Appears in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (ACL), 2012.

Big Data versus the Crowd: Looking for Relationships in All the Right Places

Ce Zhang **Feng Niu** **Christopher Ré** **Jude Shavlik**
Department of Computer Sciences
University of Wisconsin-Madison, USA
{czhang,leonn,chrisre,shavlik}@cs.wisc.edu

Abstract

Classically, training relation extractors relies on high-quality, manually annotated training data, which can be expensive to obtain. To mitigate this cost, NLU researchers have considered two newly available sources of less expensive (but potentially lower quality) labeled data from distant supervision and crowd sourcing. There is, however, no study comparing the relative impact of these two sources on the precision and recall of post-learning answers. To fill this gap, we empirically study how state-of-the-art techniques are affected by scaling these two sources. We use corpus sizes of up to 100 million documents and tens of thousands of crowd-source labeled examples. Our experiments show that increasing the corpus size for distant supervision has a statistically significant, positive impact on quality (F1 score). In contrast, human feedback has a positive and statistically significant, but lower, impact on precision and recall.

1 Introduction

Relation extraction is the problem of populating a *target relation* (representing an entity-level relationship or attribute) with facts extracted from natural-language text. Sample relations include people's titles, birth places, and marriage relationships.

Traditional relation-extraction systems rely on manual annotations or domain-specific rules provided by experts, both of which are scarce resources that are not portable across domains. To remedy these problems, recent years have seen interest in the *distant supervision* approach for relation extraction (Wu and Weld, 2007; Mintz et al., 2009). The input to distant supervision is a set of *seed facts* for the target relation together with an (unlabeled) text corpus, and the output is a set of (noisy) annotations that can be used by any machine learning technique to train a statistical model for the target relation. For example, given the target relation `birthPlace`(person, place) and a seed fact `birthPlace`(John, Springfield), the sentence "*John and his wife were born in Springfield in 1946*" (S1) would qualify as a positive training example.

Distant supervision replaces the expensive process of manually acquiring annotations that is required by direct supervision with resources that already exist in many scenarios (seed facts and a text corpus). On the other hand, distantly labeled data may not be as accurate as manual annotations. For example, "*John left Springfield when he was 16*" (S2) would also be considered a positive example about place of birth by distant supervision as it contains both John and Springfield. The hypothesis is that the broad coverage and high redundancy in a large corpus would compensate for this noise. For example, with a large enough corpus, a distant supervision system may find that patterns in the sentence S1 strongly correlate with seed facts of `birthPlace` whereas patterns in S2 do not qualify as a strong indicator. Thus, intuitively the quality of distant supervision should improve as we use larger corpora. However, there has been no study on the impact of corpus size on distant supervision for relation extraction. Our goal is to fill this gap.

Besides "big data," another resource that may be valuable to distant supervision is crowdsourc-

ing. For example, one could employ crowd workers to provide feedback on whether distant supervision examples are correct or not (Gormley et al., 2010). Intuitively the crowd workforce is a perfect fit for such tasks since many erroneous distant labels could be easily identified and corrected by humans. For example, distant supervision may mistakenly consider "*Obama took a vacation in Hawaii*" a positive example for `birthPlace` simply because a database says that Obama was born in Hawaii; a crowd worker would correctly point out that this sentence is not actually indicative of this relation.

It is unclear however which strategy one should use: scaling the text corpus or the amount of human feedback. Our primary contribution is to empirically assess how scaling these inputs to distant supervision impacts its result quality. We study this question with input data sets that are orders of magnitude larger than those in prior work. While the largest corpus (Wikipedia and New York Times) employed by recent work on distant supervision (Mintz et al., 2009; Yao et al., 2010; Hoffmann et al., 2011) contain about 2M documents, we run experiments on a 100M-document (50X more) corpus drawn from ClueWeb.[1] While prior work (Gormley et al., 2010) on crowdsourcing for distant supervision used thousands of human feedback units, we acquire tens of thousands of human-provided labels. Despite the large scale, we follow state-of-the-art distant supervision approaches and use deep linguistic features, e.g., part-of-speech tags and dependency parsing.[2]

Our experiments shed insight on the following two questions:

1. *How does increasing the corpus size impact the quality of distant supervision?*
2. *For a given corpus size, how does increasing the amount of human feedback impact the quality of distant supervision?*

We found that increasing corpus size consistently and significantly improves recall and F1, despite reducing precision on small corpora; in contrast, human feedback has relatively small impact on precision and recall. For example, on a TAC corpus with 1.8M documents, we found that increasing the corpus size ten-fold consistently results in statistically significant improvement in F1 on two standardized relation extraction metrics (t-test with p=0.05). On the other hand, increasing human feedback amount ten-fold results in statistically significant improvement on F1 only when the corpus contains at least 1M documents; and the magnitude of such improvement was only one fifth compared to the impact of corpus-size increment.

