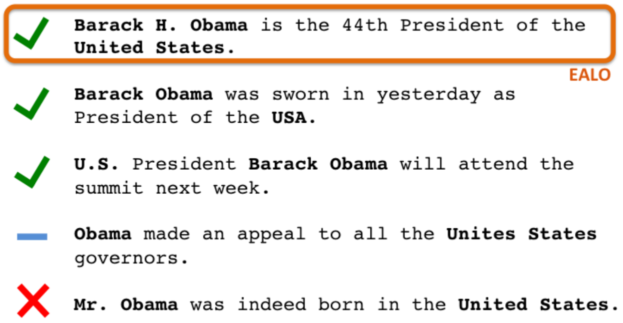
Assuming Facts Are Expressed More Than Once

Justin Betteridge and **Alan Ritter** and **Tom Mitchell**
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, USA
{jbetter, rittera, tom.mitchell}@cs.cmu.edu
http://rtw.ml.cmu.edu

Abstract

Distant supervision (DS) is a method for training sentence-level information extraction models using only an unlabeled corpus and a knowledge base (KB). Fundamental to many DS approaches is the assumption that KB facts are expressed at least once (EALO) in the text corpus. Often, however, KB facts are actually expressed in the corpus *many* times, in which cases EALO-based systems underuse the available training data. To address this problem, we introduce the "expressed at least α percent" (EALA) assumption, which asserts that expressions of KB facts account for up to α% of the corresponding mentions. We show that for the same level of precision as the EALO approach, the EALA approach achieves up to 66% higher recall on category recognition and 53% higher recall on relation recognition.



✓ **Barack H. Obama** is the 44th President of the **United States.**

✓ **Barack Obama** was sworn in yesterday as President of the **USA.**

✓ **U.S.** President **Barack Obama** will attend the summit next week.

— **Obama** made an appeal to all the **Unites States** governors.

✗ **Mr. Obama** was indeed born in the **United States.**

Figure 1: Sentences mentioning BarackObama and UnitedStates. The EALO assumption only asserts that at least one of these expresses the fact PresidentOf(BarackObama,UnitedStates), when in reality 3 out of these 5 sentences actually do.

1 Introduction

Distant supervision (DS) (Craven and Kumlien 1999; Morgan et al. 2004; Bunescu and Mooney 2007; Mintz et al. 2009; Riedel, Yao, and McCallum 2010; Hoffmann et al. 2011; Surdeanu et al. 2012; Ritter et al. 2013) has recently emerged as a popular way to scale up Information Extraction (IE) systems beyond the handfuls of semantic predicates for which expensive, manually-annotated training data exist. This trend is highly significant because in order to enable widespread adoption for truly intelligent, text-based applications, such as semantic search, question answering (QA), and deep, knowledge-based natural language understanding (NLU), IE systems must be successfully scaled to tens of thousands of semantic predicates, both unary (categories) and binary (relations).

The key idea behind distant supervision is that training examples for sentence-level extraction models can be automatically distilled from an unlabeled text corpus through the use of a repository of out-of-context facts, called a knowledge base (KB). For example, suppose the goal is to learn a model for recognizing instances of the PresidentOf relation and that a KB contains the fact PresidentOf(BarackObama,UnitedStates). The goal of distant supervision, then, is to somehow use this fact to automatically identify which sentences in an unlabeled text corpus can provide useful signals to the learning process.

The first step is consider the set of sentences from the corpus that mention the instance[1] ⟨BarackObama,UnitedStates⟩. Suppose that the sentences shown in Figure 1 comprise that set. Precisely how this set of sentences is used to learn model parameters is what differentiates various DS algorithms.

