Exploiting Parallel News Streams for Unsupervised Event Extraction

Congle Zhang, Stephen Soderland & Daniel S. Weld
Computer Science & Engineering
University of Washington
Seattle, WA 98195, USA
{clzhang, soderlan, weld}@cs.washington.edu

Abstract

Most approaches to *relation extraction*, the task of extracting ground facts from natural language text, are based on machine learning and thus starved by scarce training data. Manual annotation is too expensive to scale to a comprehensive set of relations. Distant supervision, which automatically creates training data, only works with relations that already populate a knowledge base (KB). Unfortunately, KBs such as FreeBase rarely cover event relations (*e.g. "person travels to location"*). Thus, the problem of extracting a wide range of events — e.g., from news streams — is an important, open challenge.

This paper introduces NEWSSPIKE-RE, a novel, unsupervised algorithm that discovers event relations and then learns to extract them. NEWSSPIKE-RE uses a novel probabilistic graphical model to cluster sentences describing similar events from parallel news streams. These clusters then comprise training data for the extractor. Our evaluation shows that NEWSSPIKE-RE generates high quality training sentences and learns extractors that perform much better than rival approaches, more than doubling the area under a precision-recall curve compared to Universal Schemas.

1 Introduction

Relation extraction, the process of extracting structured information from natural language text, grows increasingly important for Web search and question answering. Traditional supervised approaches, which can achieve high precision and recall, are limited by the cost of labeling training data and are unlikely to scale to the thousands of relations on the Web. Another approach, distant supervision (Craven and Kumlien, 1999; Wu and Weld, 2007), creates its own training data by matching the ground instances of a Knowledge base (KB) (*e.g.* Freebase) to the unlabeled text.

Unfortunately, while distant supervision can work well in some situations, the method is limited to relatively *static* facts (*e.g.*, *born-in(person, location)* or *capital-of(location,location)*) where there is a corresponding knowledge base. But what about dynamic *event relations* (also known as *fluents*), such as *travel-to(person, location)* or *fire(organization, person)*? Since these time-dependent facts are ephemeral, they are rarely stored in a pre-existing KB. At the same time, knowledge of real-time events is crucial for making informed decisions in fields like finance and politics. Indeed, news stories report events almost exclusively, so learning to extract events is an important open problem.

This paper develops a new unsupervised technique, NEWSSPIKE-RE, to both discover event relations and extract them with high precision. The intuition underlying NEWSSPIKE-RE is that the text of articles from two different news sources are not independent, since they are each conditioned on the same real-world events. By looking for rarely described entities that suddenly "spike" in popularity on a given date, one can identify paraphrases. Such *temporal correspondence* (Zhang and Weld, 2013) allow one to cluster diverse sentences, and the resulting clusters may be used to form training data in order to learn event extractors. Furthermore, one can also exploit parallel news to obtain direct *negative* evidence. To see this, suppose one day the news includes the following: (a) *"Snowden travels to Hong Kong, off southeastern China."* (b) *"Snowden cannot stay in Hong Kong as Chinese officials will not allow ..."* Since news stories are usually coherent, it is highly unlikely that *travel to* and *stay in* (which is negated) are synonymous. By leveraging such direct negative phrases, we can learn extractors capable of distinguishing heavily co-occurring but semantically different phrases, thereby avoiding many extraction errors. Our NEWSSPIKE-RE system encapuslates these intuitions in a novel graphical model making

Transactions of the Association for Computational Linguistics, vol. 3, pp. 117–129, 2015. Action Editor: Hal Daumé III.
Submission batch: 10/2014; Revision batch 1/2015; Published 2/2015.

the following contributions:

- We develop a method to discover a set of distinct, salient event relations from news streams.
- We describe an algorithm to exploit parallel news streams to cluster sentences that belong to the same event relations. In particular, we propose the *temporal negation heuristic* to avoid conflating co-occurring but non-synonymous phrases.
- We introduce a probabilistic graphical model to generate training for a sentential event extractor without requiring any human annotations.
- We present detailed experiments demonstrating that the event extractors, learned from the generated training data, significantly outperform several competitive baselines, *e.g.* our system more than doubles the area under the micro-averaged, PR curve (0.80 vs. 0.30) compared to Riedel's Universal Schema (Riedel et al., 2013).

2 Previous Work

Supervised learning approaches have been widely developed for event extraction tasks such as MUC-4 and ACE. They often focus on a hand-crafted ontology and train the extractor with manually created training data. While they can offer high precision and recall, they are often domain-specific (*e.g.* biological events (Riedel et al., 2011; McClosky et al., 2011) and entertainment events (Benson et al., 2011; Reichart and Barzilay, 2012)), and are hard to scale over the events on the Web.

Open IE systems extract open domain relations (*e.g.* (Banko et al., 2007; Fader et al., 2011)) and events (*e.g.* (Ritter et al., 2012)). They often perform self-supervised learning of relation-independent extractions. It allows them to scale but makes them unable to output canonicalized relations.

Distant supervised approaches have been developed to learn extractors by exploiting the facts existing in a knowledge base, thus avoiding human annotation. Wu *et al.* (2007) and Reschke *et al.* (2014) learned Infobox relations from Wikipedia, while Mintz *et al.* (2009) heuristically matched Freebase facts to texts. Since the training data generated by the heuristic matching is often imperfect, multi-instance learning approaches (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012) have been developed to combat this problem. Unfortunately, most facts existing in the KBs are static facts like geographical or biographical data. They fall short of learning extractors for fluent facts such as sports results or travel and meetings by a person.

Bootstrapping is another common extraction technique (Brin, 1999; Agichtein and Gravano, 2000; Carlson et al., 2010; Nakashole et al., 2011; Huang and Riloff, 2013). This typically takes a set of seeds as input, which can be ground instances or key phrases. The algorithms then iteratively generate more positive instances and phrases. While there are many successful examples of bootstrapping, the challenge is to avoid semantic drift. Large-scale systems, therefore, often require extra processing such as manual validation between the iterations or additional negative seeds as the input.