We find that the quality of distant supervision tends to be *recall gated*, that is, for any given relation, distant supervision fails to find all possible linguistic signals that indicate a relation. By expanding the corpus one can expand the number of patterns that occur with a known set of entities. Thus, as a rule of thumb for developing distant supervision systems, one should first attempt to expand the training corpus and then worry about precision of labels only after having obtained a broad-coverage corpus.

Throughout this paper, it is important to understand the difference between *mentions* and *entities*. Entities are conceptual objects that exist in the world (e.g., Barack Obama), whereas authors use a variety of wordings to refer to (which we call "mention") entities in text (Ji et al., 2010).

2 Related Work

The idea of using entity-level structured data (e.g., facts in a database) to generate mention-level training data (e.g., in English text) is a classic one: researchers have used variants of this idea to extract entities of a certain type from webpages (Hearst, 1992; Brin, 1999). More closely related to relation extraction is the work of Lin and Patel (2001) that uses dependency paths to find answers that express the same relation as in a question.

Since Mintz et al. (2009) coined the name *"distant supervision,"* there has been growing interest in this technique. For example, distant supervision has been used for the TAC-KBP slot-filling tasks (Surdeanu et al., 2010) and other relation-extraction tasks (Hoffmann et al., 2010; Carlson et al., 2010; Nguyen and Moschitti, 2011a; Nguyen and Moschitti, 2011b). In contrast, we study how increasing input size (and incorporating human feedback) improves the result quality of distant supervision.

We focus on logistic regression, but it is interesting future work to study more sophisticated prob-

[1] `http://lemurproject.org/clueweb09.php/`

[2] We used 100K CPU hours to run such tools on ClueWeb.

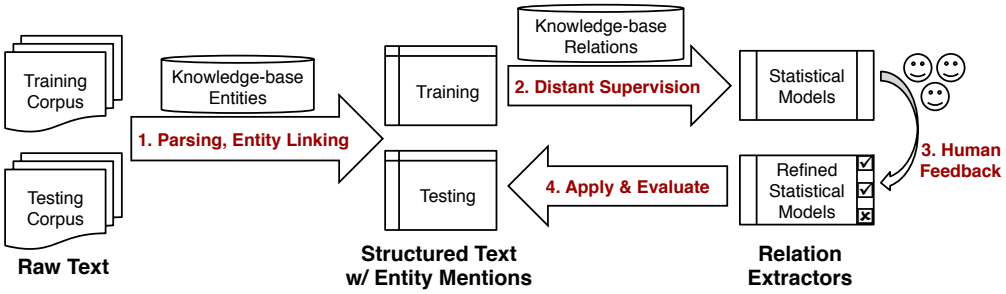


Figure 1: The workflow of our distant supervision system. Step 1 is preprocessing; step 4 is final evaluation. The key steps are distant supervision (step 2), where we train a logistic regression (LR) classifier for each relation using (noisy) examples obtained from sentences that match Freebase facts, and human feedback (step 3) where a crowd workforce refines the LR classifiers by providing feedback to the training data.

abilistic models; such models have recently been used to relax various assumptions of distant supervision (Riedel et al., 2010; Yao et al., 2010; Hoffmann et al., 2011). Specifically, they address the noisy assumption that, if two entities participate in a relation in a knowledge base, then all co-occurrences of these entities express this relation. In contrast, we explore the effectiveness of increasing the training data sizes to improve distant-supervision quality.

