Doctoral Dissertation

Relation Extraction: Perspective from Weakly Supervised Methods

Phi Van Thuy

September 15, 2019

Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Phi Van Thuy

Thesis Committee:

Professor Yuji Matsumoto	(Supervisor)
Professor Satoshi Nakamura	(Co-supervisor)
Associate Professor Masashi Shimbo	(Co-supervisor)
Assistant Professor Hiroyuki Shindo	(Co-supervisor)

Relation Extraction: Perspective from Weakly Supervised Methods*

Phi Van Thuy

Abstract

Relation extraction is the task of recognizing and extracting semantic relations over entities expressed in text. Existing supervised systems for relation extraction require a large amount of labeled relation-specific data. However, in practice, most relation extraction tasks do not have any supervised training data available.

In this study, we focus on two main weakly supervised approaches, namely bootstrapping and distantly supervised relation extraction methods, which reduce the cost of obtaining labeled examples in supervised learning. The first part of the study addresses the subtasks of automatic seed selection for bootstrapping relation extraction, and noise reduction for distantly supervised relation extraction. Ours is the first work that formulates them as ranking problems, and propose methods that can be applied for both subtasks. Experiments show that our proposed methods achieve a better performance than the baseline systems in these subtasks.

The second part of the dissertation investigates distant supervision, a weakly supervised algorithm that automatically generates training examples by aligning free text with a knowledge base. We propose a novel neural model that combines

*Doctoral Dissertation, Graduate School of Information Science, Nara Institute of Science and Technology, September 15, 2019.

a bidirectional gated recurrent unit model with a form of hierarchical attention that is better suited to relation extraction. We demonstrate that an additional attention mechanism called piecewise attention, which builds itself upon segment level representations, significantly enhances the performance of the distantly supervised relation extraction task. In addition, we propose a contextual inference method that can infer the most likely positive examples in bags with very limited contextual information. The experimental results show that our proposed methods outperform state-of-the-art baselines on benchmark datasets.

Keywords:

relation extraction, weak supervision, bootstrapping, automatic seed selection, distant supervision, noise reduction, piecewise attention, contextual inference method

Acknowledgements

I consider myself lucky to have had great advisors during my graduate life. First of all, I would like to express my sincere appreciation and gratitude to Professor Yuji Matsumoto for welcoming me to his laboratory in 2014, and for his guidance during my research. His support and inspiring suggestions have been precious for the development of my master thesis in 2016, as well as this doctoral dissertation. It has been almost 5 years since I came to Japan, and I had a great time at Computational Linguistics Laboratory, in Nara Institute of Science and Technology (NAIST).

I would also like to thank Professor Satoshi Nakamura, Associate Professor Masashi Shimbo, and Assistant Professor Hiroyuki Shindo for their kind instructions and very valuable comments on my work. Thanks so much for all the great, thoughtful feedback on my research and this dissertation.

I would also like to thank my family for their unflagging love and unconditional support throughout my life and my studies.

Finally, there are my friends and labmates, especially Joan Santoso from Institut Teknologi Sepuluh Nopember, Indonesia; Van-Hien Tran and my friends and labmates at NAIST. Thank you all for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last five years.

Thank you very much, everyone!

Dedicated to my family and to all my teachers.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

An important information extraction task is relation extraction, whose goal is to recognize and extract semantic relations over entities expressed in text. Automatic extraction of semantic relations is a challenging task and has been studied by many researchers during recent years. It is also a crucial step towards applications in several fields, such as question answering, text summarization, machine translation, information retrieval and others.

Popular approaches for relation extraction include: 1) designing some linguistic rules to capture patterns in text, or 2) developing a supervised relation extraction system based on syntactic and semantic features extracted from the text given a set of positive and negatives relation examples. Both of these approaches have their own inevitable weaknesses. In the rule-based methods, patterns need to be manually defined for domain-specific semantic relations. These rules are usually hard to maintain and adapt to new domains. On the contrary, existing supervised systems for relation extraction require a large amount of labeled relation-specific data. However, in practice, most relation extraction tasks do not

have any supervised training data available.

In this study, we focus on two main weakly supervised approaches, namely *bootstrapping* and *distantly supervised* relation extraction methods, which significantly reduce the expensive cost for data labeling and human effort required.

Bootstrapping for relation extraction [1–3] is a class of minimally supervised methods frequently used in machine learning: initialized by a small set of examples called *seeds*, to represent a particular semantic relation, the bootstrapping system operates iteratively to acquire new instances of a target relation. Bootstrapping only requires a limited number of seeds to start with and harvests more instances from unlabeled data.

Another approach, called "*distant supervision*" [4], does not require any labels on the text. The assumption of distant supervision is that if two entities participate in a known Freebase relation, any sentence that contains those two entities might express that relation. Although distant supervision is still limited by the quality of training data, it can extract semantic relations between entities in a large amount of plain text weakly supervised by external knowledge bases, such as Freebase [5] and Wikidata [6, 7].

The high-level overview of my research is shown in Figure 1.1. The first part of this study addresses the subtasks of automatic seed selection for bootstrapping relation extraction, and noise reduction for distantly supervised relation extraction. Ours is the first work that formulates them as ranking problems, and propose methods that can be applied for both subtasks. Our methods are inspired by ranking instances and patterns computed by the HITS algorithm, and selecting cluster centroids using K-means, latent semantic analysis (LSA), or the non-negative matrix factorization (NMF) method. Experiments show that our proposed methods achieve a better performance than the baseline systems in these subtasks.

The second part of the dissertation investigates distant supervision, a weakly

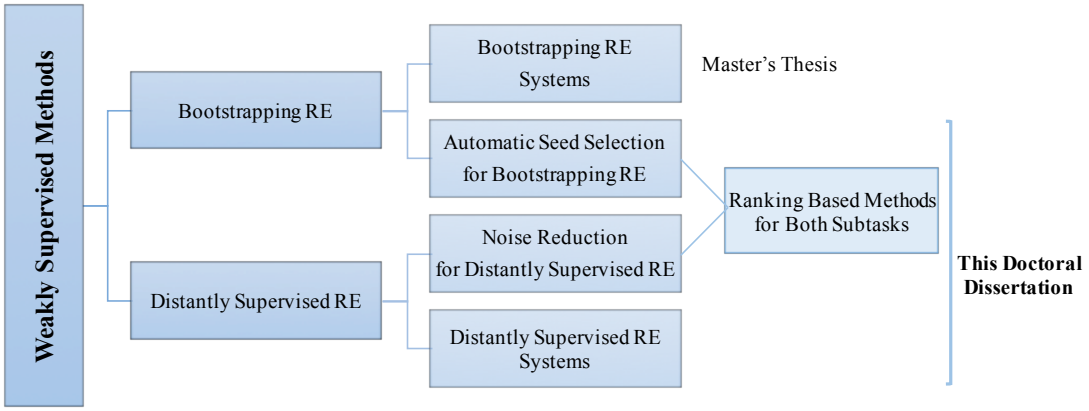

Figure 1.1: A high-level overview of my research.

supervised algorithm that automatically generates training examples by aligning free text with a knowledge base. We propose a novel neural model that combines a bidirectional gated recurrent unit (BiGRU) model with a form of hierarchical attention that is better suited to relation extraction. We demonstrate that an additional attention mechanism called piecewise attention, which builds itself upon segment level representations, significantly enhances the performance of the distantly supervised relation extraction task. In addition, we propose a contextual inference method that can infer the most likely positive examples in bags with very limited contextual information. The experimental results show that our proposed methods outperform state-of-the-art baselines on benchmark datasets.

1.2 Contribution

The main contributions of this dissertation are as follows:

- Methods for automatic seed selection for bootstrapping relation extraction and noise reduction for distant supervised relation extraction.
- An annotated dataset of 5,727 part-whole relations, which contains 8 sub-

types for the bootstrapping relation extraction system.

- Experimental results showing that the proposed models outperform baselines on two datasets in the subtasks of automatic seed selection and noise reduction.
- A novel BiGRU model combined with an additional attention mechanism called *piecewise attention* for distantly supervised relation extraction.
- A contextual inference method for improving bag label prediction for distantly supervised relation extraction.
- An annotated dataset of $5,863$ sentences, which is checked by annotators for false positive examples, to guarantee the quality of the distant supervision testing data.
- Experimental results showing that the proposed models outperform various state-of-the-art baselines on both original and annotated datasets for the distantly supervised relation extraction task.

1.3 Organization of the Dissertation

This dissertation is structured as follows:

Chapter 1 presents the introduction and motivation for this research.

Chapter 2 provides an overview of the weakly supervised relation extraction task and related works in common relation extraction tasks.

Chapter 3 presents common evaluation metrics used in most relation extraction systems.

Chapter 4 presents our proposed ranking-based automatic seed selection and noise reduction methods for weakly supervised relation extraction.

Chapter 5 investigates distant supervision for relation extraction using the piecewise attention and the bag-level contextual inference method.

Chapter 6 concludes the dissertation with discussions, a summary of the work, its contributions, and the direction of future research and improvements.

Chapter 2

Background on Weakly Supervised Relation Extraction

2.1 Relation Extraction Task

In general, relation extraction is defined as the task of extracting semantic relation between arguments. In the context of this study, we are interested in extracting binary relations, i.e., the relations between two entities, in the English newswire domain.

Trađitionally, relation extraction can be naturally cast as a supervised classification problem. Given a piece of text that contains two entity mentions, the goal of the relation extraction task is to determine whether that text contains a relation between the two entities [8]. Let the triple $r(e_1, e_2)$ denote a relation, where e_1 and e_2 are two entities contained in text, and r is a target relation[1]. Examples of binary relations include *located_in(Osaka, Japan)*, or *born_in(Barack_Obama,*

[1]Most of existing work in relation extraction focuses on the case where two entities e_1 and e_2 are pre-tagged in the unstructured text, and for a given pair of entities we need to determine the type of relationship that exists between the pair.

Honolulu).

Next, we introduce the background of two main weakly supervised approaches, namely *bootstrapping* and *distantly supervised* relation extraction methods, which significantly reduce the expensive cost for data labeling and human effort required.

2.2 Bootstrapping Relation Extraction

2.2.1 Background

Bootstrapping relation extraction [1–3] is a class of minimally supervised methods frequently used in machine learning: initialized by a small set of examples called *seeds*, to represent a particular semantic relation, the bootstrapping system operates iteratively to acquire new instances of a target relation.