Early DS approaches (Bunescu and Mooney 2007; Mintz et al. 2009) were based on the *distant supervision assumption* (Riedel, Yao, and McCallum 2010):

> *All mentions of an instance in the training corpus express the corresponding KB fact(s).*

Such systems would simply treat all the sentences in Figure 1 as labeled examples of PresidentOf mentions. However, the last mention in Figure 1 certainly does not express or directly imply the PresidentOf relation. To address the

[1] In this paper, we use the following general terminology to emphasize that the approaches we discuss apply equally well to both categories and relations. A *predicate* is either a category or a relation. A *predicate instance*, or just *instance*, is either an entity (a category instance) or a pair of entities (a relation instance). A *mention* is a reference to an instance in a particular sentence.

inevitable noise caused by using this assumption, Riedel et al. (2010) introduced the "expressed at least once" (EALO) assumption:

> *Each KB fact associated with an instance is expressed by at least one mention in the training corpus.*

Approaches based on this principle (Riedel, Yao, and McCallum 2010; Hoffmann et al. 2011) would intelligently select only a single mention from Figure 1 to treat as a labeled PRESIDENTOF mention. As this hypothetical example illustrates, however, there may be many more mentions that could provide useful signals to the learning process. Thus, although the EALO assumption addresses precision errors introduced by the more liberal distant supervision assumption, it leads to underutilization of the available training data and subsequent recall errors. This is especially true with category recognition (which was not addressed by the research cited in this paragraph) because category instances typically have many corpus mentions.

To address this problem, we make the following contributions:

- We propose a new assumption, called the "expressed at least α percent" (EALA) assumption, which constrains the system to use more than one mention when updating parameters.
- We present a model and algorithm for implementing this new assumption.
- We empirically evaluate whether the EALA assumption improves performance over the EALO assumption on both the category and relation recognition tasks. In Section 4, we show that for the same level of precision as the EALO approach, the EALA approach achieves significantly higher recall on both category and relation recognition.
- We also demonstrate that our distant supervision algorithm can be successfully scaled up using a distributed implementation to train on millions of training examples.

2 Related Work

Traditional approaches to information extraction (Grishman and Sundheim 1996; Doddington, Mitchell, and Przybocki 2004) rely on manually annotated corpora to train machine learning models or develop linguistic rules. While this approach has been quite successful, reliance on manually constructed data limits their applicability across domains. Therefore, researchers have looked at many different ways to induce predicate mention recognizers without manually labeled training data.

Distant supervision, also called *weak* or *minimal* supervision, was introduced by Craven and Kumlien (1999). Most of the previous work on distant supervision has addressed learning to recognize either relations (Craven and Kumlien 1999; Bunescu and Mooney 2007; Mintz et al. 2009; Riedel, Yao, and McCallum 2010; Hoffmann et al. 2011; Surdeanu et al. 2012; Min et al. 2013) or categories (Morgan et al. 2004; Whitelaw et al. 2008; Huang and Riloff 2010), and only very recently, both (Ritter et al. 2013). In this paper, we simultaneously address both categories and relations with the same distant supervision framework.

The task of recognizing, or extracting, semantic predicate instances at the corpus level has also received much attention. Hearst (1992) pioneered a solution for this task using textual patterns to learn instances of semantic categories. Brin (1998) described one of the first systems for automatically learning instances of a binary relation. Etzioni et al. (2005) leveraged Hearst's patterns in a bootstrapping fashion to learn category facts from the Web. Carlson et al. (2010) use coupling constraints between predicates to reduce semantic drift. Nakashole et al. (2011) describe recent work in large-scale fact extraction. Riedel et al. (2013) have recently introduced a matrix factorization approach which achieves state-of-the art performance on the aggregate-level relation extraction task.

Open IE (Banko et al. 2007) is another related task in which a system extracts instances of relations that are not predefined. Rather than relying on a schema or ontology to define the relations to be extracted, OpenIE methods use linguistic patterns to extract strings describing the relation directly from the text. This approach is flexible and covers a broad range of relations, but doesn't resolve relation mentions to a common representation, often resulting in multiple representations expressing essentially the same meaning.