Sheng et al. (2008) and Gormley et al. (2010) study the quality-control issue for collecting training labels via crowdsourcing. Their focus is the collection process; in contrast, our goal is to quantify the impact of this additional data source on distant-supervision quality. Moreover, we experiment with one order of magnitude more human labels. Hoffmann et al. (2009) study how to acquire end-user feedback on relation-extraction results posted on an augmented Wikipedia site; it is interesting future work to integrate this source in our experiments. One technique for obtaining human input is active learning. We tried several active-learning techniques as described by Settles (2010), but did not observe any notable advantage over uniform sampling-based example selection.[3]

3 Distant Supervision Methodology

Relation extraction is the task of identifying relationships between mentions, in natural-language text, of entities. An example relation is that two persons are married, which for mentions of entities x and y is denoted $R(x, y)$. Given a corpus C containing mentions of named entities, our goal is to learn a classifier for $R(x, y)$ using linguistic features of x and y, e.g., dependency-path information. The problem is that we lack the large amount of labeled examples that are typically required to apply supervised learning techniques. We describe an overview of these techniques and the methodological choices we made to implement our study. Figure 1 illustrates the overall workflow of a distant supervision system. At each step of the distant supervision process, we closely follow the recent literature (Mintz et al., 2009; Yao et al., 2010).

3.1 Distant Supervision

Distant supervision compensates for a lack of training examples by generating what are known as *silver-standard examples* (Wu and Weld, 2007). The observation is that we are often able to obtain a structured, but incomplete, database D that instantiates relations of interest and a text corpus C that contains mentions of the entities in our database. Formally, a database is a tuple $D = (E, \bar{R})$ where E is a set of entities and $\bar{R} = (R_1 \ldots, R_N)$ is a tuple of instantiated predicates. For example, R_i may contain pairs of married people.[4] We use the facts in R_i combined with C to generate examples.

Following recent work (Mintz et al., 2009; Yao et al., 2010; Hoffmann et al., 2011), we use Freebase[5] as the knowledge base for seed facts. We use two text corpora: (1) the TAC-KBP[6] 2010 corpus that

[3] More details in our technical report (Zhang et al., 2012).

[4] We only consider binary predicates in this work.

[5] http://freebase.com

[6] KBP stands for "Knowledge-Base Population."

consists of 1.8M newswire and blog articles[7], and (2) the ClueWeb09 corpus that is a 2009 snapshot of 500M webpages. We use the TAC-KBP slot filling task and select those TAC-KBP relations that are present in the Freebase schema as targets (20 relations on people and organization).

One problem is that relations in D are defined at the entity level. Thus, the pairs in such relations are not embedded in text, and so these pairs lack the linguistic context that we need to extract features, i.e., the features used to describe examples. In turn, this implies that these pairs cannot be used directly as training examples for our classifier. To generate training examples, we need to map the entities back to mentions in the corpus. We denote the relation that describes this mapping as the relation $\mathrm{EL}(e, m)$ where $e \in E$ is an entity in the database D and m is a mention in the corpus C. For each relation R_i, we generate a set of (noisy) positive examples denoted R_i^+ defined as $R_i^+ =$

$$\{(m_1, m_2) \mid R(e_1, e_2) \wedge EL(e_1, m_1) \wedge EL(e_2, m_2)\}$$

As in previous work, we impose the constraint that both mentions $(m_1, m_2) \in R_i^+$ are contained in the same sentence (Mintz et al., 2009; Yao et al., 2010; Hoffmann et al., 2011). To generate negative examples for each relation, we follow the assumption in Mintz et al. (2009) that relations are disjoint and sample from other relations, i.e., $R_i^- = \cup_{j \neq i} R_j^+$.

3.2 Feature Extraction

Once we have constructed the set of possible mention pairs, the state-of-the-art technique to generate feature vectors uses linguistic tools such as part-of-speech taggers, named-entity recognizers, dependency parsers, and string features. Following recent work on distant supervision (Mintz et al., 2009; Yao et al., 2010; Hoffmann et al., 2011), we use both lexical and syntactic features. After this stage, we have a well-defined machine learning problem that is solvable using standard supervised techniques. We use *sparse logistic regression* (ℓ_1 regularized) (Tibshirani, 1996), which is used in previous studies. Our feature extraction process consists of three steps:

1. Run Stanford CoreNLP with POS tagging and named entity recognition (Finkel et al., 2005);
2. Run dependency parsing on TAC with the Ensemble parser (Surdeanu and Manning, 2010) and on ClueWeb with MaltParser (Nivre et al., 2007)[8]; and
3. Run a simple entity-linking system that utilizes NER results and string matching to identify mentions of Freebase entities (with types).[9]

The output of this processing is a repository of structured objects (with POS tags, dependency parse, and entity types and mentions) for sentences from the training corpus. Specifically, for each pair of entity mentions (m_1, m_2) in a sentence, we extract the following features $F(m_1, m_2)$: (1) the word sequence (including POS tags) between these mentions after normalizing entity mentions (e.g., replacing "John Nolen" with a place holder PER); if the sequence is longer than 6, we take the 3-word prefix and the 3-word suffix; (2) the dependency path between the mention pair. To normalize, in both features we use lemmas instead of surface forms. We discard features that occur in fewer than three mention pairs.

