling system. In *Proceedings of the Text Analytics Conference*.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *EMNLP*.

Fei Wu and Daniel S Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of the sixteenth ACM conference on information and knowledge management*. ACM.

Wei Xu, Le Zhao, Raphael Hoffman, and Ralph Grishman. 2013. Filling knowledge base gaps for distant supervision of relation extraction. In *ACL*.

Ce Zhang, Feng Niu, Christopher Ré, and Jude Shavlik. 2012. Big data versus the crowd: Looking for relationships in all the right places. In *ACL*.