Finally, a number of researchers (Bunescu and Mooney 2007; Riedel, Yao, and McCallum 2010; Hoffmann et al. 2011; Surdeanu et al. 2012; Min et al. 2013) have recognized that distant supervision is a form of *multiple-instance learning* (MIL) (Dietterich, Lathrop, and Lozano-Perez 1997). In this work we extend the notion of positive bags in MIL to be those which contain *at least* α% positive instances as opposed to *at least one* positive instance, showing significant improvement in recall for both relation and category recognition.

This work is most closely related to that of Hoffmann et al. (2011), who introduced the MultiR algorithm, which is based on the EALO assumption, because we use the same general latent-variable approach in implementing our model.

3 Expressed At Least α Percent

In this work, we introduce the following generalization of the EALO assumption:

> *The KB facts corresponding to an instance are expressed*[2] *by at least* α% *of the training corpus mentions of that instance.*

We refer to this assumption as the "expressed at least α percent" (EALA) assumption and show in Section 4 that it leads to improved performance over the EALO assumption.

3.1 Latent Variable Model

The task addressed by this work is, given a set of prespecified predicates $\mathcal{P}$, to predict the most appropriate predicate label for each mention of each instance in an input corpus of text. As in the MultiR framework, each instance is

[2] in equal proportion, to be consistent with the EALO assumption

inherently associated with all of its mentions from the input corpus, and the labels on the mentions of a particular instance are assumed to be (collectively) independent of any other mention labels. Therefore, we present the model in the context of a single instance i with its corresponding set of mentions and mention labels.

Let M be the number of mentions of i in the corpus. Each mention is represented by an observed vector $\mathbf{X}_j$ of binary features, with $\mathbf{X} = \{\mathbf{X}_j\}$ containing the feature vectors for all the mentions of i. The primary output of the system is the latent predicate label $Z_j \in \mathcal{P}$ for each mention $\mathbf{X}_j$, with $\mathbf{Z} = \{Z_j\}$ containing the labels for all the mentions of i. Crucially, the model also includes a set $\mathbf{Y} = \{Y_p\}$ of aggregate-level binary labels, one for each of the $L = |\mathcal{P}|$ predicates in $\mathcal{P}$. Each variable Y_p represents whether the corpus supports the notion that a particular predicate p is "true" for instance i, i.e., that a sufficient number of i's mentions have been labeled as p (this will be made more concrete below). $\mathbf{Y}$ is unobserved during prediction, but observed during training, as we will further explain below.

Our approach is based on the following conditional distribution (the normalization constant $Z_{\boldsymbol{\theta}}(\mathbf{x})$ is omitted for brevity):

$$P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{z}|\mathbf{x}) \propto \Phi_{\alpha}^{EALA}(\mathbf{y}, \mathbf{z}) \prod_j \Phi_{\boldsymbol{\theta}}^{ft}(\mathbf{x}_j, z_j)$$

The workhorse of the model is $\Phi_{\boldsymbol{\theta}}^{ft}$, a parameterized, log-linear factor between the features and the label of a particular mention:

$$\Phi_{\boldsymbol{\theta}}^{ft}(\mathbf{x}_j, z_j) = \exp\left\{\sum_k \theta_k \phi_k(\mathbf{x}_j, z_j)\right\}$$

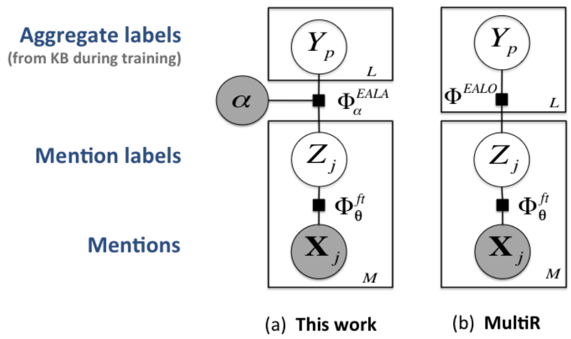
where each k corresponds to a particular feature/label combination $\{f, l\}$, and ϕ_k is a indicator function returning 1 only if f is active in $\mathbf{x}_j$ and $z_j = l$, and 0 otherwise. The weight vector $\boldsymbol{\theta} = \{\theta_k\}$ constitutes the parameters of the model, which the system learns during training.