Unsupervised approaches have been developed for relation discovery and extractions. These algorithms are usually based on some clustering assumptions over a large unlabeled corpus. Common assumptions include the distributional hypothesis used by (Hasegawa et al., 2004; Shinyama and Sekine, 2006), latent topic assumption by (Yao et al., 2012; Yao et al., 2011), and low rank assumption by (Takamatsu et al., 2011; Riedel et al., 2013). Since the assumptions largely rely on co-occurrence, previous unsupervised approaches tend to confuse correlated but semantically different phrases during extraction. In contrast to this, our work largely avoids these errors by exploiting the temporal negation heuristic in parallel news streams. In addition, unlike many unsupervised algorithms requiring human effort to canonicalize the clusters, our work automatically discovers events with readable names.

Paraphrasing techniques inspire our work. Some techniques, such as DIRT (Lin and Pantel, 2001) and Resolver (Yates and Etzioni, 2009), are based on the distributional hypothesis. Another common approach is to use parallel corpora, including news streams (Barzilay and Lee, 2003; Dolan et al., 2004; Zhang and Weld, 2013), multiple translations of the same story (Barzilay and McKeown, 2001) and bilingual sentence pairs (Ganitkevitch et al., 2013) to generate the paraphrases. Although these algorithms create many good paraphrases, they can not be directly used to generate enough training data to train a relation extractor for two reasons: first, the semantics of the paraphrases is often context dependent; second, the generated paraphrases are often in

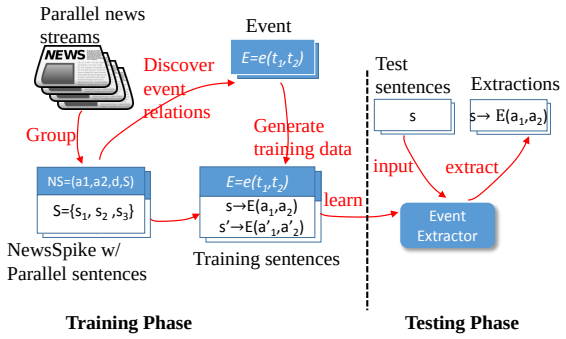


Figure 1: During its training phase, NEWSSPIKE-RE first groups parallel sentences as *NewsSpikes*. Next, the system automatically discovers a set of event relations. Then, a probabilistic graphical model clusters sentences from the NewsSpike as training data for each discovered relation, which is used to learn sentential event extractors. During the testing phase, the extractor takes test sentences as input and predicts event extractions.

small clusters and it remains challenging to merge them for the purpose of training an extractor. Our work extends previous paraphrasing techniques, notably that of Zhang and Weld (2013), but we focus on generating high-quality, positive and negative training sentences for the discovered events in order to learn extractors with high precision and recall.

3 System Overview

News articles report an enormous number of events every day. Our system, NEWSSPIKE-RE, aligns paralel news streams to indentify and extract these events as shown in Figure 1. NEWSSPIKE-RE has both training and test phases. Its training phase has two main steps: event-relation discovery and training-set generation. Section 4 describes our event relation discovery algorithm, which processes time-stamped news articles to discern a set of salient, distinct event relations in the form of $E = e(t_1, t_2)$, where e is a representative *event phrase* and t_i are types of the two arguments. NEWSSPIKE-RE generates the event phrases using an Open Information Extraction (IE) system (Fader et al., 2011), and uses a fine-grained entity recognition system FIGER (Ling and Weld, 2012) to generate type descriptors such as "company ", "politician", and "medical treatment".

The second part of NEWSSPIKE-RE's training phase, described in Section 5, is a method for building extractors for the discovered event relations. Our approach is motivated by the intuition, adapted from Zhang and Weld (2013), that articles from different news sources typically use different sentences to describe the same event, and that corresponding sentences can be identified when they mention a unique pair of real-world entities. For example, when an unusual entity pair (*Selena*, *Norway*) is suddenly seen in three articles on a single day:

> *Selena traveled to Norway to see her ex-boyfriend.*
> *Selena arrived in Norway for a rendezvous with Justin.*
> *Selena's trip to Norway was no coincidence.*

It is likely that all three refer to the same event relation, *travel-to(person, location)*[1], and can be used as positive training examples for the relation. As in Zhang & Weld (2013), we group *parallel sentences* sharing the same argument pair and date in a structure called a *NewsSpike*. However, we include all sentences mentioning the arguments (*e.g. Selena's trip to Norway*) in the NewsSpike (not just those yielding OpenIE extractions), and use the lexicalized dependency path between the arguments (*e.g.* <-[poss]-trip-[prep-to]->[2], as the event phrase. In this way, we can generalize extractors beyond the scope of OpenIE. Formally, a NewsSpike is a tuple, (a_1, a_2, d, S), where a_1 and a_2 are arguments (*e.g. Selena*), d is a date, and S is a set of argument-labeled sentences $\{(s, a_1, a_2, p) \ldots\}$ in which s is a sentence with arguments a_i and event phrase p.

It's important that non-synonomous sentences like *"Selena stays in Norway"* should be excluded from the training data for *travel-to(person, location)* even if a *travel-to* event did apply to that argument pair. In order to select only the synonomous sentences, we develop a probabilistic graphical model, described in Section 5.2, to accurately assign sentences from NewsSpikes to each discovered event relation E. Given this annotated data, NEWSSPIKE-RE trains extractors using a multiclass logistic regression classfier.

During the testing phase, NEWSSPIKE-RE accepts arbitrary sentences (no date-stamp required), uses FIGER to identify possible arguments, and uses the classifier to predicts which events (if any) hold between an argument pair. We describe the extraction process in Section 6.

Note that NEWSSPIKE-RE is an unsupervised al-

[1]For clarity in the paper, we refer to this relation as *travel-to*, even though the phrase *arrive in* is actually more frequent and is selected as the name of this relation by our event discovery algorithm, as shown in Table 2.

[2]This dependency path will be referred to as "'s trip to".

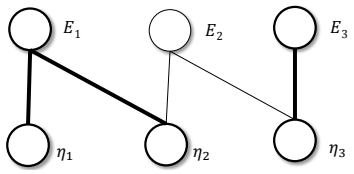


Figure 2: A simple example of the edge-cover algorithm with K=2, where E_i are event relations and η_j are NewsSpikes. The optimal solution selects E_1 with edges to η_1 and η_2, and E_3 with edge to η_3. These two event relations cover all the NewsSpikes.

gorithm that requires no manual labelling of the training instances. Like distant supervision, the key is to automatically generate the training data, at which point a traditional supervised classifier may be applied to learn an extractor. Because distant supervision creates very noisy annotations, researchers often use specialized learners that model the correctness of a training example with a latent variable (Riedel et al., 2010; Hoffmann et al., 2011), but we found this unnecessary, because NEWSSPIKE-RE creates high quality training data.