Early bootstrapping methods are DIPRE [2] and Snowball [3], which rely on a few seeds and make use of bootstrapping to obtain patterns that express relations between entities in a large web-based text corpus. DIPRE represents the occurrences of seeds as three contexts of strings (words before the first entity, words between the two entities, and words after the second entity), and generates extraction patterns, i.e., *extractors*, by clustering contexts based on string matching. Snowball is inspired by DIPRE but it computes a TF-IDF representation of each context. Snowball's recall is generally higher than DIPRE's, while the precision of both techniques is comparable. Pantel and Pennacchiotti [9] propose a bootstrapping algorithm called *Espresso* to learn binary semantic relations, such as hypernym and meronym, by using the PMI-based pattern rankings. Recently, Ittoo and Bouma [10] use a minimally-supervised approach to extract part-whole relations from text iteratively. The novelty in their approach lies in using Wikipedia as a knowledge base, from which they first acquire a set of reliable patterns that express part–whole relations. They use different seed sets for

different subtypes of part-whole relations, and achieved an overall precision of 80%.

Bootstrapping only requires a limited number of seeds to start with and harvests more instances from unlabeled data. Selecting "*good*" seeds is one of the most important steps to reduce *semantic drift*, where the relations found by the system move further and further away from the original semantic relations defined by the seed sets, due to the ambiguity of the participating entities (words or phrases). However, seed selection is not yet well understood as pointed out by Kozareva and Hovy [11], since previous work mainly used random seed selection strategies, or manually chose the most frequent examples of the desired relations. Few semi-automatic or automatic seed selection methods have been proposed for a variety of tasks: word sense disambiguation [12, 13], named entity recognition [14], single-relation extraction [11].

We previously applied the bootstrapping algorithm to the part-whole relation extraction task [15] (described in subsection 2.2.2). The subtask of automatic seed selection for bootstrapping relation extraction is investigated in Chapter 4.

2.2.2 Our Prior Work on Part-Whole Relation Extraction

The *Espresso* algorithm [9] is a well-known bootstrapping system for extracting pairs of entities in a particular relationship. It takes as input a few seed instances and iteratively learns surface patterns to acquire more instances, and has proved to be effective by significantly improving recall while keeping high precision. The Espresso bootstrapping algorithm iterates between the following three phases: *Pattern Induction*, *Pattern Ranking/Selection*, and *Instance Extraction*.

In the *Pattern Induction* phase, the Espresso algorithm takes as input a set of instances I and produces as output a set of patterns P that connects the seed instances in a given corpus.

In the *Pattern Ranking/Selection* phase, the Espresso system creates a *Pattern Ranker*, and selects top-k patterns based on the pattern reliability score for the next phase. The reliability of a pattern p, $r_\pi(p)$ is the average strength of association across input i in the set of instances I, weighted by the reliability of each instance i:

$$r_\pi(p) = \frac{\sum_{i \in I}(\frac{pmi(i,\ p)}{max_{pmi}} * r_{\iota(i)})}{|I|} \tag{2.1}$$

where $r_\iota(i)$ is the reliability of instance i (defined below) and max_{pmi} is the maximum pointwise mutual information between all patterns and all instances. The pointwise mutual information (PMI) between instance i = *(x, y)* and pattern p is measured using the following formula:

$$pmi(i,\ p) = \log \frac{|x,\ p,\ y|}{|x,\ *,\ y||*,\ p,\ *|} \tag{2.2}$$

where $|x,\ p,\ y|$ is the frequency of the pattern p linked with the instance (x,y), and the asterisk (*) represents a wildcard. Then, $pmi(i,\ p)$ is multiplied with the discounting factor used in [16] to mitigate a bias towards infrequent events.

In the *Instance Extraction* phase, the Espresso algorithm retrieves from the corpus the set of instances I that match any of the patterns in P, then creates an *Instance Ranker*, and selects the top-m instances based on the instance reliability score. The reliability of an instance i, $r_{\iota(i)}$, is defined as:

$$r_{\iota(i)} = \frac{\sum_{p \in P}(\frac{pmi(i,\ p)}{max_{pmi}} * r_\pi(p))}{|P|} \tag{2.3}$$

In our previous work [15], we improved the Espresso system for the *part-whole* relation extraction task by integrating a word embedding approach into its iterations.

The key idea of our proposed *Espresso+Word2vec* model is utilizing an additional ranker component, namely *Similarity Ranker* in the *Instance Extraction*

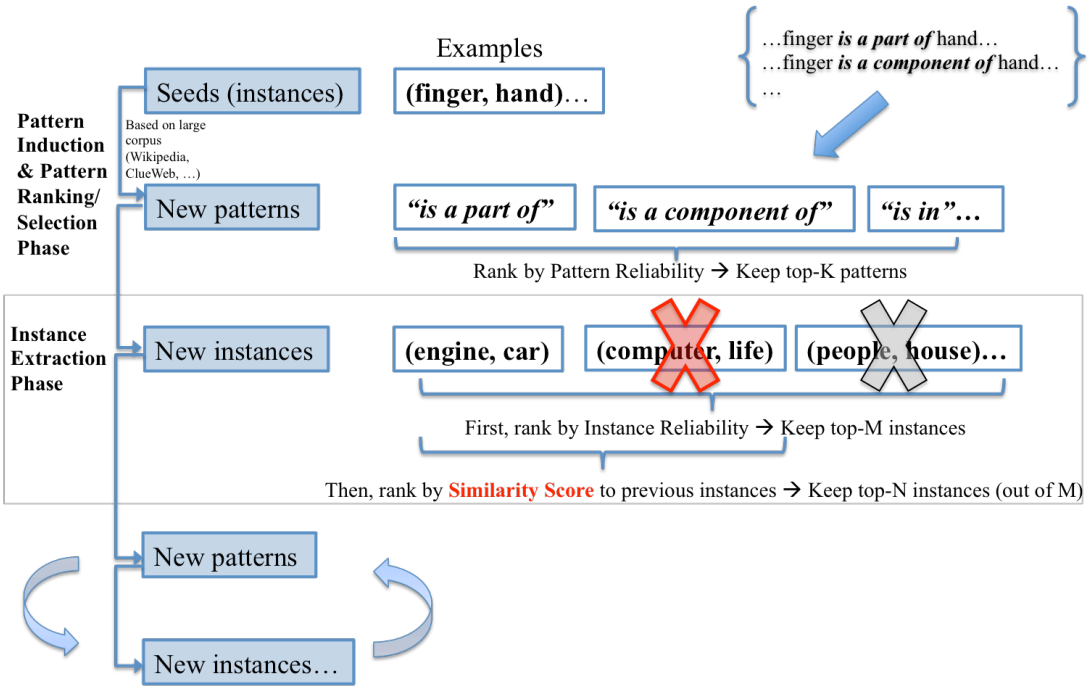

Figure 2.1: Illustration of our proposed model (*Espresso+Word2vec*) for extracting part-whole relations.

phase of the Espresso system. This ranker component uses the embedding offset information between instance pairs of part-whole relations. For each new instance, our ranker calculates the average similarity score between this instance and previous instances. The similarity score of an instance i, $SIM(i)$, is defined as:

$$SIM(i) = \frac{\sum_{j \in I_{Previous}} Cos\_sim(i,j)}{|I_{Previous}|} \tag{2.4}$$

where $Cos\_sim(i,j)$ is the cosine similarity between two instances, and $I_{Previous}$ is the set of extracted instances. Our calculation is mainly based on the recently proposed Word2vec model. We use the *word2vec* tool, and pre-trained vectors published by Google[2].

Figure 2.1 provides an illustration of our proposed model. In the Instance

[2]https://code.google.com/p/word2vec/

Extraction phase, the Espresso bootstrapping algorithm ranks instances first by the instance reliability, and removes unrelated pairs, e.g. *(people, house)*. This is intended to address the "*semantic drift*" phenomenon, as the proposed system filtered out the noisy instance *(people, house)* instead of keeping it for the next iteration. Then, the Similarity Ranker ranks the remaining instances and keeps top-n instances that have the highest similarity score. In our illustration, the instance *(computer, life)* is eliminated to keep a high precision over iterations.

The experiments show that our proposed *Espresso+Word2vec* system achieved a precision of 84.9% for harvesting instances of the part-whole relation, and outperformed the original Espresso system. *Espresso+Word2vec* is also the main system used in Chapter 4 for evaluating our different automatic seed selection strategies.

2.3 Distantly Supervised Relation Extraction

2.3.1 Background

Another weakly supervised approach that we focus on is distant supervision, which exploits existing knowledge bases instead of annotated texts as the source of supervision. The original assumption of distant supervision is that if two entities participate in a known Freebase relation, any sentence that contains those two entities might express that relation [4]. This assumption indicated that all sentences containing a known relation (e.g., in Freebase) might be potential *true positive* relation mentions. It is too strong and may cause the issue of incorrect labels. Consequently, it will deteriorate the performance of a model trained on such noisy data. *At-least-one* models make a relaxed distant supervision assumption [17]: *if two entities participate in a relation, at least one sentence that mentions these two entities might express that relation.* In this case, at least one

mention is considered as a true positive.

Ridel et al. [17], Hoffmann et al. [18], and Surdeanu et al. [19] introduced a series of models casting distant supervision as a multiple-instance learning problem [20]. In this multi-instance setting, the training set contains many entity-pair bags, and each bag consists of many *relation mentions*. Each relation mention is an occurrence of a pair of entities with the source sentence[3]. The labels of the bags are known; however, the labels of the relation mentions in these bags are unknown.

A distant supervision system has several key differences from traditional supervised relation extraction systems. First, the primary goal of a distant supervision system is to determine whether a relation between a given pair of entities is expressed *somewhere* in the text, and not necessarily where it is expressed [17]. In other words, a distant supervision system should predict labels for *relations* (i.e., entity pair labels), not *relation mentions* (i.e., sentence labels). By contrast, the objective of standard supervised relation extraction systems is to classify relation mentions (i.e., a sentence mentioned a specific entity pair). One of the most important benefits of focusing on relations instead of relation mentions is that it allows us to aggregate evidence for a relation from several places in the corpus. Second, in standard supervised learning, the gold annotations of all training sentences are given, whereas in distant supervision, only entity pair labels are provided. This, however, may serve as a challenge because distant supervision generates many noisy mentions that do not support target relations.

2.3.2 Labeling Procedure of Distant Supervision

We are given a corpus C and a knowledge base K that contains known triples (e_1, r, e_2) in which $r \in R$ (the set of relations we are interested in) and (e_1, e_2)

[3]Although some previous work used the term *instance* to indicate the sentence containing two entities, we used the original term *relation mention* as used in [17].