The heart of the model is Φ^{EALA}, which explicitly encodes the EALA assumption and therefore is fundamental to the entire approach. Although this factor is not parameterized, it constrains $P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{z}|\mathbf{x})$ by assigning zero probability to configurations of $\mathbf{y}$ and $\mathbf{z}$ that violate the EALA assumption:

$$\Phi_{\alpha}^{EALA}(\mathbf{y}, \mathbf{z}) = \begin{cases} 1 & \text{if } \forall p : y_p = 1, n_{\mathbf{z},p} \geq \alpha M / n_{\mathbf{y},1} \\ 0 & \text{otherwise} \end{cases}$$

where $n_{\mathbf{z},p} = |\mathbf{z}_{(p)}|$ is the number of mention labels in $\mathbf{z}$ whose value is p and $n_{\mathbf{y},1} = |\mathbf{y}_{(1)}|$ is the number of aggregate labels in $\mathbf{y}$ whose value is 1. Figure 2(a) shows a graphical representation of the model. We also show a representation of the MultiR approach in Figure 2(b) (using our notation) to highlight the similarities and differences between the two models.

Assuming the set of parameters $\boldsymbol{\theta}$ has already been learned, the prediction task is to find the most probable mention labels $\hat{\mathbf{z}}$:



(a) This work (b) MultiR

Figure 2: Graphical models for (a) this work and (b) MultiR (Hoffmann et al. 2011)

$$\hat{\mathbf{z}} = \arg\max_{\mathbf{z}} \; P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{z}|\mathbf{x})$$
$$= \arg\max_{\mathbf{z}} \; \prod_j \Phi_{\boldsymbol{\theta}}^{ft}(\mathbf{x}_j, z_j)$$

where the last simplification is possible because $\mathbf{y}$ is essentially a deterministic function of $\mathbf{z}$ and can therefore be ignored when computing $\hat{\mathbf{z}}$.

3.2 Training

To learn the model parameters $\boldsymbol{\theta}$, our system maximizes the conditional probability of aggregate labels $\mathbf{y}^*$ derived from a knowledge base, given the observed mention features $\mathbf{x}$, or equivalently, the log of that probability, subject to L_1 and L_2 regularization[3]:

$$O(\boldsymbol{\theta}) = \sum_i O_i(\boldsymbol{\theta}) - \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 - \tau\|\boldsymbol{\theta}\|_1$$
$$O_i(\boldsymbol{\theta}) = \ln P_{\boldsymbol{\theta}}(\mathbf{y}^*|\mathbf{x})$$
$$= \ln \sum_{\mathbf{z}} P_{\boldsymbol{\theta}}(\mathbf{y}^*, \mathbf{z}|\mathbf{x})$$

This objective is maximized using stochastic gradient ascent, where the primary terms of the gradient are:

$$\frac{\partial}{\partial \theta_k} \ln O_i(\boldsymbol{\theta}) = \mathbb{E}_{P_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y},\mathbf{x})}[\sum_j \phi_k(\mathbf{x}_j, z_j)]$$
$$- \mathbb{E}_{P_{\boldsymbol{\theta}}(\mathbf{y},\mathbf{z}|\mathbf{x})}[\sum_j \phi_k(\mathbf{x}_j, z_j)]$$

However, because it is not feasible to compute these expectations exactly, we follow the MultiR approach and replace the expectations with maximizations. This requires computing the following two settings of mention labels:

$$\mathbf{z}^* = \arg\max_{\mathbf{z}} \; P_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{y})$$
$$\hat{\mathbf{z}} = \arg\max_{\mathbf{z}} \; P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{z}|\mathbf{x})$$

[3] For efficiency, instead of regularizing paramters after each example, we do so only once every 100 examples, in the prox-grad style (Martins et al. 2011).