4 Discovering Salient Events

The first step of NEWSSPIKE-RE is to discover a set of event relations in the form of $E = e(t_1, t_2)$, where e is an event phrase, and t_i are fine-grained argument types generated by FIGER, augmented with the important types "number" and "money", which are recognized by the Stanford name entity recognition system (Finkel et al., 2005). To be most useful, the discovered event relations should cover salient events that are frequently reported in the news articles. Formally, we say that a NewsSpike $\eta = (a_1, a_2, d, S)$ mentions $E = e(t_1, t_2)$ if the types of a_i are t_i for each i, and one of its sentence has e as the event phrase between the arguments. To maximize the salience of the events, NEWSSPIKE-RE will prefer event relations that are *"mentioned"* by more NewsSpikes.

In addition, the set of event relations should be distict. For example, if the relation *travel-to(person, location)* is already in the set, then *visit(person, location)* should not be selected as a separate relation. To reduce overlap, discovered event relations should not be mentioned by the same NewsSpike.

Let $\mathcal{E}$ be all candidate event relations, $\mathcal{N}$ be all NewsSpikes. Our goal is to select the K most salient relations from $\mathcal{E}$, minimizing overlap between relations. We can frame this task as a variant of the bipartite graph edge-cover problem. Let a bipartite graph G have one node E_i for each event relation in $\mathcal{E}$ and one node η_j for each NewsSpike in $\mathcal{N}$. There is an edge between E_i and η_j if η_j mentions E_i. The edge-cover problem is to select a largest subset of edges subject to (1) at most K nodes of E_i are chosen and all edges incident to them are chosen as the covered edges; (2) each node of η_j is incident to at most one edge. The first constraint guarantees that there are exactly K event relations discovered; the second constraint ensures that no NewsSpike participates in two event relations. Figure 2 shows the optimized solution of a simple graph with $K = 2$, which can cover 3 edges with 2 event relations that have no overlapping NewsSpikes.

Since both the objective function and constraints are linear, we can optimize this edge-cover problem with integer linear programming (Nemhauser and Wolsey, 1988). By solving the optimization problem, NEWSSPIKE-RE finds a salient set of event relations incident to the covered edges. The discovered relations with K set to 30 are shown in Table 2 in Section 7. In addition, the covered edges bring us the initial mapping between the event types and NewsSpikes, which is used to train the probablistic model in Section 5.3.

5 Generating the Training Sentences

After NEWSSPIKE-RE has discovered a set of event relations, it then generates training instances to learn an extractor for each relation. In this section, we present our algorithm for generating the training sentences. As shown in Figure 1, the generator takes N NewsSpikes $\{\eta_i = (a_{1i}, a_{2i}, d_i, S_i) | i = 1 \dots N\}$ and K event relations $\{E_k = e_k(t_{1k}, t_{2k}) | k = 1 \dots K\}$ as input. For every event relation, E_k, the generator identifies a subset of sentences from $\cup_{i=1}^{N} S_i$ expressing the event relation as training sentences. In this section, we first characterize the paraphrased event phrases and the parallel sentences in NewsSpikes. Then we show how to encode this heuristic in a probabilistic graphical model that jointly paraphrases the event phrases and identifies a set of training sentences.

5.1 Exploiting Properties of Parallel News

Previous work (Zhang and Weld, 2013) proposed several heuristics that are useful to find similar sentences in a NewsSpike. For example, the temporal functionality heuristic says that sentences in a

NewsSpike with the same tense tend to be paraphrases. Unfortunately, these methods are too weak to generate enough data for training high quality event extractors: (1) they are "in-spike heuristics" that tend to generate small clusters from individual NewsSpikes. It remains unclear how to merge similar events occuring on different days and between different entities to increase cluster size. (2) they included heuristics to "gain precision at the expense of recall" (*e.g.* news articles do not state the same fact twice), because it is hard to obtain direct negative phrases inside one NewsSpike. In this paper, we exploit news streams in a cross-spike, global manner to obtain accurate positive and negative signals. This allows us to dramatically improve recall while maintaining high precision.

Our system starts from the basic observation that the parallel sentences tend to be coherent. So if a NewsSpike $\eta = (a_1, a_2, d, S)$ is an instance of an event relation $E = e(t_1, t_2)$, the event phrases in its parallel sentences tend to be paraphrases. But sometimes the sentences in the NewsSpike are related but not paraphrases. For example, one day "Snowden will stay in Hong Kong ..." appears together with "Snowden travels to Hong Kong ...". Although the fact *stay-in(Snowden, Hong Kong)* is true, it is harmful to include "Snowden will stay in Hong Kong" in the training for *travel-to(person, location)*.

Detecting paraphrases remains a challenge to most unsupervised approaches because they tend to cluster heavily co-occurring phrases which may turn out to be semantically different or even antonymous. (Zhang and Weld, 2013) presented a method to avoid confusion between antonym and synonyms in NewsSpikes, but did not address the problem of related but different phrases like *travel to* and *stay in* in a NewsSpike.

To handle this, our method rests on a simple observation: when you read "Snowden travels to Hong Kong" and "Snowden cannot stay in Hong Kong as Chinese officials do not allow ..." in the same NewsSpike, it is unlike that *travel to* and *stay in* are synonymous event phrases because otherwise the two news stories are describing the opposite event. This observation leads to:

Temporal Negation Heuristic. *Two event phrases p and q tend to be semantically different if they co-occur in the NewsSpike but one of them is in negated form.*

The temporal negation heuristic helps in two ways: (1) it provides some direct negative phrases for the event relations; NEWSSPIKE-RE uses these to heuristically label some variables in the model. (2) It creates some useful features to implement a form of transitvity. For example, if we find that *live in* and *stay in* are frequently co-occurring and the temporal negation heuristic tells us that *travel to* and *stay in* are not paraphrases, this is evidence that *live in* is unlikely to be a paraphrase of *travel to*, even if they are heavily co-occurring.

The following section describes our implementation that uses these properties to generate high quality training. Our goal is the following: a sentence (s, a_1, a_2, p) from NewsSpike $\eta = (a_1, a_2, d, S)$ should be included in the training data for event relation $E = e(t_1, t_2)$ if the event phrase p is a paraphrase of e and the event relation E happens to the argument pair (a_1, a_2) at time d.