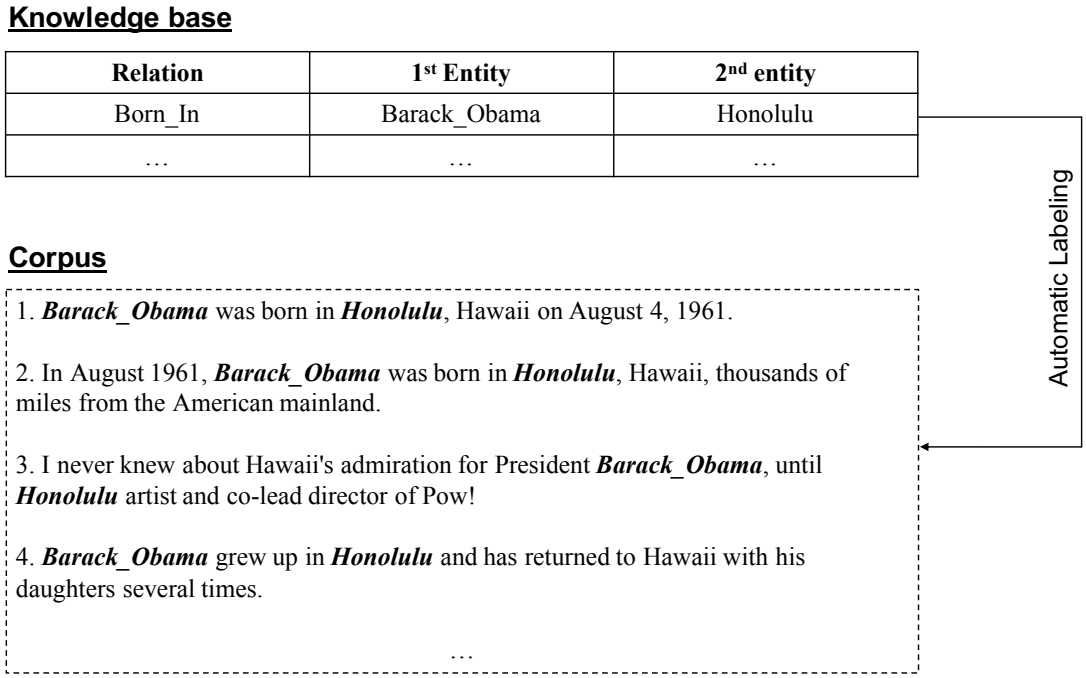

Figure 2.2: Labeling procedure of distant supervision.

is an entity pair that expresses the relation r. The labeling procedure of distant supervision for the relation extraction task is as follows:

- We align K to C; and for a triple (e_1, r, e_2) in K, all sentences (relation mention candidates) in C that simultaneously mention both entities e_1 and e_2 constitute a bag and are deemed as having the relation r. This generates a dataset that has labels on the entity-pair (bag) level with (possibly noisy) positive examples.
- Previous works typically assumed that if the argument entity pair (e_1, e_2) does not appear in K as holding a relation, all of the corresponding relation mentions in C are automatically annotated as negative examples (i.e., with "NA" labels).

Figure 2.2 shows a simple example of the labeling procedure of distant supervision in the relation extraction task. In the knowledge base, e.g., Freebase, e_1

$=$ *Barack_Obama* and $e_2 =$ *Honolulu* are two related entities, and $r =$ *Born_In* is the target relationship between them. According to the assumption of distant supervision, all sentences in the corpus, e.g., Wikipedia texts, that contain both entities $e_1 =$ *Barack_Obama* and $e_2 =$ *Honolulu* are considered to be (possibly noisy) positive training examples.

The data generated by the labeling procedure above can then be used by supervised learning algorithms to train relation extraction models.

Chapter 3

Common Evaluation Metrics for Relation Extraction

To test the effect of proposed relation extraction models, and to analyse the performance of the systems in general, several different evaluation metrics are used in the relation extraction task. This chapter gives details of common evaluation metrics used in most relation extraction systems.

3.1 Precision, Recall and F1 Score

Since the relation extraction task can be naturally cast as a supervised classification problem, evaluation metrics like *Precision*, *Recall* and *F1 Score* are used for performance evaluation.

The *Precision* measures the fraction of automatically extracted relations which

were correct over all the predicted relations in the testing set:

$$Precision(P) = \frac{\text{Number of correctly extracted relations}}{\text{Total number of extracted relations}} \tag{3.1}$$

The *Recall* measures the fraction of relations that were extracted over all actual relations that exists and should be extracted in the text:

$$Recall(R) = \frac{\text{Number of correctly extracted relations}}{\text{Actual number of relations}} \tag{3.2}$$

The *F1 Score* is the harmonic mean of *Precision* and *Recall*. In our study, we give equal importance to *Precision* and *Recall*:

$$\text{F1 Score} = \frac{2PR}{P+R} \tag{3.3}$$

3.2 Held-Out and Human Evaluation

In the absence of labeled testing data, evaluating weakly and semi-supervised methods is a slightly different process although the underlying metrics remain the same (Precision, Recall and F1 Score) [21].

In bootstrapping relation extraction, a small sample drawn randomly from the output is treated as a representative of the output and manually checked by human for actual relations. Then, the precision is calculated using the Equation 3.1. Calculating the recall is difficult given the large volume of relationships that are extracted.

In distantly supervised relation extraction, we can evaluate labels in two ways: (1) by holding out part of the Freebase relation data during training, and comparing newly extracted relation mentions against this held-out data, and (2) having

humans who look at each positively labeled entity pair and determine whether the relation actual holds between the two entities [4]. The former evaluation is an automated method. The Wikipedia texts will be aligned and annotated with the Freebase knowledge base to create a set of testing relation mentions. The Freebase relations will serve as a gold benchmark. This automated evaluation will make it possible to see the relative difference in performance between the different relation extraction systems using the Precision, Recall, and F1 Score. We can also report the precision/recall curves in the experiments.

As distant supervision may produce incorrect labels due to its automatic labeling procedure, the human evaluation is conducted to manually check the newly discovered relation mentions. Then, we can report the precision of top-k outputs with high confidence produced by relation extraction models. Note that we can not calculate the recall because the label of each relation mention in a particular bag is not provided. However, in combination with the automated evaluation, this human evaluation will give a detailed insight in the behaviour of the relation extraction systems.

Chapter 4

Ranking-Based Automatic Seed Selection and Noise Reduction for Weakly Supervised Relation Extraction

In bootstrapping relation extraction, selecting "*good*" seeds is one of the most important steps to reduce *semantic drift*, which is a typical phenomenon of the bootstrapping process. In distantly supervised relation extraction, another weakly supervised approach, noise reduction methods can reduce the issue of incorrect labels in positively labeled data generated based on the distant supervision assumption, which affects the performance of supervised learning.

This chapter presents the methods to figure out seeds for bootstrapping relation extraction, and filter out the noise from distant supervision. The main novelty of our work is defining the problems of automatic seed selection and noise reduction as the ranking problem. These problems are finding the ranking of data points (seeds or noisy examples), therefore suitable methods need to

score the data points based on a particular ranking criterion. From this insight, we propose various strategies such as K-means, Hypertext-induced topic search (HITS), latent semantic analysis (LSA), and non-negative matrix factorization (NMF) to solve the two problems with one method. For the experiments, we provide an annotated dataset of subtypes of part-whole relations. We compare our novel methods on both subtasks of seed selection and noise reduction, and these outperform the baselines.

4.1 Related Work

In our work, we propose methods that can be applied for both automatic seed selection and noise reduction by formulating these tasks as ranking problems according to different ranking criteria. We provide an overview of previous work related to seed selection approaches for bootstrapping algorithms, and noise reduction methods for distant supervision in this section.

4.1.1 Automatic Seed Selection for Bootstrapping Relation Extraction

Seed selection approaches for bootstrapping can be divided into manual, random, and automatic methods. Several works used a manual seed selection methodology, e.g., proposed by Hearst [22], Agichtein and Gravano [3] and Pantel et. al. [9]. Ittoo et. al. [10] and Phi and Matsumoto [15] also used a manual strategy, combing with the information of the frequencies of the target relations.

Rather than using the manual approach, seeds can be chosen at random [23, 24].

As manually selecting the seeds requires tremendous effort, some research

proposed methods to select the seed automatically. Eisner and Karakos [12] used a "strapping" approach to evaluate many candidate seeds automatically for a word sense disambiguation task. Kozareva and Hovy [11] proposed a method for measuring seed quality using a regression model and applied it to the extraction of unary semantic relations, such as"*people*" and "*city*". Kiso et al. [13] suggested a Hyperlink-Induced Topic Search (HITS) based approach to ranking the seeds, based on Komachi et al.'s analysis [25] of the Espresso algorithm [26]. Movshovitz-Attias and Cohen [14] generated a ranking based on pointwise mutual information (PMI) to pick up the seeds from existing resources in the biomedical domain.

We follow existing definitions of bootstrapping in previous studies [27–30]. Given the seed set of a target relation, the goal of the bootstrapping method is to find *instances* similar to initial seeds by harvesting *instances* and *patterns* iteratively over large corpora, e.g., Wikipedia or ClueWeb. We are interested in extracting semantic relationships between two entities in the English newswire domain.

4.1.2 Noise Reduction for Distantly Supervised Relation Extraction

The distant supervision assumption is too strong and leads to wrongly labeled data that affects performance. Many studies focused on methods of noise reduction in distant supervision. Intxaurrondo et al. [31] filtered out noisy mentions from the distantly supervised dataset using their frequencies, PMI, or the similarity between the centroids of all relation mentions and each individual mention. Xiang et al. [32] introduced ranking-based methods according to different strategies to select effective training groups. Li et al. [33] proposed three novel heuristics that use lexical and syntactic information to remove noise in the biomedical domain.

Then, the data generated by the noise reduction process can be used by su-

pervised learning algorithms to train relation extraction models.

4.2 Problem Formulation

Let R^* be the set of target relations. The goal is to find instances, or pairs of entities, upon which the relation holds. For each target relation $r \in R^*$, we assume there is a set D_r of triples representing the relation r. The triples in D_r have the form (e_1, p, e_2), where e_1 and e_2 denote entities, and p denotes the *pattern* that connects the two entities. A pair of entities (e_1, e_2) is called an *instance*. This terminology is similar to the one used in open information extraction systems, such as *Reverb* [34]. For example, in triple (*Barack Obama, was born in, Honolulu*), (*Barack Obama, Honolulu*) is the instance, and *"was born in"* is the pattern.

The two tasks we address are defined as follows:

Seed Selection for Bootstrapping Relation Extraction: In automatic seed selection, a set R^* of target relations and sets of instance-pattern triples $D_r = \{(e_1, p, e_2)\}$ representing each target relation $r \in R^*$ are given as input. These triples are extracted from existing corpus or database, e.g., WordNet. With these inputs, the task is to choose *good* seeds from the instances appearing in D_r for each $r \in R^*$, such that they work effectively in bootstrapping relation extraction.

Noise Reduction for Distantly Supervised Relation Extraction: In noise reduction for distantly supervised relation extraction, the input is the target relations R^* and the sets D_r of triples[1] generated automatically by distant supervision

[1] To be precise, in each triple (e_1, s, e_2) generated by distant supervision, s is not a pattern but a sentence that contains entities e_1 and e_2. However, we can easily convert each instance-sentence triple (e_1, s, e_2) to an instance-pattern triple (e_1, p, e_2) by looking for a pattern p that

for each relation $r \in R^*$. Because the data is generated automatically by distant supervision, D_r may contain noise, i.e., triples (e_1, p, e_2) for which relation r does not actually hold between e_1 and e_2. The goal of noise reduction is to filter out these noisy triples, so that they do not deteriorate the quality of the triple classifier trained subsequently.