3.3 Crowd-Sourced Data

Crowd sourcing provides a cheap source of human labeling to improve the quality of our classifier. In this work, we specifically examine feedback on the result of distant supervision. Precisely, we construct the union of $R_1^+ \cup \ldots R_N^+$ from Section 3.1. We then solicit human labeling from Mechanical Turk (MTurk) while applying state-of-the-art quality control protocols following Gormley et al. (2010) and those in the MTurk manual.[10]

These quality-control protocols are critical to ensure high quality: spamming is common on MTurk and some turkers may not be as proficient or careful as expected. To combat this, we replicate each question three times and, following Gormley

[7] http://nlp.cs.qc.cuny.edu/kbp/2010/

[8] We did not run Ensemble on ClueWeb because we had very few machines satisfying Ensemble's memory requirement. In contrast, MaltParser requires less memory and we could leverage Condor (Thain et al., 2005) to parse ClueWeb with MaltParser within several days (using about 50K CPU hours).

[9] We experiment with a slightly more sophisticated entity-linking system as well, which resulted in higher overall quality. The results below are from the simple entity-linking system.

[10] http://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf

et al. (2010), plant gold-standard questions: each task consists of five yes/no questions, one of which comes from our gold-standard pool.[11] By retaining only those answers that are consistent with this protocol, we are able to filter responses that were not answered with care or competency. We only use answers from workers who display overall high consistency with the gold standard (i.e., correctly answering at least 80% of the gold-standard questions).

3.4 Statistical Modeling Issues

Following Mintz et al. (2009), we use logistic regression classifiers to represent relation extractors. However, while Mintz et al. use a single multi-class classifier for all relations, Hoffman et al. (2011) and use an independent binary classifier for each individual relation; the intuition is that a pair of mentions (or entities) might participate in multiple target relations. We experimented with both protocols; since relation overlapping is rare for TAC-KBP and there was little difference in result quality, we focus on the binary-classification approach using training examples constructed as described in Section 3.1.

We compensate for the different sizes of distant and human labeled examples by training an objective function that allows to tune the weight of human versus distant labeling. We separately tune this parameter for each training set (with cross validation), but found that the result quality was robust with respect to a broad range of parameter values.[12]

4 Experiments

We describe our experiments to test the hypotheses that the following two factors improve distant-supervision quality: increasing the

(1) corpus size, and

(2) the amount of crowd-sourced feedback.

We confirm hypothesis (1), but, surprisingly, are unable to confirm (2). Specifically, when using logistic regression to train relation extractors, increasing corpus size improves, consistently and significantly, the precision and recall produced by distant supervision, regardless of human feedback levels. Using the methodology described in Section 3, human feedback has limited impact on the precision and recall produced from distant supervision by itself.

[11] We obtain the gold standard from a separate MTurk submission by taking examples that at least 10 out of 11 turkers answered yes, and then negate half of these examples by altering the relation names (e.g., `spouse` to `sibling`).

[12] More details in our technical report (Zhang et al., 2012).

4.1 Evaluation Metrics

Just as direct training data are scarce, ground truth for relation extraction is scarce as well. As a result, prior work mainly considers two types of evaluation methods: (1) randomly sample a small portion of predictions (e.g., top-k) and manually evaluate precision/recall; and (2) use a held-out portion of seed facts (usually Freebase) as a kind of "distant" ground truth. We replace manual evaluation with a standardized relation-extraction benchmark: TAC-KBP 2010. TAC-KBP asks for extractions of 46 relations on a given set of 100 entities. Interestingly, the Freebase held-out metric (Mintz et al., 2009; Yao et al., 2010; Hoffmann et al., 2011) turns out to be heavily biased toward distantly labeled data (e.g., increasing human feedback *hurts* precision; see Section 4.6).