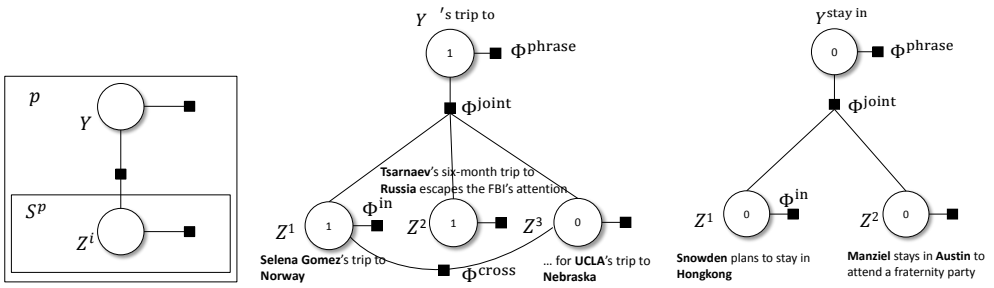
5.2 Joint Cluster Model

As discussed above, to identify a high quality set of training sentences from NewsSpikes, one needs to combine evidence that event phrases are paraphrases with evidence from NewsSpikes. For this purpose, we define an undirected graphical model to jointly reason about paraphrasing the event phrases and identifying the training sentences from NewsSpikes. We first list the notation used in this section:

E	event relation
$p \in P$	event phrases
$s \in S^p$	sentences w/ the event phrase p
Y^p	Is p a paraphrase for E?
Z_p^s	Is s w/ p good training for E?
Φ	factors

Let P be the union of all the event phrases from every NewsSpike. For each $p \in P$, let S^p be the set of sentences having p as its event phrase.

Figure 3(a) shows the model in plate form. There are two kinds of random variables corresponding to phrases and sentences, respectively. For each event relation $E = e(t_1, t_2)$, there exists a connected component for every event phrase $p \in P$ that models (1) whether p is a paraphrase of e or not (modeled using Boolean phrase variables, Y^p); and (2) whether each sentence of S^p is a good training sentence for E (modeled using $|S^p|$ Boolean sentence variables $\{Z_p^s | s \in S^p\}$. Intuitively, the goal of the model is to find the set of good training sentences, with



(a) (b) (c)

Figure 3: (a) The connected components depicted as plate model, where each Y is a Boolean variable for a relation phrase and each Z is a Boolean variable for a training sentence for with that phrase; (b) and (c) are example connected components for the event phrases *'s trip to* and *stay in* respectively. The goal of the model is to set $Y = 1$ for good paraphrases of a relation and to set $Z = 1$ for good training sentences.

$Z_p^s = 1$. The union of such sentences over the different phrases, $\cup_p\{s|Z_p^s = 1\}$, defines the training sentences for the event. Figure 3(b) and 3(c) show two example connected-components for the event phrases *'s trip to* and *stay in* respectively.

Now, we can define the joint distribution over the event phrases and the sentences. The joint distribution is a function defined on factors that encode our observations about NewsSpikes as features and constraints. The *phrase factor* Φ^{phrase} is a log-linear function attaching to Y^p with the paraphrasing features, such as whether p and e co-occur in the NewsSpikes, or whether p shares the same head word with e. They are used to distinguish whether p is a good event phrase.

A sentence should not be identified as a good training sentence if it does not contain a positive event phrase. For example, if $Y^{\text{stay in}}$ in Figure 3(b) takes the value of 0, thus all sentences with the event phrase *stay in* should also take the value of 0. We implement this constraint with a *joint factor* Φ^{joint} among Y^p and Z_p^s variable.

In addition, good training sentences occur when the NewsSpike is an event instance. To encode this observation, we need to featurize the NewsSpikes and let them bias the assignments. Our model implements this with two types of log-linear factors: (1) the unary *in-spike factor* Φ^{in} depends on the sentence variables and contains features about the corresponding NewsSpike. The factor is used to distinguish whether the NewsSpike is an instance of $e(t_1, t_2)$, such as whether the argument types of the NewsSpike match the designated types t_1, t_2; (2) the pairwise *cross-spike factors* Φ^{cross} connect pairs of sentences. This uses features such as whether the pair of NewsSpikes for the two sentences have high textual similarity, and whether two NewsSpikes contain negated event phrases.

We define the joint distribution for the connected component for p as follows. Let $\mathbf{Z}$ be the vector of sentence variables, let $\mathbf{x}$ be the features. The joint distribution is:

$$p(Y = y, \mathbf{Z} = \mathbf{z}|\mathbf{x}; \Theta) \stackrel{\text{def}}{=} \frac{1}{Z_x}\Phi^{\text{phrase}}(y, \mathbf{x}) \times \Phi^{\text{joint}}(y, \mathbf{z}) \prod_s \Phi^{\text{in}}(z^s, \mathbf{x}) \prod_{s,s'} \Phi^{\text{cross}}(z^s, z^{s'}, \mathbf{x})$$

where the parameter vector Θ is the weight vector of the features in Φ^{in} and Φ^{cross}, which are log-linear functions. The joint factors Φ^{joint} is zero when $Y^p = 0$ but some $Z_p^s = 1$. Otherwise, it is set to 1. We use integer linear programming to perform MAP inference on the model, finding the predictions $y, \mathbf{z}$ that maximize the probability.

5.3 Learning from Heuristic Labels

We now present the learning algorithm for our joint cluster model. The goal of the learning algorithm is to set Θ for the log-linear functions in the factors in a way that maximizes the likelihood estimation. We do this in a totally unsupervised manner, since manual annotation is expensive and not scalable to large numbers of event relations.

The weights are learned in three steps: (1) NEWSSPIKE-RE creates a set of heuristic labels for a subset of variables in the graphical model; (2) it uses the heuristic labels as supervision for the model; (3) it updates Θ with the perceptron learning algorithm. The weights are used to infer the values of the variables that don't have heuristic labels. The procedure is summarized in Figure 4.

For each event relation $E = e(t_1, t_2)$, NEWSSPIKE-RE creates heuristic labels as follows:

Input: NewsSpikes and the connected components of the model;

Heuristic Labels:

1. find positive and negative phrases and sentences P^+, P^-, S^+, S^-;
2. label the connected componenets accordingly and create $\{(Y_i^{\text{label}}, \mathbf{Z}_i^{\text{label}}) \mid_{i=1}^{M}\}$.