Formulation as Ranking Tasks: As we can see from the task definitions above, both seed selection and noise reduction are the task of selecting triples from a given collection. Indeed, the two tasks essentially have a similar goal in terms of the ranking-based perspective.

We thus formulate them as the task of ranking instances (in seed selection) or triples (in noise reduction), given a set of (possibly noisy) triples. In the seed selection task, we use the k highest ranked instances as the seeds for bootstrapping relation extraction. Likewise, in noise reduction for distant supervision, we only use the k highest ranked triples from the distant supervision-generated data to train a classifier. Note that the value of k in noise reduction may be much larger than in seed selection.

4.3 Approaches to Automatic Seed Selection and Noise Reduction

In this section, we propose several methods that can be applied for both automatic seed selection and noise reduction tasks, inspired by ranking relation instances and patterns computed by the HITS algorithm, and picking cluster centroids using the K-means, latent semantic analysis (LSA), or non-negative matrix factorization (NMF) method.

connects two entities in sentence s.

4.3.1 K-means-based Approach

The first method we describe is a K-means-based approach. It is described as follows:

1. Determine the number k of instances/triples that should be selected[2].
2. Run the K-means clustering algorithm to partition all instances in the input triples (see Section 4.2) into k clusters. Each data point is represented by the embedding vector difference between its entities; e.g., the instance $I = (\textit{Barack\_Obama}, \textit{Honolulu})$ corresponds to: $\mathbf{vec}(I) = \mathbf{vec}(\textit{"Barack\_Obama"}) - \mathbf{vec}(\textit{"Honolulu"})$, where $\mathbf{vec}(x)$ is the embedding vector of the word x (e.g., *Barack_Obama*). We use pre-trained vectors published by [35]. The K-means algorithm stops when the assignments do not change from one iteration to the next.
3. The instance closest to the centroid is selected in each cluster. Given that the number of clusters is k, the same number of instances/triples will be chosen.

4.3.2 HITS-based Approach

Hypertext-induced topic search (HITS) [36], also known as the *hubs-and-authorities* algorithm, is a link analysis method for ranking web pages. In HITS, a good hub is a page that points to many good authorities and vice versa; a good authority is a page that is pointed to by many good hubs. These hubs and authorities form a *bipartite* graph, where we can compute the hubness score of each node. Kiso et

[2] Depending on the task, instances or triples will be selected: instances for the automatic seed selection task, and triples for the noise reduction task. As instances are pairs of entities which are included in triples, we can simply convert between the instance and the triple, and apply a proposed method to both tasks.

al. [13] proposed a graph-based approach for selecting seeds using the rankings of instances calculated by the HITS algorithm, and applied to the word sense disambiguation task.

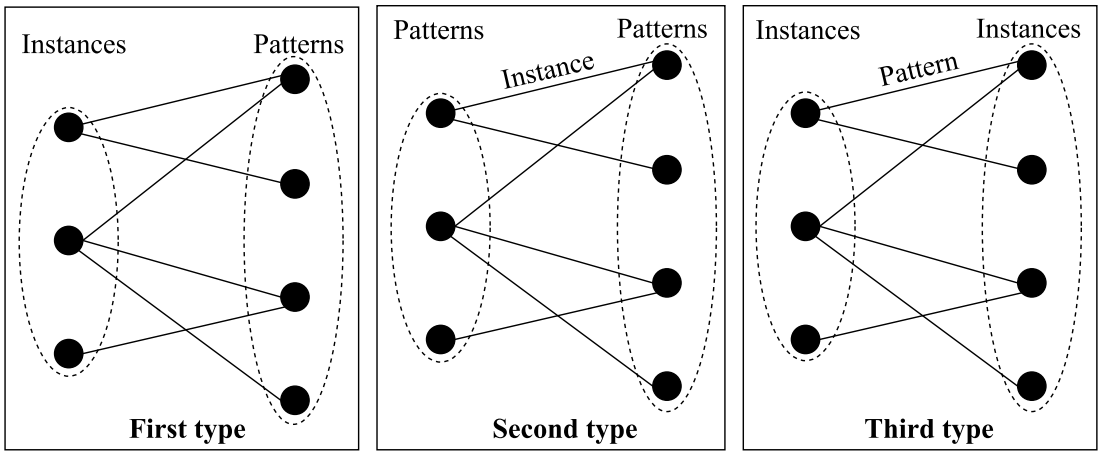

Figure 4.1: Graph representations of instances and patterns using the HITS algorithm.

In our task, let **A** be the instance-pattern co-occurrence matrix. We can compute the hubness score for each instance on the bipartite graph of instances and patterns induced by the matrix **A**. Inspired by the way HITS ranks hubs and authorities, our HITS-based seed selection strategy can be explained as follows:

1. Determine the number k of triples that should be selected.

2. Build the bipartite graph of instances and patterns based on the instance-pattern co-occurrence matrix **A**. Figure 4.1 presents three possible ways of building a bipartite graph. For the first type of graph, we consider each instance/pattern as a node in the graph. This representation is similar to that used by [13]. In the second graph representation, patterns and instances are treated as nodes and edges, respectively. Similarly, instances and patterns are treated as nodes and edges, respectively in the last representation.

3. For the first and third types, we simply retain the top-k instances with the highest hubness scores as the outputs (we sort the instances in descending

order based on their hubness scores). For the second type, k instances associated with the highest scoring patterns are chosen (we first sort the patterns in descending order based on their hubness scores).

4.3.3 HITS- and K-means-based Approach

By ranking instances and patterns computed by the HITS algorithm, and picking clusters' centroids using the K-means method, we can also select automatically top-k triples. In the combined method of HITS and K-means algorithms, we first rank the instances and patterns based on their bipartite graph and then run K-means to cluster instances in our annotated dataset. However, instead of choosing the instance nearest to the centroid, we retain the one that has the highest HITS hubness score in each cluster.

4.3.4 LSA-based Approach

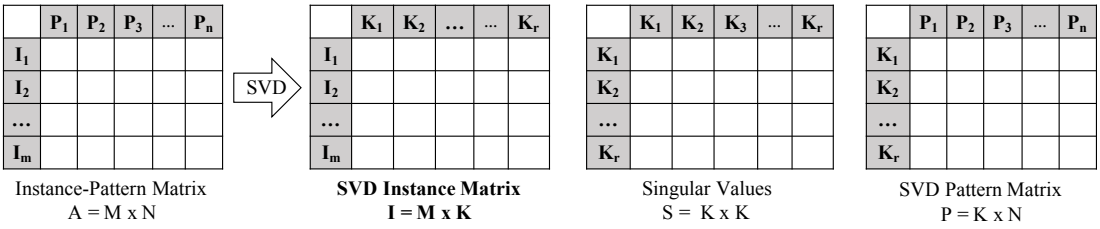

Figure 4.2: Decomposition of the instance-pattern matrix **A** using the LSA method.

Latent semantic analysis (LSA) [37] is also a widely used method for the automatic clustering of data along multiple dimensions. In our task, the instance-pattern co-occurrence matrix **A** is likely to have several thousands of rows and columns. Singular value decomposition (SVD) is used to construct a low-rank approximation of the instance-pattern co-occurrence matrix **A**. As illustrated in Figure 4.2, the SVD projection is performed by decomposing the matrix $\mathbf{A} \in$

$\mathbb{R}^{M\times N}$ into the product of three matrices, namely an SVD instance matrix $\mathbf{I} \in \mathbb{R}^{M\times K}$, a diagonal matrix of singular values $\mathbf{S} \in \mathbb{R}^{K\times K}$, and an SVD pattern matrix $\mathbf{P} \in \mathbb{R}^{K\times N}$:

$$\mathbf{A} \approx \mathbf{ISP}^T \tag{4.1}$$

Our LSA-based seed selection strategy is as follows:

1. Specify the desired number k of triples.
2. Use the LSA algorithm to decompose the instance-pattern co-occurrence matrix $\mathbf{A}$ into three matrices $\mathbf{I}$, $\mathbf{S}$, and $\mathbf{P}$. We set the number of LSA dimensions to $K = k$.
3. We can consider LSA as a form of soft clustering, with each column of the SVD instance matrix $\mathbf{I}$ corresponding to a cluster. Then, we select the k instances that have the highest absolute values from each column of $\mathbf{I}$.

4.3.5 NMF-based Approach

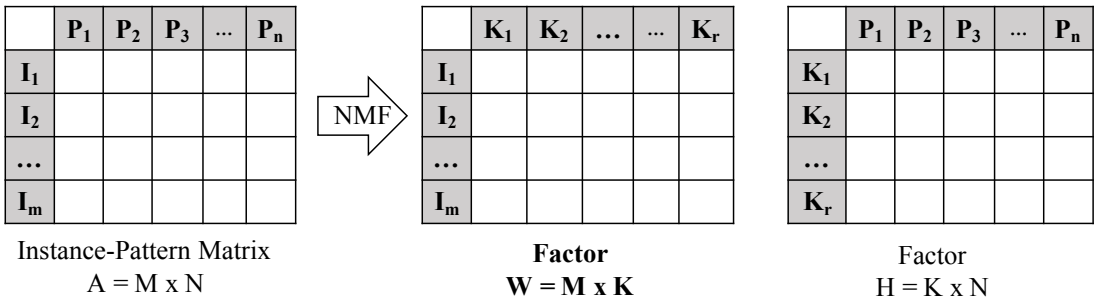

Figure 4.3: Illustration of approximate non-negative matrix factorization (NMF).

Non-negative matrix factorization (NMF) [38, 39] is another method for approximate non-negative matrix factorization, as shown in Figure 4.3. The non-negative data matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ is represented by two non-negative factors

$\mathbf{W} \in \mathbb{R}^{M \times K}$ and $\mathbf{H} \in \mathbb{R}^{K \times N}$, which, when multiplied, approximately reconstruct $\mathbf{A}$:

$$\mathbf{A} \approx \mathbf{WH} \tag{4.2}$$

The non-negativity constraints make the representation purely additive (allowing no subtractions), in contrast to many other linear representations such as principal component analysis (PCA) and independent component analysis. The non-negativity constraint is also the main difference between NMF and LSA.

Similarly to the LSA-based method, we set the NMF parameter K to k, the desired number of instances to select. We use the Projected Gradient NMF as it has good speed and performance for large-scale problems [40, 41]. We then select the k instances that have the highest values from each column of $\mathbf{W}$.