4.2 Experimental Setup

Our first group of experiments use the 1.8M-doc TAC-KBP corpus for training. We exclude from it the 33K documents that contain query entities in the TAC-KBP metrics. There are two key parameters: the corpus size (#docs) M and human feedback budget (#examples) N. We perform different levels of down-sampling on the training corpus. On TAC, we use subsets with $M = 10^3, 10^4, 10^5$, and 10^6 documents respectively. For each value of M, we perform 30 independent trials of uniform sampling, with each trial resulting in a training corpus D_i^M, $1 \leq i \leq 30$. For each training corpus D_i^M, we perform distant supervision to train a set of logistic regression classifiers. From the full corpus, distant supervision creates around 72K training examples.

To evaluate the impact of human feedback, we randomly sample 20K examples from the input corpus (we remove any portion of the corpus that is used in an evaluation). Then, we ask three different crowd workers to label each example as either positive or negative using the procedure described in Section 3.3. We retain only credible answers using the gold-standard method (see Section 3.3), and use them as the pool of human feedback that we run experiments with. About 46% of our human labels are negative. Denote by N the number of examples that

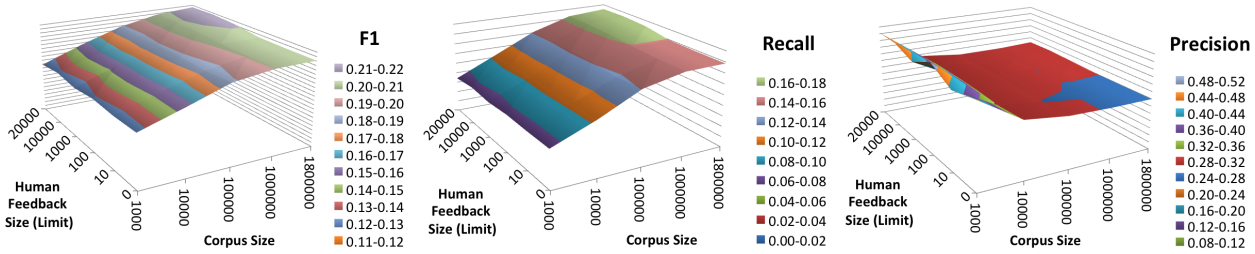


Figure 2: Impact of input sizes under the TAC-KBP metric, which uses documents mentioning 100 predefined entities as testing corpus with entity-level ground truth. We vary the sizes of the training corpus and human feedback while measuring the scores (F1, recall, and precision) on the TAC-KBP benchmark.

we want to incorporate human feedback for; we vary N in the range of $0, 10, 10^2, 10^3, 10^4$, and 2×10^4. For each selected corpus and value of N, we perform without-replacement sampling from examples of this corpus to select feedback for up to N examples. In our experiments, we found that on average an M-doc corpus contains about $0.04M$ distant labels, out of which $0.01M$ have human feedback. After incorporating human feedback, we evaluate the relation extractors on the TAC-KBP benchmark. We then compute the average F1, recall, and precision scores among all trials for each metric and each (M,N) pair. Besides the KBP metrics, we also evaluate each (M,N) pair using Freebase held-out data. Furthermore, we experiment with a much larger corpus: ClueWeb09. On ClueWeb09, we vary M over $10^3, \ldots, 10^8$. Using the same metrics, we show at a larger scale that increasing corpus size can significantly improve both precision and recall.

4.3 Overall Impact of Input Sizes

We first present our experiment results on the TAC corpus. As shown in Figure 2, the F1 graph closely tracks the recall graph, which supports our earlier claim that quality is recall gated (Section 1). While increasing the corpus size improves F1 at a roughly log-linear rate, human feedback has little impact until both corpus size and human feedback size approch maximum M, N values. Table 1 shows the quality comparisons with minimum/maximum values of M and N.[13] We observe that increasing the corpus size significant improves per-relation recall

	$M = 10^3$	$M = 1.8 \times 10^6$
$N = 0$	0.124	0.201
$N = 2 \times 10^4$	0.118	0.214

Table 1: TAC F1 scores with max/min values of M/N.

and F1 on 17 out of TAC-KBP's 20 relations; in contrast, human feedback has little impact on recall, and only significantly improves the precision and F1 of 9 relations – while hurting F1 of 2 relations (i.e., `MemberOf` and `LivesInCountry`).[14]

(a) Impact of corpus size changes.