Learning: Update Θ with the perceptron learning algorithm.

Output: the values of all variables in the connected components with the MAP inference.

Figure 4: Learning from Heuristic Labels

(1) P^+: the temporal functionality heuristic (Zhang and Weld, 2013) says that if an event phrase p co-occurs with e in the NewsSpikes, it tends to be a paraphrase of e. We add the most frequently co-occurring event phrases to P^+. P^+ also includes e itself. (2) P^-: the temporal negation heuristic says that if p and e co-occur in the NewsSpike but one of them is in its negated form, p should be negatively labeled. We add those event phrases to P^-. If a phrase p appears in both P^+ and P^-, we remove it from both sets. (3) S^+: we first get the positive NewsSpikes from the solution of the edge-cover problem in section 4. We treat the NewsSpike η as positive if the edge between η and E is covered. Next, every sentence with $p \in P^+$ is added into S^+. (4) S^-: since the event relations discovered in section 4 tend to be distinct relations, a sentence is treated as negative sentence for E if it is heuristically labeled as positive for $E' \neq E$. In addition, S^- includes all sentences with $p \in P^-$.

With P^+, P^-, S^+, S^-, we define the heuristic labeled set to be $\{(Y_i^{\text{label}}, \mathbf{Z}_i^{\text{label}}) \mid_{i=1}^{M}\}$, where M is the number of the connected components with the corresponding event phrases $p \in P^+ \cup P^-$; $Y_i^{\text{label}} = 1$ if $p \in P^+$ and $Y_i^{\text{label}} = 0$ if $p \in P^-$. $\mathbf{Z}^i$ is labeled similarly, but note that if the sentence in the connected component doesn't exist in $S^+ \cup S^-$, NEWSSPIKE-RE doesn't include the corresponding variable in $\mathbf{Z}_i^{\text{label}}$. With $\{(Y_i^{\text{label}}, \mathbf{Z}_i^{\text{label}}) \mid_{i=1}^{M}\}$, learning can be done with maximum likelihood estimation as $L(\Theta) = \log \prod_i p(Y_i = y_i^{\text{label}}, \mathbf{Z}_i = \mathbf{z}_i^{\text{label}} \mid \mathbf{x}_i, \Theta)$. Following (Collins, 2002), we use a fast perceptron learning approach to update Θ. It consists of iterating two steps: (1) MAP inference given the current weight; (2) penalizing the weights if the inferred assignments are different from the heuristic labeled assignments.

6 Sentential Event Extraction

As shown in Figure 1, we learn the extractors from the generated training sentences. Note that most distant supervised (Hoffmann et al., 2011; Surdeanu et al., 2012) approaches use multi-instance, aggregate-level training (*i.e.* the supervision comes from labeled sets of instances instead of individually labeled sentences). Coping with the noise inherent in these multi-instance bags remains a big challenge for distant supervision. In contrast, our sentence-level training data is more direct and minimizes noise. Therefore, we implement the event extractor as a simple multi-class, L2-regularized logistic regression classifier.

For features of the classifier, we use the lexicalized dependency paths, the OpenIE phrases, the minimal subtree of the dependency parse and the bag-of-words between the arguments. We also augment them with fine grained argument types produced by FIGER (Ling and Weld, 2012). The event extractor that is learned can take individual test sentences (s, a_1, a_2) as input and predict whether that sentence expresses the event between (a_1, a_2).

7 Empirical Evaluation

Our evaluation addresses two questions. Section 7.2 considers whether our training generation algorithm identifies accurate and diverse sentences. Then, Section 7.3 investigates whether the event extractor, learned from the training sentences, outperforms other extraction approaches.

7.1 Experimental Setup

We follow the procedure described in (Zhang and Weld, 2013) to collect parallel news streams and generate the NewsSpikes: first, we get news seeds and query the Bing newswire search engine to gather additional, time-stamped, news articles on a similar topic; next, we extract OpenIE tuples from the news articles and group the sentences that share the same arguments and date into NewsSpikes. We collected the news stream corpus from March 1st 2013 to July 1st 2014. We split the dataset into two parts: in the training phrase, we use the news streams in 2013 (named NS13) to generate the training sentences. NS13 has 33k NewsSpikes containing 173k sentences.

We evaluated the extraction performance on news articles collected in 2014 (named NS14). In this

way, we make sure the test sentences are unseen during training. There are 15 million sentences in NS14. We randomly sample 100k unique sentences having two different arguments recognized by the name entity recognition system.

For our event discovery algorithm, we set the number of event relations to be 30 and ran the algorithm on NS13. The algorithm takes 6 seconds to run on a 2.3GHz CPU. Note that most previous unsupervised relation discovery algorithms require additional manual post-processing to assign names to the output clusters. In contrast, NEWSSPIKE-RE discovers the event relations fully automatically and the output is self-explanatory. We list them together with the by-event extraction performance in Table 2. From the table, we can see that most of the discovered event relations are salient with little overlap between relations.

While we arbitrarily set K to 30 in our experiments, there is no inherent limit to the number of relation phrases as long as the news corpus provides sufficient support to learn an extractor for each relation. In future, we plan to explore much larger sets of event relations to see if the extraction accuracy is maintained.

The joint cluster model that identifies training sentences for each event relation $E = e(t_1, t_2)$ uses cosine similarity between the event phrase p of a sentence and the canonical phrases of each relation as features in the phrase factors in Figure 3(a). It also includes the cosine similarity between p and a set of "anti-phrases" for the event relation which are recognized by the temporal negation heuristic.

For the in-spike factor, we measure whether the fine-grained argument types of the sentence returned from the FIGER system matches the required t_i respectively. In addition, we implement the features from (Zhang and Weld, 2013) to measure whether the sentence is describing the event of the NewsSpike. For the cross-spike factors, we use textual similarity features between the two sets of parallel sentences to measure the distance between the pair of NewsSpikes.

7.2 Quality of the Generated Training Set

The key to a good learning system is a high-quality training set. In this section, we compare our joint model against pipeline systems that consider paraphrases and argument type matching sequentially, based on the following paraphrasing techniques.

system	all			diverse		
	#	mi.	ma.	#	mi.	ma.
Basic	43,718	.50	.62	12,701	.38	.51
Yates09	15,212	.78	.76	586	.48	.50
Ganit13	14,420	.74	.71	1,210	.53	.53
Zhang13	14,804	.76	.75	890	.63	.61
NEWSSPIKE-RE	20,105	**.88**	**.89**	2,156	**.71**	**.72**
w/o cross	16,463	.86	.86	1,883	.67	.69
w/o neg	33,548	.76	.81	4,019	.64	.68

Table 1: Quality of the generated training sentences (count, micro- and macro- accuracy), where "all" includes sentences with all event phrases and "diverse" are those with distinct event phrases.