4.4 Experiments

4.4.1 Datasets and Settings

We provide an annotated dataset of part-whole relations as a reliable resource for selecting seeds. Our dataset was collected from Wikipedia and ClueWeb, and annotated by two annotators. One of its special characteristics is that the part-whole relation is a collection of relations, not a single relation [42, 43]. To the best of our knowledge, there are no datasets available for all fine-grained subtypes of the part-whole relation so far[3]. We use the part-whole taxonomy described in [15] since it is well-structured, clearly-presented, and it contains all subtypes in previous ontological studies. From that taxonomy, part-whole

[3]WordNet also provides a number of semantic relations, such as *synonymy, hypernonymy, and meronymy*. From examples of *meronymy, or the part-whole relation*, part-whole pairs are divided into *Part-Of, Member-Of*, and *Substance-Of* sub-categories. Nevertheless, they do not cover the variety of the part-whole relation.

relations include *Component-Of, Member-Of, Portion-Of, Stuff-Of, Located-In, Contained-In, Phase-Of*, and *Participates-In*.

Table 4.1: Statistics of our part-whole dataset.

Subtype	**Freq**
Component-Of	643 (11.23%)
Member-Of	1,272 (22.21%)
Portion-Of	555 (9.69%)
Stuff-Of	1,082 (18.89%)
Located-In	534 (9.32%)
Contained-In	272 (4.75%)
Phase-Of	497 (8.68%)
Participates-In	872 (15.23%)
TOTAL	5,727 triples

Table 4.1 gives the frequencies of each subtype of part-whole relations. There are $5,727$ instances of 8 subtypes that were annotated with the same labels by both annotators.

Then, we use selected seed sets as the initial seeds for *bootstrapping relation extraction* systems. We use *"Espresso+Word2vec"* [44], which is an improved version for the original Espresso algorithm [26] (described in subsection 2.2.2). *"Espresso+Word2vec"* outperformed the Espresso system for harvesting part-whole relations by utilizing the *Similarity Ranker*, which uses the embedded vector difference between instance pairs of relations. The performance is measured with *Precision@N* [45], $N = 50$. In total, $5,000$ instances are checked by annotators to ascertain whether they express part-whole relations. We vary the number k of seeds between 5 and 50 with a step of 5 to report the average P@50 of each seed selection method.

For the *noise reduction* task, we use the training and testing set developed by

Riedel et al. [17], which contains 53 relation classes. This dataset was generated by aligning Freebase relations with the New York Times corpus. After removing noisy triples from the dataset using the proposed methods, we use the filtered data to train two kinds of convolutional neural networks (CNN) (the CNN model in [46] and the PCNN model in [47]) with at-least-one multi-instance learning (ONE) used in [47], and the sentence-level attention (ATT) used in [48]. Finally, we report the area under the precision-recall (AUCPR) of each noise reduction method.

4.4.2 Performance on Automatic Seed Selection Task

The performances of the seed selection methods are presented in Table 4.2.

Table 4.2: Performance of seed selection methods.

Method	**Average P@50**
K-means	0.96
HITS_Graph1	0.90
HITS_Graph2	0.85
HITS_Graph3	0.90
HITS+K-means_Graph1	0.92
HITS+K-means_Graph2	0.85
HITS+K-means_Graph3	0.94
LSA	0.90
NMF	0.89
Random	0.75

For the HITS-based and HITS+K-means-based methods, we display the P@50 (P@N is the precision at N outputs) with three types of graph representation

as shown in Section 4.3.2. We use random seed selection as the baseline for comparison.

As Table 4.2 shows, the random method achieved a precision of 0.75. The relation extraction system that uses the random method has the worst average P@50 among all seed selection strategies. The HITS-based method's P@50s when using Graph1 and Graph3 are confirmed to be better than when using Graph2. This indicates that relying on reliable instances is better than reasoning over patterns (recall that for the Graph2, we first choose the patterns, then select the instances associated with those patterns), as there is a possibility that a pattern can be ambiguous, and therefore, instances linked to that pattern can be incorrect. The K-means-based seed selection method provides the best average P@50 with a performance of 0.96. The HITS+K-means-based method performs better than using only the HITS strategy, while the LSA-based and NMF-based methods have a comparable performance.

4.4.3 Performance on Noise Reduction Task

Recall that the K-means-based method achieves a high P@50 for the seed selection method. Our assumption is that each cluster may represent a set in which elements have similar semantic properties. However, we observed that as the number of relations is relatively high and there is no distinct definition between some relations in the distantly labeled data (e.g., the following three relations are quite similar: */location/country/capital*, */location/province/capital*, and */location/us_state/capital*, we decided not to perform the K-means-based method for our noise reduction task.

The performances of the HITS-based, LSA-based, and NMF-based noise reduction methods are presented in Table 4.3. We experimentally set the portion of retained data from the distantly labeled data to 90%, given that the performance can be affected if too many sentences are removed from the original data.

Table 4.3: Performance (Area Under the Precision-Recall - AUCPR) of each noise reduction method; in bold are the best scores.

System	**Original**	**+HITS**	**+LSA**	**+NMF**	**+Ensemble**
CNN+ONE	0.180	**0.183**	0.173	0.178	0.181
CNN+ATT	0.234	0.235	0.235	0.233	**0.236**
PCNN+ONE	0.231	0.234	0.233	0.234	**0.235**
PCNN+ATT	0.248	0.253	0.250	0.252	**0.255**

We also perform experiments with an ensemble method that combines the HITS-based and LSA-based strategies to merge rankings from their outputs, with half of the triples coming from the LSA-based method and the other half from the HITS-based method. Table 4.3 indicates that our proposed methods improved the performance of all CNN and PCNN models. Our ensemble method achieved the best improvements for three out of four systems, except that the HITS-based method obtained the best score for *CNN+ONE* (the CNN model with at-least-one multi-instance learning).

All the experimental results in this chapter confirmed that the idea of merging the two important subtasks (automatic seed selection and noise reduction), which lie between two main weakly supervised methods, is effective, and it helps in automating relation extraction systems and reducing the effort of developing and maintaining models to deal with separate subtasks. Our proposed methods showed their efficiency in terms of ranking the data points (seeds or noisy triples), and achieved higher performance than the baselines.

4.5 Conclusion

In this chapter, we formulated the seed selection and noise reduction subtasks as ranking problems. In addition, we proposed several methods, inspired by ranking instances and patterns computed by the HITS algorithm, and selecting clusters' centroids using the K-means, LSA, or NMF method. Experiments demonstrated that our proposed methods improved the baselines in both subtasks.

Chapter 5

Distant Supervision for Relation Extraction via Piecewise Attention and Bag-Level Contextual Inference

Distant supervision is a class of weakly supervised methods [49] and has become a popular approach for relation extraction to alleviate the lack of labeled examples in supervised learning. Distant supervision is an effective approach to scale relation extraction to very large corpora that contain thousands of relations without any labels on the text.

The term "*distant supervision*" was formally used by Mintz et al. [4] as a method of utilizing existing structured facts for obtaining training data without the manual labeling of examples. For the relation extraction task, distant supervision makes use of an already existing knowledge base such as Freebase or a domain-specific knowledge base to label entity pairs automatically in the text. This is then used to extract features and train a machine learning classifier. The

original "*distant supervision assumption*" is that *if two entities participate in a known Freebase relation, any sentence that contains these two entities might express that relation.* For example, Freebase contains the fact that *<Tokyo, is the capital of, Japan>*. We consider this fact and label each pair of "*Tokyo*" and "*Japan*" that appear in the same sentence as a positive example for the "*/location/country/capital*" relation. By aligning knowledge base facts with texts, distant supervision provides coherent positive training examples and avoids the high cost and human effort of manual annotation. Such large datasets allow for learning more complex models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). However, distant supervision often introduces noise to the generated training data. This approach can generate *false positives*, as not every mention of an entity pair in a sentence means that a relation is also expressed. As a result, distant supervision is still limited by the quality of the training data, and noise existing in positively labeled data may affect the performance of the supervised learning.

Recently, neural networks have been widely explored in distantly supervised relation extraction and achieved state-of-the-art results. Zeng et al. [47] treated relation extraction as a problem of multi-instance learning to relax the strong assumption of distant supervision: they assumed that "at least one document in the bag expresses the relation of the entity pair." Then, they divided the original input sentence into three segments by the positions of two entities and used piecewise max pooling to automatically learn relevant features using a piecewise CNN (*PCNN*). Lin et al. [48] addressed the shortcoming of the previous model, which only used the most relevant sentence from the bag. They proposed using sentence-level attention to capture the importance of each sentence, and then leveraging large amounts of useful data and information that is expressed by all sentences in each bag. Currently, *PCNN+ATT*, proposed by Lin et al. [48], is one of the state-of-the-art neural-network-based relation extraction models.

In this chapter, we propose a novel neural relation extraction model that combines a bidirectional gated recurrent unit (BiGRU) sequence model with a

form of hierarchical attention that is better suited to relation extraction. Our model consists of two attention modules: a piecewise attention that builds itself upon segment-level representations, and a sentence-level attention that builds itself upon sentence-level representations in each bag. Our piecewise attention not only captures crucial segments in each sentence but also reflects forward and backward directions of a sentence for better understanding the target relations between two entities.

The primary goal of relation extraction under distant supervision is to determine the relation for a given bag, i.e., between a given pair of entities. Hence, we propose using a contextual inference method that can infer the most likely positive examples of an entity pair in bags with very limited contextual information (i.e., for a bag with only a few sentences). Our inference method increases the number of positive examples and intentionally covers more contexts for target bags by using the similarity between entity pairs in positively labeled data. In addition, we provide an annotated dataset for the distantly supervised relation extraction task, which is based on the most commonly used dataset developed by Riedel et al. [17], and report on the actual performance of several relation extraction models.

5.1 Distantly Supervised Relation Extraction Task

The distantly supervised relation extraction task is usually decomposed into two steps. First, all sentences (or relation mentions) that contain mentions of two entities e_1 and e_2 are obtained following the labeling procedure described in subsection 2.3.2. Then, these sentences become the input to a relation extraction model, which then produces a set of relations which hold between the sentences [50]. We focus on the second step, i.e., classifying a set of pairs of entities into the target relations that they express.

The distantly supervised relation extraction task can be formalized as follows. We are given a training set T that contains N entity-pair bags $(B_1, B_2, ..., B_N)$. The n-th bag consists of n_b sentences (or relation mentions) $\{x_1, x_2, ..., x_{n_b}\}$ and the relation label r for a given entity pair (e_1, e_2). An relation extraction model M is trained with training set T to select valid sentences based on r for each bag. In the testing phase, our goal is to predict which relation types are expressed in the unseen bags, given all sentences in which both entities are mentioned in a large collection of unlabeled documents.

5.2 Related Work

The distantly supervised relation extraction task aims at identifying the semantic relation of a sentence set expressed toward an entity pair or a bag level [4]. Ridel et al. [17], Hoffmann et al. [18], and Surdeanu et al. [19] introduced a series of models casting distant supervision as a multiple-instance learning problem [20] to relax its original strong assumption.