$M \backslash N$	0	10	10^2	10^3	10^4	2e4
$10^3 \rightarrow 10^4$	+	+	+	+	+	+
$10^4 \rightarrow 10^5$	+	+	+	+	+	+
$10^5 \rightarrow 10^6$	+	+	+	+	+	+
$10^6 \rightarrow$ 1.8e6	0	0	0	+	+	+

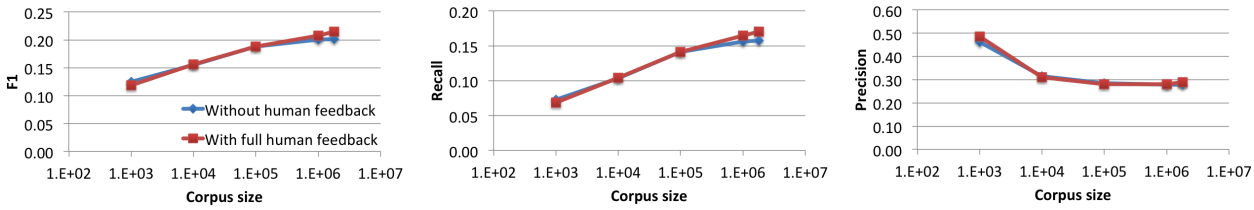
(b) Impact of feedback size changes.

$N \backslash M$	10^3	10^4	10^5	10^6	1.8e6
$0 \rightarrow 10$	0	0	0	0	0
$10 \rightarrow 10^2$	0	0	0	+	+
$10^2 \rightarrow 10^3$	0	0	0	+	+
$10^3 \rightarrow 10^4$	0	0	0	0	+
$10^4 \rightarrow$ 2e4	0	0	0	0	-
$0 \rightarrow$ 2e4	0	0	0	+	+

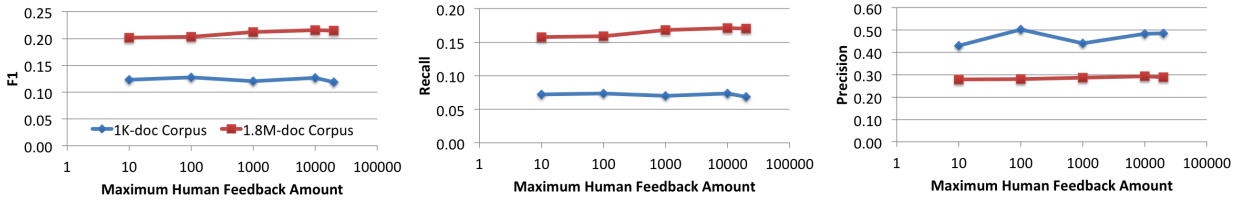
Table 2: Two-tail t-test with d.f.=29 and p=0.05 on the impact of corpus size and feedback size changes respectively. (We also tried p=0.01, which resulted in change of only a single cell in the two tables.) In (a), each column corresponds to a fixed human-feedback budget size N. Each row corresponds to a jump from one corpus size (M) to the immediate larger size. Each cell value indicates whether the TAC F1 metric changed significantly: + (resp. -) indicates that the quality increased (resp. decreased) significantly; 0 indicates that the quality did not change significantly. Table (b) is similar.

[13] When the corpus size is small, the total number of examples with feedback can be smaller than the budget size N – for example, when $M = 10^3$ there are on average 10 examples with feedback even if $N = 10^4$.

[14] We report more details on per-relation quality in our technical report (Zhang et al., 2012).



(a) Impact of corpus size changes.



(b) Impact of human feedback size.

Figure 3: Projections of Figure 2 to show the impact of corpus size and human feedback amount on TAC-KBP F1, recall, and precision.

4.4 Impact of Corpus Size

In Figure 3(a) we plot a projection of the graphs in Figure 2 to show the impact of corpus size on distant-supervision quality. The two curves correspond to when there is no human feedback and when we use all applicable human feedback. The fact that the two curves almost overlap indicates that human feedback had little impact on precision or recall. On the other hand, the quality improvement rate is roughly log-linear against the corpus size. Recall that each data point in Figure 2 is the average from 30 trials. To measure the statistical significance of changes in F1, we calculate t-test results to compare adjacent corpus size levels given each fixed human feedback level. As shown in Table 2(a), increasing the corpus size by a factor of 10 consistently and significantly improves F1. Although precision decreases as we use larger corpora, the decreasing trend is sub-log-linear and stops at around 100K docs. On the other hand, recall and F1 keep increasing at a log-linear rate.