Algorithm 1 : Computing $\mathbf{z}^*$

```
Input: $R$: number of "rounds", $\mathbf{p}_{kb}$: active KB labels, $\hat{\mathbf{z}}$:
    predicted mention labels
 1: $\mathbf{z}^* \leftarrow \hat{\mathbf{z}}$
 2: for $p$ in $\mathbf{p}_{kb}$ do
 3:     $\mathbf{x}^p \leftarrow$ sortBy($\mathbf{x}, \Phi^{ft}_{\boldsymbol{\theta}}(\mathbf{x}_j, p)$)
 4: end for
 5: for 1 to $R$ do
 6:     for $p$ in $\mathbf{p}_{kb}$ do
 7:         $\mathbf{x}_j \leftarrow$ getTopNotAssigned($\mathbf{x}^p$)
 8:         $\mathbf{z}^*_j \leftarrow p$
 9:         markAsAssigned($\mathbf{x}_j$)
10:     end for
11: end for
12: return $\mathbf{z}^*$
```

$\hat{\mathbf{z}}$ can be computed precisely as described above.

Computing $\mathbf{z}^*$, however, is more difficult. Let $\mathbf{p}_{kb} = \{p : y^*_p = 1\}$ be the set of labels specified in the KB to the be true for instance i. In MultiR, Hoffmann et al. use Algorithm 1 with $R = 1$, which they motivate as a variant of the weighted edge cover problem. (This formalization of the algorithm with the R loop at line 5 is one of our contributions.) First, $\mathbf{z}^*$ is initialized to $\hat{\mathbf{z}}$. Then, for each label p that is active in the KB, the algorithm selects the not-previously-selected mention $\mathbf{x}_j$ which the model most confidently labeled with p and sets that mention's label z^*_j to p.

To implement the EALA rather than the EALO assumption, we use Algorithm 1 with $R = \lceil \alpha M / |\mathbf{p}_{kb}| \rceil$. This change allows each active label from the KB to "claim" *multiple* mentions until α percent of all the mentions have been claimed.

Thus, to learn model parameters $\boldsymbol{\theta}$, the system iterates through the training set, and for each instance i, it (i) instantiates $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, (ii) computes $\hat{\mathbf{z}}$, (iii) computes $\mathbf{z}^*$, and (iv) updates parameters using

$$\theta_k \leftarrow \theta_k + \eta_t(\phi^*_k - \hat{\phi}_k)$$
$$\phi^*_k = \sum_j \phi_k(\mathbf{x}_j, z^*_j)$$
$$\hat{\phi}_k = \sum_j \phi_k(\mathbf{x}_j, \hat{z}_j)$$

where $\eta_t = \eta_0 / \sqrt{t/N}$, i is the t-th example, and N is the total number of examples

3.3 Category Mention Features Under EALA

We describe the features we use in more detail in Section 4.3, but here we convey an important lesson regarding category recognition under the EALA assumption. Standard features for the task of recognizing category mentions fall into two categories: features of the noun phrase, and features of the context. Typical noun phrase features include the identity of each word in the noun phrase as well as prefixes and suffixes, and other features. Initially, we tried using such standard features in our EALA-based and EALO-based systems. However, we observed that the EALA approach's use of multiple mentions for each instance, *all of which have identical noun phrase features*, resulted in overly-inflated weights on those noun phrase features, in particular the word identities and affixes. Although it is likely that significantly more negative data could remedy this problem, time and resource constraints necessitated other solutions.

Next, although leaving out *all* noun phrase features led to the EALA approach out-performing the EALO approach, we found that basic word shape features on the noun phrases were required in order to achieve an acceptable overall level of performance. Hence, we found that in the face of non-infinite negative data, the optimal solution was to use only more general noun phrase features, such as POS and word shape.