Recently, neural networks have been widely explored in distantly supervised relation extraction and achieved state-of-the-art results [47, 48, 51]. Most existing systems model the noisy distant supervision process in the hidden layers by learning an informative sentence representation or features, and then selecting one or more valid relation mentions for relation extraction. Zeng et al. [47] divided the original input sentence into three segments by the positions of two entities, and used piecewise max-pooling to automatically learn relevant features using a piecewise CNN (*PCNN*) model. Lin et al. [48] and Ji et al. [52] addressed the shortcoming of the PCNN model, which uses only the most relevant sentence from each bag. They proposed to use sentence-level attention to dynamically calculate the weights of multiple sentences, and then leverage large amounts of useful information from all sentences in each bag. Currently, *PCNN+ATT* [48] is one of the state-of-the-art neural-network-based relation extraction models.

Zhou et al. [53] presented word-level attention integrated in a BiLSTM-based model and achieved significant improvements on SemEval2010 [54], which is a supervised dataset and cannot be used for the distantly supervised relation extraction task. Yang et al. [55] and Jat et al. [56] combined the word-level and sentence-level attention mechanisms in their single-layer BiGRU-based models and showed that these performed better than the CNN/PCNN models.

We believe that using only sentence-level or the word-level attention might not be the optimal solution because the crucial information should be distributed to different segments in the input sentence. Therefore, in this work, we develop two-layer BiGRU-based models with a combination of piecewise and sentence-level attention in order to capture the significance of each piece of text as well as the directionality of nonsymmetric relations.

We also make another contribution by proposing a novel contextual inference method that can support the bags with very few examples. In addition, previous works usually evaluated relation extraction systems in a held-out evaluation, which suffers from noise, e.g., in the Riedel dataset. Only a few works conducted manual evaluations with a small number of annotated sentences (e.g., 500 in [52]). By providing an annotated dataset of non-false positive examples, the real performance of various relation extraction systems can then be measured accurately.

5.3 Methodology

The distantly supervised relation extraction task is formulated as multi-instance learning. In this section, we introduce a novel neural relation extraction model that combines a BiGRU sequence model with a form of hierarchical attention that effectively incorporates the *piecewise* and *sentence-level* attentions. Furthermore, we propose to use a *contextual inference* method that can infer the most likely

positive examples of an entity pair in bags with limited contextual information without using any external knowledge resources or human annotations.

Our model takes input as an entity pair (e_1, e_2) and a bag $B = \{x_1, x_2, ..., x_{n_b}\}$ for (e_1, e_2), and predicts the probability $p(r|e1, e2)$ corresponding to the relation label r, $\forall r \in R$ (R is the set of relation labels). Our model consists of two main components:

- **Sentence Encoder** Given a sentence in $x \in B$, which contains two target entities, the sentence encoder outputs a distributed representation $\mathbf{x}$ of the sentence.
- **Bag Encoder** Given the encoding of each sentence in the bag for the entity pair (e_1, e_2), the bag encoder aims to learn a representation of the given bag, which is fed to a softmax classifier.

We briefly present the components of our model below. Each component will be described in detail in subsequent sections.

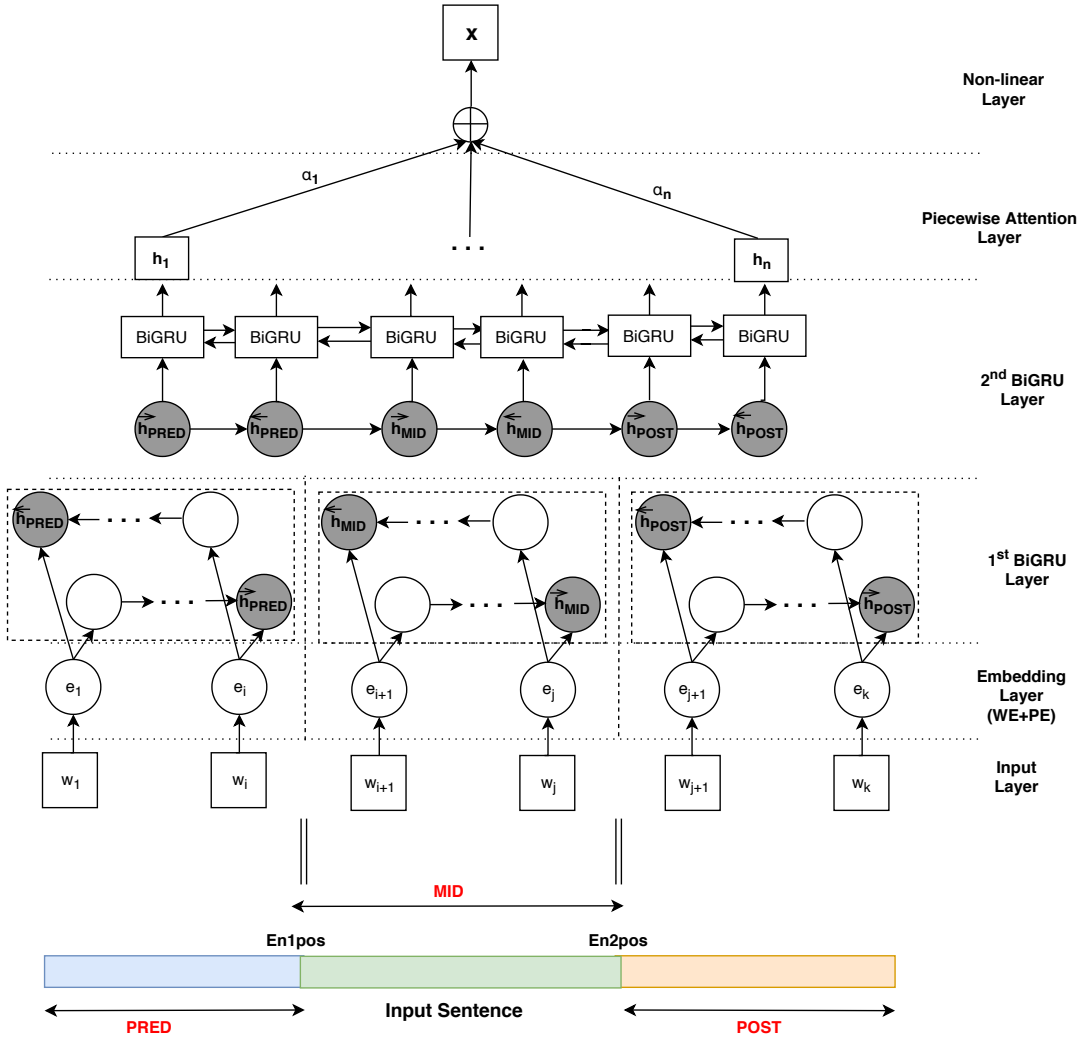

Figure 5.1: Architecture of our BiGRU model with piecewise attention used for sentence encoder.

5.3.1 Sentence Encoder

The overall architecture of the sentence encoder is depicted in Fig. 5.1, with the original sentence as the input to our model. Our sentence encoder has an embedding layer, two BiGRU layers, and a piecewise attention layer. These key modules are analyzed as follows.

Embedding Layer

Following previous work, we transform each input word of the source sentence into a combination of *word embedding* and *position embedding* in the embedding layer.

Word embeddings (WEs) aim to represent words as low-dimensional dense vectors. They can capture syntactic and semantic properties of words, such as in [35]. An embedding lookup table is first used to map words in the sentence into real-valued vectors. Word representations are encoded by column vectors in an embedding matrix $\mathbf{E} \in \mathbb{R}^{d^w \times |V|}$, where d^w is the dimensionality of the embedding space and $|V|$ is the size of the vocabulary.

Position embedding (PE) [51] is used to specify the positional information of the current word with respect to two target entities e_1 and e_2. Therefore, we define two lookup tables with two position embedding matrices $\mathbf{P_1}$ and $\mathbf{P_2}$, where $\mathbf{P_i} \in \mathbb{R}^{d^p \times |L|}$ (L is the maximum distance between any words of the sentence and two entities, and d^p is the dimension of the position embedding). $\mathbf{P_1}$ and $\mathbf{P_2}$ are randomly initialized. We then transform each relative distance (from the i-th word to e_1 or e_2) into a real-valued vector by looking up the position embedding matrices.

We concatenate the word and position embeddings as the input of the network. For a given sentence composed of k words, $x = \{w_1, w_2, ..., w_k\}$, we transform each word w_i into a real-valued vector. Then, x is fed into the next layer as $\mathbf{e}^x = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_k]$. If the size of the word representation is d_w and that of the position representation is d_p, then the size of a word vector is $d_w + 2d_p$.

1st BiGRU Layer

The role of the sentence encoder is to read the input sentence and construct an informative sentence representation. RNNs have been widely exploited to deal with variable-length sequence input. RNNs can learn long dependencies, but in practice they tend to be biased toward their most recent inputs in the sequence [57]. Long short-term memory networks (LSTMs) [58] incorporate a memory cell to combat this issue and avoid the vanishing gradient problem.

A gated recurrent unit (GRU) [59] is a simpler variant of the LSTM and was found to achieve better performance than the LSTM on some tasks [60]. A single-direction GRU has one drawback of not using the contextual information from the future words. A BiGRU exploits both the previous and future contexts by processing the sequence in two directions, and generates two independent sequences of GRU output vectors. Given the input sequence $\mathbf{e}^x = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_k]$, we employ a BiGRU as the recurrent unit, where the GRU is defined as

$$\mathbf{z}_i = \sigma(\mathbf{W}_z[\mathbf{e}_i; \mathbf{h}_{i-1}]), \tag{5.1}$$

$$\mathbf{r}_i = \sigma(\mathbf{W}_r[\mathbf{e}_i; \mathbf{h}_{i-1}]), \tag{5.2}$$

$$\tilde{\mathbf{h}}_i = \tanh(\mathbf{W}_h[\mathbf{e}_i; \mathbf{r}_i \odot \mathbf{h}_{i-1}]), \tag{5.3}$$

$$\mathbf{h}_i = (1 - \mathbf{z}_i) \odot \mathbf{h}_{i-1} + \mathbf{z}_i \odot \tilde{\mathbf{h}}_i, \tag{5.4}$$

where $\mathbf{W}_z$, $\mathbf{W}_r$ and $\mathbf{W}_h$ are weight matrices, σ is a sigmoid function, and $\odot$ is an element-wise multiplication operator. Initially, for $t = 0$, the output vector is $\mathbf{h}_0 = 0$.