4.5 Impact of Human Feedback

Figure 3(b) provides another perspective on the results under the TAC metric: We fix a corpus size and plot the F1, recall, and precision as functions of human-feedback amount. Confirming the trend in Figure 2, we see that human feedback has little

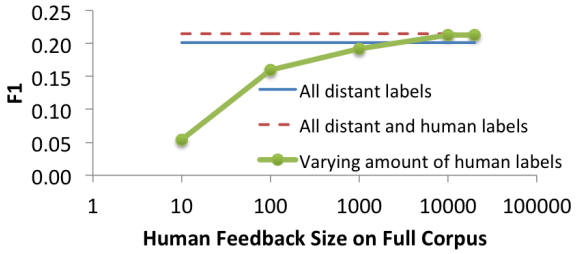


Figure 4: TAC-KBP quality of relation extractors trained using different amounts of human labels. The horizontal lines are comparison points.

impact on precision or recall with both corpus sizes.

We calculate t-tests to compare adjacent human feedback levels given each fixed corpus size level. Table 2(b)'s last row reports the comparison, for various corpus sizes (and, hence, number of distant labels), of (i) using no human feedback and (ii) using *all* of the human feedback we collected. When the corpus size is small (fewer than 10^5 docs), human feedback has no statistically significant impact on F1. The locations of $+$'s suggest that the influence of human feedback becomes notable only when the corpus is very large (say with 10^6 docs). However, comparing the slopes of the curves in Figure 3(b) against Figure 3(a), the impact of human feedback is substantially smaller. The precision graph in Figure 3(b) suggests that human feedback does not no-

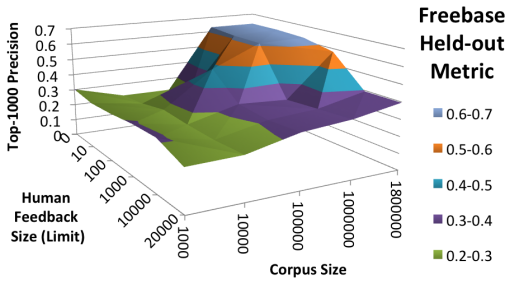


Figure 5: Impact of input sizes under the Freebase held-out metric. Note that the human feedback axis is in the reverse order compared to Figure 2.

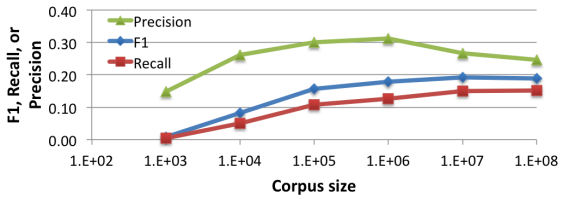


Figure 6: Impact of corpus size on the TAC-KBP quality with the ClueWeb dataset.

tably improve precision on either the full corpus or on a small 1K-doc corpus. To assess the quality of human labels, we train extraction models with human labels only (on examples obtained from distant supervision). We vary the amount of human labels and plot the F1 changes in Figure 4. Although the F1 improves as we use more human labels, the best model has roughly the same performance as those trained from distant labels (with or without human labels). This suggests that the accuracy of human labels is not substantially better than distant labels.

4.6 Freebase Held-out Metric

In addition to the TAC-KBP benchmark, we also follow prior work (Mintz et al., 2009; Yao et al., 2010; Hoffmann et al., 2011) and measure the quality using held-out data from Freebase. We randomly partition both Freebase and the corpus into two halves. One database-corpus pair is used for training and the other pair for testing. We evaluate the precision over the 10^3 highest-probability predictions on the test set. In Figure 5, we vary the size of the corpus in the train pair and the number of human labels; the precision reaches a dramatic peak when we the corpus size is above 10^5 *and* uses little human feedback. This suggests that this Freebase held-out metric is biased toward solely relying on distant labels alone.

4.7 Web-scale Corpora

To study how a Web corpus impacts distant-supervision quality, we select the first 100M English webpages from the ClueWeb09 dataset and measure how distant-supervision quality changes as we vary the number of webpages used. As shown in Figure 6, increasing the corpus size improves F1 up to 10^7 docs ($p = 0.05$), while at 10^8 the two-tailed significance test reports no significant impact on F1 ($p = 0.05$). The dip in precision in Figure 6 from 10^6 to either 10^7 or 10^8 is significant ($p = 0.05$), and it is interesting future work to perform a detailed error analysis. Recall from Section 3 that to preprocess ClueWeb we use MaltParser instead of Ensemble. Thus, the F1 scores in Figure 6 are not comparable to those from the TAC training corpus.