3.4 Distributed Implementation

To facilitate subsequent scaling up of distant supervision both in terms of the number of predicates and the size of the corpus, our system is implemented in Hadoop[4]. Our system learns model parameters in parallel on different shards of the dataset and then averages them after each iteration (McDonald, Hall, and Mann 2010). In all of our experiments reported below, each system was trained for 10 iterations using 20 shards.

4 Evaluation

4.1 Corpus

For the unlabeled text corpus in our experiments we used a random sample of 25% of the documents in a dependency-parsed version of Wikipedia provided by the Hazy [5] research group. This corpus consists of 821,186 English web pages that were downloaded in November 2011. The preprocessing annotations provided in the Hazy collection include POS tags and NER labels generated using the Stanford CoreNLP tools [6] and dependency parse annotations generated using the Malt parser. [7]

4.2 Knowledge Base

For the KB in our experiments we used instances downloaded from Freebase in June 2013. To determine the set of categories to consider, we ranked the Freebase categories (entity types) in descending order by the number of instances (entities) they had which were mentioned in Wikipedia (using only canonical entity names). Then, going through this ranked list from top to bottom and looking at the list of mentioned instances for each category (along with each instance's mention count), we discarded those categories for which the majority of the mentions were very likely to refer to a different category. For example, the Freebase category FILM contains many instances, such as "9", whose mentions most likely do not refer to the FILM by that name. In our experiments, we used the top 50 categories that were not discarded by this procedure.

[4] http://hadoop.apache.org

[5] http://hazy.cs.wisc.edu/hazy/

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

[7] http://www.maltparser.org

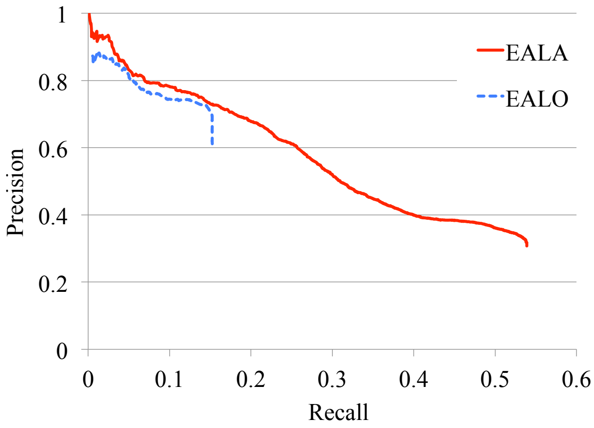


Figure 3: Precision-recall curves using the EALO and EALA assumptions on the category recognition task.

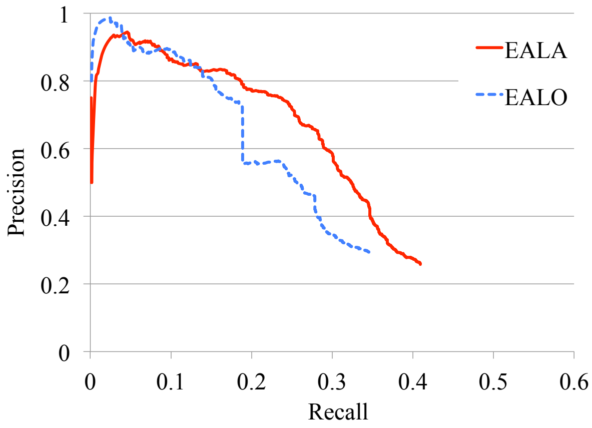


Figure 4: Precision-recall curves using the EALO and EALA assumptions on the relation recognition task.

To determine the list of relations, we followed a completely analogous procedure, with the stipulation that a relation instance mention occurs only if the canonical entity names of both arguments are found in the same sentence with no more than 6 tokens between them.[8] 40 relations were not discarded by the same procedure and therefore included in our experiments.