Inspired by the *PCNN* model [47], we divide the original input sentence x into three segments by the positions of two entities e_1 and e_2. Fig. 5.1 illustrates these three segments, namely *PRED*, *MID*, and *POST* in our model. Let $En1pos$ and $En2pos$ be the positions of two entities in x. The input sequence $\mathbf{e}^x$ =

$[\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_k]$ of the BiGRU layer is divided into three independent subsequences:

$$\mathbf{e}_x^{\text{PRED}} = [\mathbf{e}_1, ..., \mathbf{e}_{En1pos}], \tag{5.5}$$

$$\mathbf{e}_x^{\text{MID}} = [\mathbf{e}_{En1pos}, ..., \mathbf{e}_{En2pos}], \tag{5.6}$$

$$\mathbf{e}_x^{\text{POST}} = [\mathbf{e}_{En2pos}, ..., \mathbf{e}_k]. \tag{5.7}$$

The repetitions of entities in Eq. (5.5), Eq. (5.6), and Eq. (5.7) mark the opening or closing of a coherent piece of text, and help our models extract informative distinct features over these adjacent text spans. Then, the first BiGRU layer processes each segment (PRED|MID|POST) separately. Concretely, the BiGRU consists of a forward GRU and a backward GRU. The forward GRU reads the input from left to right and generates a sequence of hidden states, e.g., ($\overrightarrow{\mathbf{h}}_1$, ..., $\overrightarrow{\mathbf{h}}_{En1pos}$) for $\mathbf{e}_x^{\text{PRED}}$. The backward GRU reads the input in reverse from right to left, and results in another sequence of hidden states, e.g., ($\overleftarrow{\mathbf{h}}_1$, ..., $\overleftarrow{\mathbf{h}}_{En1pos}$) for $\mathbf{e}_x^{\text{PRED}}$. The i-th hidden state is defined as

$$\overrightarrow{\mathbf{h}}_i = \text{GRU}(\mathbf{e}_i, \overrightarrow{\mathbf{h}}_{i-1}), \tag{5.8}$$

$$\overleftarrow{\mathbf{h}}_i = \text{GRU}(\mathbf{e}_i, \overleftarrow{\mathbf{h}}_{i+1}). \tag{5.9}$$

2nd BiGRU Layer

The 1st BiGRU model sequentially takes each word in the input sentence, extracts its information, and embeds it into a semantic vector. Owing to its ability to capture long-term memory, the BiGRU accumulates increasingly richer information as it goes through the sentence. The entire representation can be obtained as the final hidden state of the last word or time step. We retain the final forward and backward hidden states of each segment separately from the 1st BiGRU, and then feed them into the 2nd BiGRU layer.

Let $\overrightarrow{\mathbf{h}}_{PRED}$ and $\overleftarrow{\mathbf{h}}_{PRED}$ be the two final hidden states of the forward and backward directions generated for the *PRED* segment, respectively, and similarly

for the other segments. As shown in Fig. 5.1, we put these hidden states together in order of their occurrences in the input sentence to establish a direction-aware sequence:

$$(\overrightarrow{\mathbf{h}}_{PRED}, \overleftarrow{\mathbf{h}}_{PRED}, \overrightarrow{\mathbf{h}}_{MID}, \overleftarrow{\mathbf{h}}_{MID}, \overrightarrow{\mathbf{h}}_{POST}, \overleftarrow{\mathbf{h}}_{POST}). \tag{5.10}$$

The 2nd BiGRU takes the above sequence as the entire input, and can build up progressively higher-level representations of sequence data. Thus, it is more effective than the single-layer BiGRU encoder.

Piecewise Attention Layer

The attention mechanism was introduced by [61] in order to stress the target words step by step in machine translation. Recently, it was transferred to other tasks including distantly supervised relation extraction. Lin et al. [48] proposed a sentence-level attention scheme to select informative sentences from each bag. Jat et al. [56] recently introduced a model with sentence-level attention integrated with word-level attention to further explore the importance of different words in each sentence.

The word-level attention mechanism is a straightforward method to extract specific words that are important to the meaning of a sentence. However, a drawback of this method as an approach for distantly supervised relation extraction is that it is difficult to take the directionality of target relations into account. For example, we may know that two entities e_1 and e_2 should be related in a relation r (the relation is not symmetric in general), but we cannot really infer whether the triple (e_1, r, e_2) or (e_2, r, e_1) is correct without focusing on the right context in a given sentence.

All of the segments in an input sentence might provide necessary information to relation extraction. However, it is obvious that not all segments contribute equally to the sentence meaning for different relations. For example, considering

three cases from the Riedel dataset with two entities are in boldface, and the important segments are underlined:

<S1>. (*/people/person/nationality*) mr. burns said the indian foreign secretary , **shiv_shankar_menon**$_{<e_1>}$, had been invited to washington for talks early next month , and mr. burns planned then to travel to **india**$_{<e_2>}$.

<S2>. (*/location/location/contains*) kelly air force base closed in the 1990 's , but **san_antonio**$_{<e_1>}$ is still ringed by three air force installations as well as **brooke_army_medical_center**$_{<e_2>}$ and fort sam houston , the army 's largest base through world war ii .

<S3>. (*/people/person/children*) one , senator **evan_bayh**$_{<e_2>}$, above , son of former senator **birch_bayh**$_{<e_1>}$ of indiana , is testing the waters for a possible presidential bid in 2008 .

In the sentence <S1>, the left segment is more important than others to reflect the relation type */people/person/nationality*. In the sentence <S2>, the middle and right segments might provide the necessary information to the relation type */location/location/contains*. The right context in <S2> also supplement more useful information for predicting target relations. In the last example, the middle segment is the most important part related to the relation type */people/person/children*. In addition, the direction of the relation between two entities *birch_bayh* and *evan_bayh* in the sentence <S3> should be taken into account properly.

In our model, we therefore integrate a direction-aware attention layer over the 2nd BiGRU network to tackle the above challenges. We propose an additional attention mechanism called *piecewise attention*, which builds itself upon segment-level representations to improve the performance of the distantly super-

vised relation extraction task. Our piecewise attention not only captures crucial segments in each sentence but also reflects the direction of the target relation in each segment.

As shown in Fig. 5.1, we obtain hidden state representations of the sentence by feeding the sequence (10) into the 2nd BiGRU:

$$\{\mathbf{h}_1, ..., \mathbf{h}_6\} = \text{BiGRU}(\{\overrightarrow{\mathbf{h}}_{PRED}, ..., \overleftarrow{\mathbf{h}}_{POST}\}), \tag{5.11}$$

where

$$\mathbf{h}_j = [\overrightarrow{\mathbf{h}}_j \oplus \overleftarrow{\mathbf{h}}_j]; j = 1, 2, ..., 6, \tag{5.12}$$

and the number of hidden states produced by the 2nd BiGRU is 6, which is equal to the number of components of the input to the BiGRU in Eq. (5.11). Here, we use the element-wise sum (the symbol $\oplus$ in Eq. (5.12)) to combine the forward and backward pass outputs.

Next, we apply the attention mechanism at the segment level to assign a weight α_i to each hidden vector $\mathbf{h}_i$ generated by the BiGRU network, and pay more attention to the informative segment. The piecewise attention α_i is given by

$$\mathbf{h'}_i = \tanh(\mathbf{h}_i), \tag{5.13}$$

$$\alpha_i = \frac{\exp(\mathbf{w}^\top \mathbf{h'}_i)}{\sum_k \exp(\mathbf{w}^\top \mathbf{h'}_k)}, \tag{5.14}$$

where $\mathbf{w}$ is a parameter vector to be trained, and $\mathbf{w}^\top$ is a transpose of $\mathbf{w}$.

Finally, we aggregate the representation of these direction-aware segments to construct the sentence representation. The final sentence vector $\mathbf{x}$ is computed as a weighted sum of hidden states $\{\mathbf{h}_1, ..., \mathbf{h}_6\}$ as follows:

$$\mathbf{x} = \sum_{i=1}^{6} \alpha_i \mathbf{h}_i. \tag{5.15}$$

5.3.2 Bag Encoder

Following previous work [48], we use selective attention to deemphasize noisy sentences in the given bag. By using the sentence-level attention over sentences, a representation for the entire bag is learned. The details are described below.

Sentence-Level Attention Layer

In our model, the piecewise attention and the sentence-level attention are complemented to deemphasize the noisy samples. The sentence-level attention layer assigns higher weights to valid sentences and lower weights to invalid ones in a particular bag $B = \{x_1, x_2, ..., x_{n_b}\}$. The sentence-level attention β_i for the sentence vector $\mathbf{x}_i$ can be computed by

$$s_i = \mathbf{x}_i^\top \mathbf{A}\mathbf{r}, \tag{5.16}$$

$$\beta_i = \frac{\exp(s_i)}{\sum_k \exp(s_k)}, \tag{5.17}$$

where $\mathbf{A}$ denotes a diagonal weight matrix, $\mathbf{r}$ is a parameter vector related to relation r, and the query-based function s_i scores how well the input sentence x_i and the relation r match.

The final representation $\mathbf{b}$ for a given bag is computed as a weighted sum of its sentence vectors $\{\mathbf{x}_1, ..., \mathbf{x}_{n_b}\}$:

$$\mathbf{b} = \sum_{i=1}^{n_b} \beta_i \mathbf{x}_i. \tag{5.18}$$

where n_b is the number of sentences in the n-th bag.

Classification and Training

The bag vector $\mathbf{B}$ extracted from the segments and sentences of a bag B is a high-level representation of that bag and can be used as features for relation

classification. Then, $\mathbf{B}$ is passed to a softmax layer to predict the probability distribution corresponding to the relation labels. The conditional probability of the i-th relation is

$$p(r_i|B;\theta) = \frac{\exp(\mathbf{o}_i)}{\sum_k \exp(\mathbf{o}_k)}, \tag{5.19}$$

where θ denotes all parameters of our model, and $\mathbf{o} = \mathbf{Mb} + \mathbf{d}$ comprises the scores of all possible relations ($o \in \mathbb{R}^{|N|}$, where $\mathbf{M}$ is the representation matrix, d is a bias vector, and N denotes the number of relations).

We define the objective function using cross-entropy at the bag level [48]:

$$J(\theta) = \sum_{i=1}^{n_b} \log p(r_i|B;\theta) \tag{5.20}$$

In addition, we adopt the dropout strategy [62] and use stochastic gradient descent (SGD) to optimize our models.

5.3.3 Bag-Level Contextual Inference Method

The advantage of distantly supervised relation extraction lies in aggregating features from multiple sentences for the same entity pair. However, in many cases, there are insufficient number of sentences for a particular entity pair because of the limited coverage of the text corpus (e.g., when aligning the knowledge base with that corpus, we can not acquire many sentences for rare entity names, such as person/location names). For example, in the testing set developed by Riedel et al. [17], which is the most widely used dataset for the distantly supervised relation extraction task, there are $74,857$ entity pairs that correspond to only one sentence around $3/4$ overall entity pairs [48]. Therefore, it is desirable to infer more sentences for that entity pair. In addition, few sentence may not cover the diversity of the context for predicting the bag's label. More contexts may increase the confidence score of the prediction.

Using a small number of sentences in each test bag may affect the accuracy of the prediction in the testing phase. We therefore propose using a *contextual inference* method that can infer the most likely positive examples of an entity pair in test bags with limited contextual information without using any external corpora or knowledge bases. The target bags are those containing only one or very few sentences in the testing phase.