5 Discussion and Conclusion

We study how the size of two types of cheaply available resources impact the precision and recall of distant supervision: (1) an unlabeled text corpus from which distantly labeled training examples can be extracted, and (2) crowd-sourced labels on training examples. We found that text corpus size has a stronger impact on precision and recall than human feedback. We observed that distant-supervision systems are often *recall gated*; thus, to improve distant-supervision quality, one should first try to enlarge the input training corpus and then increase precision.

It was initially counter-intuitive to us that human labels did not have a large impact on precision. One reason is that human labels acquired from crowd-sourcing have comparable noise level as distant labels – as shown by Figure 4. Thus, techniques that improve the accuracy of crowd-sourced answers are an interesting direction for future work. We used a particular form of human input (yes/no votes on distant labels) and a particular statistical model to incorporate this information (logistic regression). It is interesting future work to study other types of human input (e.g., new examples or features) and more sophisticated techniques for incorporating human input, as well as machine learning methods that explicitly model feature interactions.

Acknowledgements

We gratefully acknowledge the support of the Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181. CR is also generously supported by NSF CAREER award under IIS-1054009, ONR award N000141210041, and gifts or research awards from Google, Greenplum, Johnson Controls, Inc., LogicBlox, and Oracle. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA, AFRL, or the US government. We are thankful for the generous support from the Center for High Throughput Computing, the Open Science Grid, and Miron Livny's Condor research group at UW-Madison. We are also grateful to Dan Weld for his insightful comments on the manuscript.

References

S. Brin. 1999. Extracting patterns and relations from the world wide web. In *Proceedings of The World Wide Web and Databases*, pages 172–183.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka Jr, and T. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Conference on Artificial Intelligence*, pages 1306–1313.

J. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 363–370.

M. Gormley, A. Gerber, M. Harper, and M. Dredze. 2010. Non-expert correction of automatically generated relation annotations. In *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 204–207.

M. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics-Volume 2*, pages 539–545.

R. Hoffmann, S. Amershi, K. Patel, F. Wu, J. Fogarty, and D.S. Weld. 2009. Amplifying community content creation with mixed initiative information extraction. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 1849–1858. ACM.

R. Hoffmann, C. Zhang, and D. Weld. 2010. Learning 5000 relational extractors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 286–295.

R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 541–550.

H. Ji, R. Grishman, H.T. Dang, K. Griffitt, and J. Ellis. 2010. Overview of the TAC 2010 knowledge base population track. In *Text Analysis Conference*.

D. Lin and P. Pantel. 2001. Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(4):343–360.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1003–1011.

T.V.T. Nguyen and A. Moschitti. 2011a. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proceeding of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 277–282.

T.V.T. Nguyen and A. Moschitti. 2011b. Joint distant and direct supervision for relation extraction. In *Proceeding of the International Joint Conference on Natural Language Processing*, pages 732–740.

J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135.

S. Riedel, L. Yao, and A. McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, pages 148–163.

B. Settles. 2010. Active learning literature survey. Technical report, Computer Sciences Department, University of Wisconsin-Madison, USA.

V.S. Sheng, F. Provost, and P.G. Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.

M. Surdeanu and C. Manning. 2010. Ensemble models for dependency parsing: Cheap and good? In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 649–652.

M. Surdeanu, D. McClosky, J. Tibshirani, J. Bauer, A.X. Chang, V.I. Spitkovsky, and C. Manning. 2010. A simple distant supervision approach for the TAC-KBP slot filling task. In *Proceedings of Text Analysis Conference 2010 Workshop*.

D. Thain, T. Tannenbaum, and M. Livny. 2005. Distributed computing in practice: The Condor experience. *Concurrency and Computation: Practice and Experience*, 17(2-4):323–356.

R. Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

F. Wu and D. Weld. 2007. Autonomously semantifying wikipedia. In *ACM Conference on Information and Knowledge Management*, pages 41–50.

L. Yao, S. Riedel, and A. McCallum. 2010. Collective cross-document relation extraction without labelled data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1013–1023.

C. Zhang, F. Niu, C. Ré, and J. Shavlik. 2012. Big data versus the crowd: Looking for relationships in all the right places (extended version). Technical report, Computer Sciences Department, University of Wisconsin-Madison, USA.