4.3 Features

Our system represents a category mention with the following features: (i) shortened phrase shape (capitalization pattern), e.g., `AxaxaxA` for "Call of the Wild," (ii) dependency paths of length one and two steps away from the noun phrase head, along with the token at the other end of the path, and (iii) trigrams from the following token sequence: up to three tokens on the left, a placeholder for the NP, up to three tokens on the right.

Our system represents relation mentions with the standard set of features for this task, which were originally defined by (Mintz et al. 2009).

4.4 Data Sets

To construct the data sets used in our experiments, we first computed features for the corpus mentions of all the instances in our KB. Any noun phrase from the corpus that was not a category instance in our KB became a negative example, i.e., an instance of the NONE category. For relation negative examples, i.e., instances of the NONE relation, we used pairs of KB category instances that were not instances of any relation in the KB. If an instance (category or relation, positive or negative) was mentioned more than 100 times in the corpus, we randomly selected only 100 of its mentions to use in our experiments. Thus, a learning example consists of an instance (a noun phrase or noun phrase pair), up to 100 of its mentions from the corpus, and one or more predicate labels: either from the KB, or NONE. In total, we used 1.3M category examples (83K/1.2M pos./neg.) containing 24.9M mentions (2.0M/22.9M pos./neg.) and 1.4M relation examples (27K/1.4M pos./neg.) containing 2.7M mentions (148K/2.5M pos./neg.). Finally, we randomly selected 10% of the examples for the test set and 10% for the development set, with the rest serving as the training set.

For efficiency during training, we used only a random sample of 50% of the negative training examples. The development and test sets, however, retained the original positive/negative ratio.

4.5 Experiments

The goal of our evaluation was to examine the effects on performance of using the EALA assumption when compared with using the EALO assumption for both the category and relation recognition tasks. To measure performance under the EALO assumption, we ran our system using Algorithm 1 with $R = 1$. This system configuration is essentially the same as the MultiR approach. Of course, for the EALA approach, we ran the system using Algorithm 1 with $R = \lceil \alpha M / |\mathbf{p}_{kb}| \rceil$.

First, we trained the EALA-based system on only the training set, and tuned α separately for categories and relations using the development set. The optimal α for categories and relations was 0.4 and 0.3, respectively. Then we combined the training and development sets, trained both systems (with the optimal α's for the EALA-based system), and evaluated their performance on the test set.

We evaluate our systems on the aggregate extraction task: after predicting the labels for each mention of an instance, we take the set of labels that were predicted for at least one mention and compare it with the set of KB labels associated with that instance. Performance is then measured in terms of precision and recall of these KB labels.

Figure 3 shows the precision of both systems, as a function of recall, on the category recognition task. Figure 4 shows the same thing for the relation recognition task.

As might be expected, the EALA assumption leads to higher recall over the EALO assumption because the system is able to learn from more training examples. Furthermore,

[8] We ignored all sentences containing more than 100 tokens.

as these results show, this boost in recall is present at comparable levels of precision. In fact, the largest improvements in recall at comparable precision, 66% increase for category recognition and 53% increase for relation recognition, are achieved at a precision of 0.61. Interestingly, the EALA assumption also achieves the highest levels of precision for category recognition but not for relation recognition.

5 Conclusion and Future Work

In summary, we have given clear evidence that the "expressed at least once" assumption, which is central to many distant supervision algorithms for information extraction, can actually be too *weak* in many cases. Our experimental results indicate that significantly higher recall can be obtained by assuming KB facts are expressed multiple times in the corpus, both for categories and for relations.

However, this study is only the first step in investigating this issue. In reality, the notion of α put forth in this paper, i.e., the percentage of corpus mentions that actually express particular meanings, is a fact-specific notion. It seems that the ideal approach would be to estimate a separate α for each KB fact and then use those estimates, which essentially are the relative frequencies of different word senses, to guide the learning procedure. However, from initial investigation in this direction, we have found that accurate estimates of word sense frequencies are very difficult to obtain, even from manually-labeled data. Hence, this line of research merits additional investigation.