Basic is based on the temporal functionality heuristic of (Zhang and Weld, 2013). It treats all event phrases appearing in the same NewsSpike as paraphrases. **Yates09** uses Resolver (Yates and Etzioni, 2009) to create clusters of phrases. Resolver measures the similarity between the phrases by means of both distributional features and textual features. We convert the sentences in NewsSpikes into tuples in the form of (a_1, p, a_2), and run Resolver on these tuples to generate the paraphrases. **Zhang13**: We used the generated paraphrase set from (Zhang and Weld, 2013). **Ganit13**: Ganitkevitch *et al.* (2013) released a large paraphrase database (PPDB) based on exploiting the bilingual parallel corpora. Note that some of these paraphrasing systems do not handle dependency paths. So when p is a dependency path, we use the surface string between the arguments as the phrase. **NewsSpike-RE**: We also conduct ablation testing on NEWSSPIKE-RE to measure the effect of the cross-spike factors and the temporal negation heuristic: **w/o Cross** uses a simpler model by removing the cross-spike factors of NEWSSPIKE-RE; **w/o Negation** uses the same joint cluster model as NEWSSPIKE-RE but removes the features and the heuristic labels coming from the temporal negation heuristic.

We measured the micro- and macro- accuracy of each system by manually labeling 1000 randomly chosen output from each system[3]. Annotators read each training sentence, and decided if it was a good example for a particular event. We also report the number of generated sentences. Since the extractor should generalize over sentences with dissimilar expressions, it is crucial to identify sentences with

[3]Two Odesk workers were asked to label the dataset, a graduate student then reconciled any disagreements.

Event	F1 @ max recall				area u/ PR curve			area u/ diverse PR curve			
	#	R13	R13P	N-RE	R13	R13P	N-RE	#	R13	R13P	N-RE
acquire(organization,person)	59	0.34	0.33	**0.58**	0.26	0.26	**0.57**	20	0.26	0.17	**0.58**
arrive in(organization,location)	95	0.11	0.40	**0.56**	0.01	0.12	**0.42**	18	0.01	0.02	**0.50**
arrive in(person,location)	130	0.61	**0.86**	0.86	0.35	0.67	**0.93**	18	0.26	0.33	**0.80**
beat(organization,organization)	178	0.42	0.85	**0.90**	0.14	0.64	**0.84**	24	0.06	0.53	**0.58**
beat(person,person)	107	0.57	0.82	**0.94**	0.21	0.53	**0.91**	14	0.08	0.25	**0.77**
buy(organization,organization)	84	0.47	0.47	**0.78**	0.25	0.50	**0.82**	34	0.19	0.40	**0.79**
defend(person,person)	41	0.37	0.38	**0.52**	0.36	0.47	**0.65**	12	0.13	0.06	**0.47**
die at(person,number)	158	0.53	0.97	**0.98**	0.31	0.93	**0.97**	17	0.33	0.83	**0.94**
die(person,time)	179	0.85	0.91	**0.97**	0.66	0.80	**0.96**	16	0.22	0.63	**0.87**
fire(organization,person)	39	0.36	0.33	**0.53**	0.32	0.45	**0.88**	8	0.20	0.10	**0.66**
hit(event,location)	33	0.00	0.42	**0.64**	0.00	**0.51**	0.48	24	0.00	0.45	**0.50**
lead(person,organization/sports_team)	119	0.77	0.86	**0.87**	0.57	0.73	**0.77**	14	0.30	0.36	**0.62**
leave(person,organization)	61	0.40	0.52	**0.59**	0.14	0.38	**0.57**	14	0.07	0.13	**0.38**
meet with(person,person)	137	0.74	0.86	**0.92**	0.48	0.73	**0.88**	14	0.28	0.56	**0.93**
nominate(person/politician,person)	44	0.12	0.38	**0.54**	0.13	0.44	**0.77**	27	0.11	0.53	**0.75**
pay(organization,money)	134	0.77	0.91	**0.93**	0.52	0.85	**0.90**	17	0.33	**0.90**	0.56
place(organization,person)	34	0.17	0.28	**0.50**	0.24	0.23	**0.95**	16	0.19	0.21	**0.94**
play(person/artist,person)	173	**0.92**	0.89	0.87	**0.88**	0.79	0.73	15	**0.63**	0.56	0.47
release(organization,person)	30	0.18	0.22	**0.60**	0.08	0.25	**0.72**	16	0.06	0.15	**0.81**
replace(person,person)	115	0.82	0.89	**0.94**	0.62	0.75	**0.87**	18	0.46	0.58	**0.89**
report(government_agency,time)	140	0.37	0.84	**0.91**	0.09	0.74	**0.83**	35	0.06	0.52	**0.70**
report(written_work,time)	130	0.64	**0.85**	0.83	0.43	**0.82**	0.74	22	0.38	**0.58**	0.51
return to(person/athlete,location)	45	0.14	0.34	**0.50**	0.03	0.30	**0.49**	21	0.08	0.23	**0.78**
shoot(person,number)	101	0.71	0.89	**0.92**	0.49	0.74	**0.84**	8	0.35	0.37	**0.48**
sign with(person,organization)	129	0.47	0.62	**0.89**	0.25	0.46	**0.85**	44	0.15	0.17	**0.91**
sign(organization,person)	110	0.45	0.71	**0.85**	0.26	0.63	**0.79**	26	0.15	0.27	**0.66**
unveil(organization,product)	88	0.43	**0.71**	0.44	0.26	**0.52**	0.30	22	0.31	0.22	**0.63**
vote(government,time)	32	0.29	0.24	**0.74**	0.32	0.25	**0.77**	19	0.35	0.22	**0.83**
win at(person,location)	100	0.24	0.68	**0.85**	0.08	0.60	**0.90**	40	0.01	0.42	**0.90**
win(person,event)	107	0.54	0.77	**0.86**	0.22	0.63	**0.77**	19	0.03	0.26	**0.78**
micro average	2,903	0.53	0.70	**0.81**	0.30	0.59	**0.80**	609	0.15	0.31	**0.71**
macro average	97	0.46	0.64	**0.76**	0.30	0.56	**0.76**	20	0.20	0.37	**0.70**

Table 2: Performance of extractors by event relation, reporting both precision and the area under the PR curve. The # column shows the number of true extractions in the pool of sampled output. NEWSSPIKE-RE (labeled N-RE) outperforms two implementations of Riedel's Universal Schemas (See Section 7.3 for details). The advantage of NEWSSPIKE-RE over Universal Schemas is greatest on a diverse test set where each sentence has a distinct event phrase.

diverse event phrases. Therefore we also measured the accuracy and the count of a "diverse" condition: only consider the subset of sentences with distinct event phrases.