6 Acknowledgments

This research has been supported in part by DARPA (under contract number FA8750-13-2-0005), Google, the National Science Foundation, under award CCF-1019104, and the Gordon and Betty Moore Foundation (eScience project). Any opinions, findings, conclusions and recommendations expressed in this paper are the authors and do not necessarily reflect those of the sponsors.

References

Banko, M.; Cafarella, M. J.; Soderland, S.; Broadhead, M.; and Etzioni, O. 2007. Open Information Extraction from the Web. In *Proceedings of IJCAI 2007*.

Brin, S. 1998. Extracting patterns and relations from world wide web. In *Proceedings of WebDB'98*, 172–183.

Bunescu, R., and Mooney, R. 2007. Learning to Extract Relations from the Web using Minimal Supervision. In *Proceedings of ACL 2007*, 576–583.

Carlson, A.; Betteridge, J.; Kisiel, B.; Settles, B.; Hruschka Jr, E.; and Mitchell, T. 2010. Toward an architecture for never-ending language learning. In *Proc. of AAAI 2010*.

Craven, M., and Kumlien, J. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of ISMB 1999*, 77–86.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Perez, T. 1997. Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence* 89:31–71.

Doddington, G.; Mitchell, A.; and Przybocki, M. 2004. The automatic content extraction (ACE) programtasks, data, and evaluation. In *Proceedings of LREC 2004*, 837–840.

Etzioni, O.; Cafarella, M.; Downey, D.; Popescu, A.; Shaked, T.; Soderland, S.; Weld, D.; and Yates, A. 2005. Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence* 165(1):91–134.

Grishman, R., and Sundheim, B. 1996. Message Understanding Conference-6: A brief history. In *Proc. of COLING 1996*, 466–471.

Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of COLING 1992*, 539–545.

Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; and Weld, D. S. 2011. Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations. In *Proceedings of ACL 2011*, 541–550.

Huang, R., and Riloff, E. 2010. Inducing domain-specific semantic class taggers from (almost) nothing. In *Proceedings of ACL 2010*, 275–285.

Martins, A. F. T.; Smith, N. A.; Aguiar, P. M. Q.; and Figueiredo, M. A. T. 2011. Structured sparsity in structured prediction. In *Proceedings of the EMNLP 2011*, 1500–1511.

McDonald, R.; Hall, K.; and Mann, G. 2010. Distributed training strategies for the structured perceptron. In *Proceedings of HLT-NAACL 2010*, 456–464.

Min, B.; Grishman, R.; Wan, L.; Wang, C.; and Gondek, D. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proc. of NAACL-HLT 2013*.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-AFNLP 2009*, 1003–1011.

Morgan, A. a.; Hirschman, L.; Colosimo, M.; Yeh, A. S.; and Colombe, J. B. 2004. Gene name identification and normalization using a model organism database. *Journal of biomedical informatics* 37(6):396–410.

Nakashole, N.; Theobald, M.; and Weikum, G. 2011. Scalable knowledge harvesting with high precision and high recall. In *Proceedings of WSDM 2011*, 227–236.

Riedel, S.; Yao, L.; Marlin, B. M.; and McCallum, A. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of HLT-NAACL 2013*.

Riedel, S.; Yao, L.; and McCallum, A. 2010. Modeling Relations and Their Mentions without Labeled Text. volume 6323 of *Lecture Notes in Computer Science*. 148–163.

Ritter, A.; Zettlemoyer, L.; Mausam; and Etzioni, O. 2013. Modeling missing data in distant supervision for information extraction. *Proceedings of TACL 2013*.

Surdeanu, M.; Tibshirani, J.; Nallapati, R.; and Manning, C. D. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of EMNLP-CoNLL 2012*.

Whitelaw, C.; Kehlenbeck, A.; Petrovic, N.; and Ungar, L. 2008. Web-scale named entity recognition. In *Proceedings of CIKM 2008*, 123–132.