Table 1 shows the accuracy and the number of training examples. The basic temporal system brings us 0.50/0.62 micro- and macro- accuracy overall and 0.38/0.51 in the diverse condition. It shows that NewsSpikes are promising resources to generate the training set, but that elaboration is necessary. Yates09 gets 0.78/0.76 accuracy overall because its textual features help it to recognize many good sentences with similar phrases. But for the diverse condition, it gets lower precision because the distributional hypothesis fails to distinguish those correlated but different phrases.

Although Ganitkevitch13 and Zhang13 leverage existing paraphrase databases, it is interesting that their accuracy is still not good. It is largely because many times the paraphrasing must depend on the context: *e.g. "Cutler hits Martellus Bennett with TD in closing seconds."* is not good for the *beat(team, team)* relation, even though *hit* is a synonym for *beat* in general. These two systems show that it is not enough to use an off-the-shelf paraphrasing database for extraction.

The ablation test shows the effectiveness of the temporal negation hypothesis: after turning off the relevant features and heuristic labels, the precision drops about 10 percentage points. In addition, the cross-spike factors bring NEWSSPIKE-RE about 22% more training sentences and also increase the accuracy.

We did bootstrap sampling to test the statistical significance of NEWSSPIKE-RE's improvement in accuracy over each comparison system and ablation of NEWSSPIKE-RE. For each system we computed the accuracy of 10 samples of 100 labeled outputs. We then ran the paired t-test over the accuracy numbers of each other system compared to

NEWSSPIKE-RE. For all but w/o cross the improvement is strongly significant with p-value less than 1%. The increase in accuracy compared to w/o cross has borderline significance (p-value 5.5%), but is a clear win with its 22% increase in training size.

7.3 Performance of the Event Extractors

Most previous relation extraction approaches either require a manually labeled training set, or work only on a pre-defined set of relations that have ground instances from KBs. The closest work to NEWSSPIKE-RE is Universal Schemas (Riedel et al., 2013), which addresses the limitation of distant supervision that the relations must exist in KBs. Their solution is to treat the surface strings, dependency paths, and relations from KBs as equal "schemas", and then to exploit the correlation between the instances and the schemas from a very large unlabeled corpus. In their paper, Riedel *et al.* evaluated only on static relations from Freebase and achieve state-of-the-art performance. But Universal Schemas can be adapted to handle events, by introducing the events as schemas and heuristically finding seed instances.

We set up a competing system (**R13**) as follows: (1) We take the NYTimes corpus published between 1987 and 2007 (Sandhaus, 2008), the dataset used by Riedel *et al.* (2013) containing 1.8 million NY Times articles; (2) The instances (*i.e.* the rows of the matrix) come from the entity pairs from the news articles; (3) There are two types of columns: some are the extraction features used by NEWSSPIKE-RE, including the lexicalized dependency paths described in Riedel *et al.*; others are event relations $E = e(t_1, t_2)$; (4) For an entity pair (a_1, a_2), if there is an OpenIE extraction (a_1, e, a_2) and the entity types of (a_1, a_2) match (t_1, t_2), we assume the event relation E is observed on that instance.

As shown in Table 1, parallel news streams are a promising resource for clustering because of the strong correlation between the instances and the event phrases. We train another version of Universal Schemas **R13P** on the parallel news streams NS13. In particular, entity pairs from different NewsSpikes are used as different rows in the matrix.

We would like to measure the precision and recall of the extractors. But note that it is impossible to fully label all the sentences, so we follow the "pooling" technique described in (Riedel et al., 2013) to create the labeled dataset. For every competing system, we sample 100 top outputs for every event relation and add this to the pool. The annotators are shown these sentences and asked to judge whether the sentence expresses the event relation or not. After that, the labeled set become "gold" and can be used to measure the precision and pseudo-recall. There are in all 6,178 distinct sentences in the pool, since some outputs are produced by multiple systems. Among them, 2,903 sentences are labeled as positive. In Table 2, the # columns show the number of true extractions in the pool for every event relation.

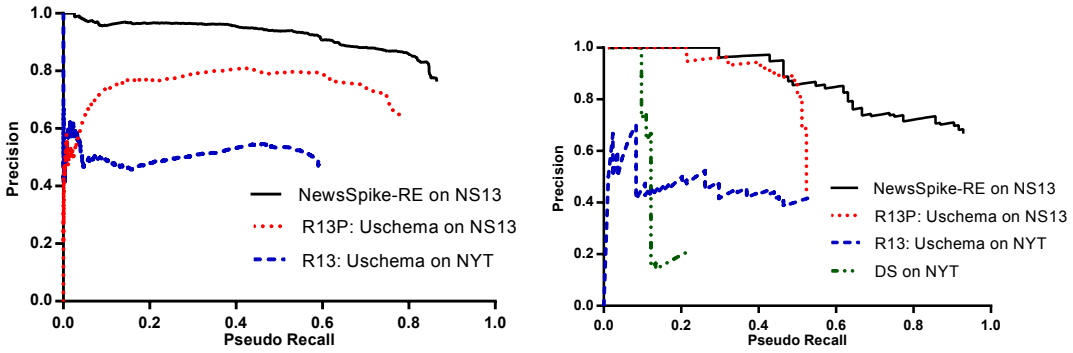
Similar to the diverse condition in Table 1, it is important that the extractor can correctly predict on diverse sentences that are dissimilar to each other. Thus we conducted a "diverse pooling": for each system, we report numbers for the sentences with different dependency paths between the arguments for every discovered event.

Figure 5(a) shows the precision pseudo-recall curve for all sentences for the three systems. NEWSSPIKE-RE outperforms the competing systems by a large margin. For example, the area under the curve (AUC) of NEWSSPIKE-RE for all sentences is 0.80 while that of R13P and R13 are 0.59 and 0.30. This is a 35% increase over R13P and 2.7 times the area compared to R13. Similar increases in AUC are observed on diverse sentences. Table 2 further lists the breakdown numbers for each event relation, as well as the micro and macro average. Although Universal Schemas had some success for several relations, NEWSSPIKE-RE achieved the best F1 for 26 out of 30 event relations; best AUC for 26 out of 30. The advantage is even greater in the diverse condition. It is interesting to see that R13P performs much better than R13, since the data coming from NYTimes is much noisier.

A closer look shows that Universal Schemas tends to confuse correlated but different phrases. NEWSSPIKE-RE, however, rarely made these errors because our model can effectively exploit negative evidence to distinguish them.

7.3.1 Comparing to Distant Supervision

Although the most event relations in Table 2 cannot be handled by the distant supervised approach, it is possible to match *buy(org,org)* to Freebase relations with appropriate database operators such as



(a) (b)

Figure 5: Precision pseudo-recall curves for (a) all event relations; (b) *buy(org, org)*, this figure includes the distant supervision algorithm MIML learned from matching the Freebase relation[5] to The New York Times. NEWSSPIKE-RE has AUC 0.80, more than doubling R13 (0.30) and 35% higher than R13P (0.59) for all event relations.

join and select (Zhang et al., 2012). To evaluate how distant supervision performs, we introduce the system **DS on NYT** based on a manual mapping of *buy(org,org)* to the join relation[4] in Freebase. Then we match its instances to NYTimes articles and follow the steps of Surdeanu *et al.* (2012) to train the extractor.

The matching to NYTimes brings us 264 positive instances having 5,333 sentences, but unfortunately the sentence-level accuracy is only 13% based on examination of 100 random sentences. Figure 5(b) shows the PR curves for all the competing systems. Distant supervision predicts the top extractions correctly because the multi-instance technique recognizes some common expressions (*e.g.* buy, acquire), but the precision drops dramatically since most positive expressions are overwhelmed by the noise.

8 Conclusions and Future Work

Popular distant supervised approaches have limited ability to handle event extraction, since fluent facts are highly time dependent and often do not exist in any KB. This paper presents a novel unsupervised approach for event extraction that exploits parallel news streams. Our NEWSSPIKE-RE system automatically identifies a set of argument-typed events from a news corpus, and then learns a sentential (micro-reading) extractor for each event.

We introduced a novel, temporal negation heuristic for parallel news streams that identifies event phrases that are correlated, but are not paraphrases. We encoded this in a probabilistic graphical model to cluster sentences, generating high quality training data to learn a sentential extractor. This provides negative evidence crucial to achieving high precision training data.

Experiments show the high quality of the generated training sentences and confirm the importance of our negation heuristic. Our most important experiment shows that we can learn accurate event extractors from this training data. NEWSSPIKE-RE outperforms comparable extractors by a wide margin, more than doubling the area under a precision-recall curve compared to Universal Schemas.

In future work we plan to implement our system as an end-to-end online service. This would allow users to conveniently define events of interest, learn extractors for each event, and return extracted facts from news streams.

Acknowledgments

We thank Hal Daume III, Xiao Ling, Luke Zettlemoyer and the reviewers. This work was supported by ONR grant N00014-12-1-0211, the WRF/Cable Professorship, a gift from Google, and the Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181. 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.

[4] /organization/organization/companies_acquired⋈/business/acquisition/company_acquired

References

Eugene Agichtein and Luis Gravano. 2000. *Snowball*: extracting relations from large plain-text collections. In *ACM DL*, pages 85–94.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 2670–2676.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, pages 16–23.

Regina Barzilay and Kathleen R McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 50–57.

Edward Benson, Aria Haghighi, and Regina Barzilay. 2011. Event discovery in social media feeds. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, pages 389–398.

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

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

Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 77–86.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Computational Linguistics*, page 350.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1535–1545.

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

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Joint Human Language Technology Conference/Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2013)*, pages 758–764.

Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. 2004. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL)*, page 415.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT-ACL)*, pages 541–550.

Ruihong Huang and Ellen Riloff. 2013. Multi-faceted event recognition with bootstrapped dictionaries. In *the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, pages 41–51.

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

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

David McClosky, Mihai Surdeanu, and Christopher D Manning. 2011. Event extraction as dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT-ACL)*, pages 1626–1635.

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

Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum. 2011. Scalable knowledge harvesting with high precision and high recall. In *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM)*, pages 227–236.

George L Nemhauser and Laurence A Wolsey. 1988. *Integer and combinatorial optimization*, volume 18. Wiley New York.

Roi Reichart and Regina Barzilay. 2012. Multi event extraction guided by global constraints. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, pages 70–79.

Kevin Reschke, Martin Jankowiak, Mihai Surdeanu, Christopher D Manning, and Daniel Jurafsky. 2014. Event extraction using distant supervision. In *Language Resources and Evaluation Conference (LREC)*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases (ECML)*, pages 148–163.

Sebastian Riedel, David McClosky, Mihai Surdeanu, Andrew McCallum, and Christopher D Manning. 2011. Model combination for event extraction in BioNLP 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 51–55.

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

Alan Ritter, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 1104–1112.

Evan Sandhaus. 2008. *The New York Times annotated corpus*. Linguistic Data Consortium.

Yusuke Shinyama and Satoshi Sekine. 2006. Preemptive information extraction using unrestricted relation discovery. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, pages 304–311.

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

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2011. Probabilistic matrix factorization leveraging contexts for unsupervised relation extraction. In *Advances in Knowledge Discovery and Data Mining*, pages 87–99.

Fei Wu and Daniel S. Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 41–50.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1456–1466.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2012. Unsupervised relation discovery with sense disambiguation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 712–720.

Alexander Yates and Oren Etzioni. 2009. Unsupervised methods for determining object and relation synonyms on the web. *Journal of Artificial Intelligence Research*, 34(1):255.

Congle Zhang and Daniel S Weld. 2013. Harvesting parallel news streams to generate paraphrases of event relations. In *Proceedings of the 2013 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP)*, pages 455–465.

Congle Zhang, Raphael Hoffmann, and Daniel S Weld. 2012. Ontological smoothing for relation extraction with minimal supervision. In *Association for the Advancement of Artificial Intelligence (AAAI)*.