For example, consider the following two sentences:

$\mathbf{s_1}$: "... ***Tokyo*** *is located in* ***Japan*** ..." <in training data>

$\mathbf{s_2}$: "... ***Paris*** *is the capital of* ***France*** ..." <in testing data>

In the above example, the sentence s_1 belongs to the bag *(Tokyo, Japan)* in the training set, and the s_2 is in the bag *(Paris, France)* in the testing set. Our assumption is that *if these two bags have a high similarity, their two entity pairs can be replaced by each other to form new sentences that may cover more contexts for the target relations*. One of the new examples can be produced by this assumption is "*Paris is located in France*."

We use the cosine function to measure the similarity of two bags. Each bag is represented by the embedding difference between its entity vectors [15], e.g., the bag *(Tokyo, Japan)* corresponds to vec("*Japan*") - vec("*Tokyo*"). The similarity between two bags (e_1, e_2) and (x_1, x_2) is defined as

$$\mathrm{Sim}((e_1, e_2), (x_1, x_2)) = \cos([vec(e_2) - vec(e_1)], [vec(x_2) - vec(x_1)]) \quad (5.21)$$

Our bag-level contextual inference method is described in Algorithm 1. We leverage the given training data to generate *artificial* sentences, and hence increase the number of positive examples for each bag in the testing phase. It is expected that the newly generated sentences will share a similar semantic meaning with the target bag and provide supporting contexts for prediction. Our inference method aims to find high-quality sentences and avoid noise added to

Algorithm 1 Bag-level contextual inference

For each target bag (e_1, e_2) in a testing set (e.g., bags with only one sentence):

1. Find top-k similar bags to (e_1, e_2) from training set according to Eq. (5.21). Each sentence s in these similar bags has the form (x_1, c, x_2), where x_1, x_2 are two entities, and c is the context in s.

2. A new artificial sentence s' is generated with the form (e_1, c, e_2) by joining (e_1, e_2) and c.

3. Retain a maximum number of sentences s' (e.g., 5) added to the bag (e_1, e_2).

4. Include the newly generated sentences s' in the bag (e_1, e_2) to support the prediction.

the target bags. It can be integrated in our proposed BiGRU-based model. To the best of our knowledge, our contextual inference method is the first approach that can infer more examples for the target relations leveraging the similarity of two bags, without using any external resources in the distantly supervised relation extraction task.

5.4 Experiments

5.4.1 Datasets and Settings

Riedel Dataset

We use the Riedel dataset introduced in [17], which is the most commonly used dataset for the distantly supervised relation extraction task. It was generated automatically by aligning New York Times (NYT) articles with the Freebase knowledge base. Articles from 2005–2006 are used as training, and articles from 2007 are used as testing. The training data contain $522,611$ sentences, $281,270$

entity pairs, and 18,252 relational facts. The testing data contain 172,448 sentences, 96,678 entity pairs, and 1,950 relational facts. In total, there are 53 relation labels including the *NA* relation in this dataset. However, this automatically generated dataset could be incorrect owing to the limitation of the distant supervision assumption.

Our Annotated Dataset

A training dataset for distant supervision is created with the following simple rule: If a sentence mentions two entities e_1 and e_2 and they are known to have a relation r (according to a knowledge base such as Freebase), the sentence must be put in a bag for the relation r between entities e_1 and e_2. Nevertheless, this rule may produce many *false positive* sentences in a bag, as e_1 and e_2 may have occurred in the same sentence merely by chance. Consequently, the existence of false positive sentences in a bag can hurt the performance of relation extraction models.

We therefore provide an annotated dataset to guarantee the quality of the data and report on the real performance of various relation extraction systems. The Riedel testing set comprises 172,448 sentences, and 6,444 of them are labeled as *positive* examples by the distant supervision assumption. As some of them appear several times, we use 5,863 unique positive examples for our annotation. To the best of our knowledge, our current work is the first that provides such a high number of annotated sentences for the distantly supervised relation extraction task.

In the first stage, we request two annotators to check independently if 5,863 sentences express the target relations. Second, the two annotators discuss the disagreed labels in order to reach a consensus. The details of the second stage of our annotation process are listed in Table 5.1. There are 1,529 sentences where both annotators are marked as "false positive" and 4,246 sentences marked as

Table 5.1: Details of the second stage of our annotation process.

		Annotator 2	
		False positive	True positive
Annotator 1	False positive	1,529	46
	True positive	42	4,246

"No" (i.e., true positive). The Cohen's kappa coefficient on our annotation is 0.96, which indicates a strong agreement between annotators. For 88 sentences (1.5%) for which the two annotators cannot reach an agreement, another participant is involved in the decision-making process. Finally, 1,575 of 5,863 sentences (26.86%) are judged as false positive by three annotators.

Experimental Settings

We follow the parameter settings that are similar to those used in previous baselines [47, 48] in order to evaluate the effectiveness of our proposed methods. We use the word embeddings trained on the NYT corpus. The entities consisting of multiple tokens are considered as a single token. The dimensions for the word embedding (WE) and position embedding (PE) are set to 50 and 5, respectively. We use the maximum relative distance $L = 100$ in the position embedding, which is randomly initialized. The BiGRU hidden unit size is set to 230. We use a dropout with probability $p = 0.5$ and learning rate $\lambda = 0.01$ for the SGD.

For the bag-level inference method, we also use the skip-gram word2vec model to measure the similarities between different bags. The target bags are those with only 1 sentence. The maximum number of sentences added to each bag is 5. We tune the top-k similar bags to the target bag when our inference method is combined with others.

For evaluation, we report on the performance of models by using a precision-recall curve and top-N precision (P@N) metrics, which were commonly used in previous works.

Compared Models

To evaluate our proposed models, we compare them against the previous baselines for the distantly supervised relation extraction task. All of the models are described as follows:

- **Mintz**: A multiclass logistic regression model [4].
- **MultiR**: A probabilistic graphical model for multi-instance learning [18].
- **MIMLrelation extraction**: A graphical model that jointly models multiple instances and multiple labels [19].
- **CNN+ONE**: A CNN-based relation extraction model [51] with multi-instance learning [47].
- **CNN+ATT**: A CNN-based relation extraction model [51] with sentence-level attention [48].
- **PCNN+ONE**: A CNN-based relation extraction model [47] that uses piecewise max-pooling to generate the sentence representation.
- **PCNN+ATT**: A piecewise max-pooling over a CNN-based model to obtain the sentence representation, followed by sentence-level attention [48].

Currently, *PCNN+ATT* is one of the state-of-the-art neural-network-based relation extraction models for this task.

- **PCNN+ATT+Inference**: The model *PCNN+ATT* combined with our *bag-level contextual inference method*.
- **BGWA**: A recent single-layer BiGRU-based relation extraction model with word-level and sentence-level attention [56].
- **2BiGRU+PATT**: Our proposed model, which uses two BiGRU layers and *piecewise* attention.
- **2BiGRU+PATT+Inference**: Our proposed model *2BiGRU+PATT* combined with the *bag-level contextual inference method*.

We refer to three feature-based systems (Mintz, MultiR, and MIMLrelation extraction) as the traditional models, and neural-network-based systems (CNN+ONE, CNN+ATT, PCNN+ONE, PCNN+ATT, PCNN+ATT+Inference, and BGWA) as the state-of-the-art models for comparison. An analysis of the results is provided in the next section.

5.4.2 Experimental Results and Analysis

Comparison with Traditional Methods

We evaluate our proposed models (2BiGRU+PATT and 2BiGRU+PATT+Inference) and compare them with three conventional feature-based methods (Mintz, MultiR, and MIMLrelation extraction) on the Riedel dataset. The precision-recall curve of each system is shown in Fig. 5.2. It is obvious that our proposed models significantly outperform all feature-based methods over the entire range of recall. When the recall is around 0.1, the performances of Mintz, MultiR, and MIMLrelation extraction drop quickly, while our models maintain high precision.

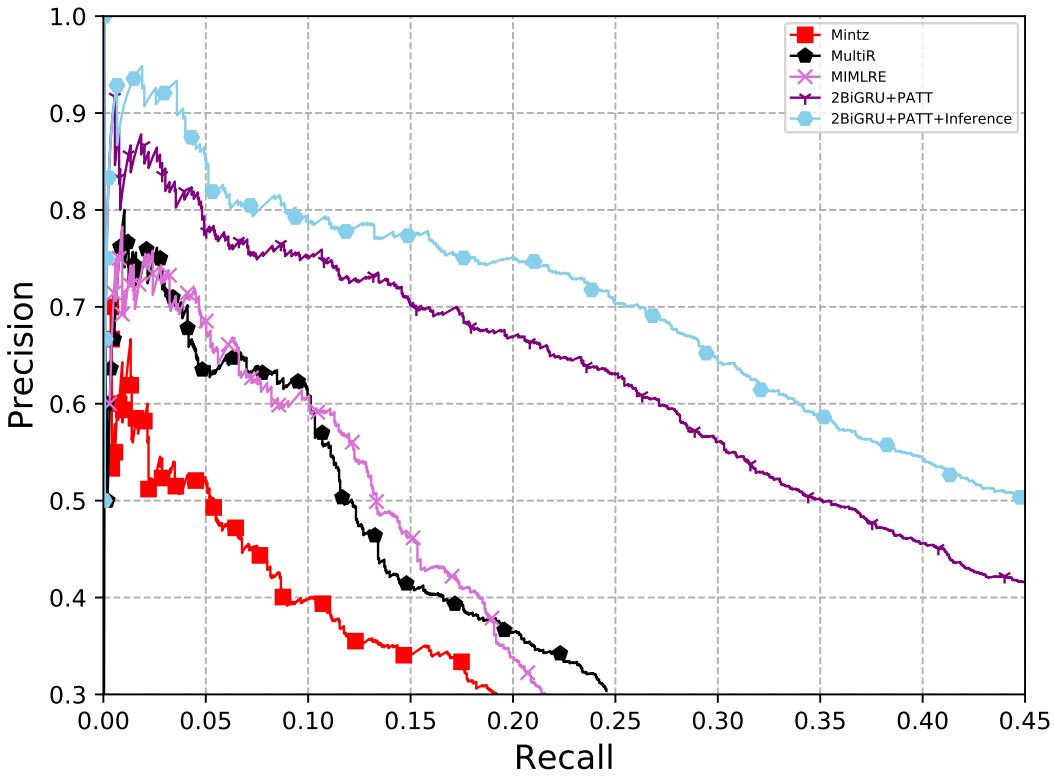

Figure 5.2: Performance comparison of the proposed model and traditional methods.

All of the feature-based methods used human-designed features, which are time consuming and labor intensive. By contrast, our models can automatically learn the intrinsic features without human intervention from a large number of training examples.

Effects of Our Proposed Methods and Comparison with State-of-the-Art Models

We compare our proposed models with two types of recent CNN-based models: the CNN model in [51] and the PCNN model in [47]) with at-least-one multi-instance learning (+ONE) used in [47] and the sentence-level attention (+ATT)

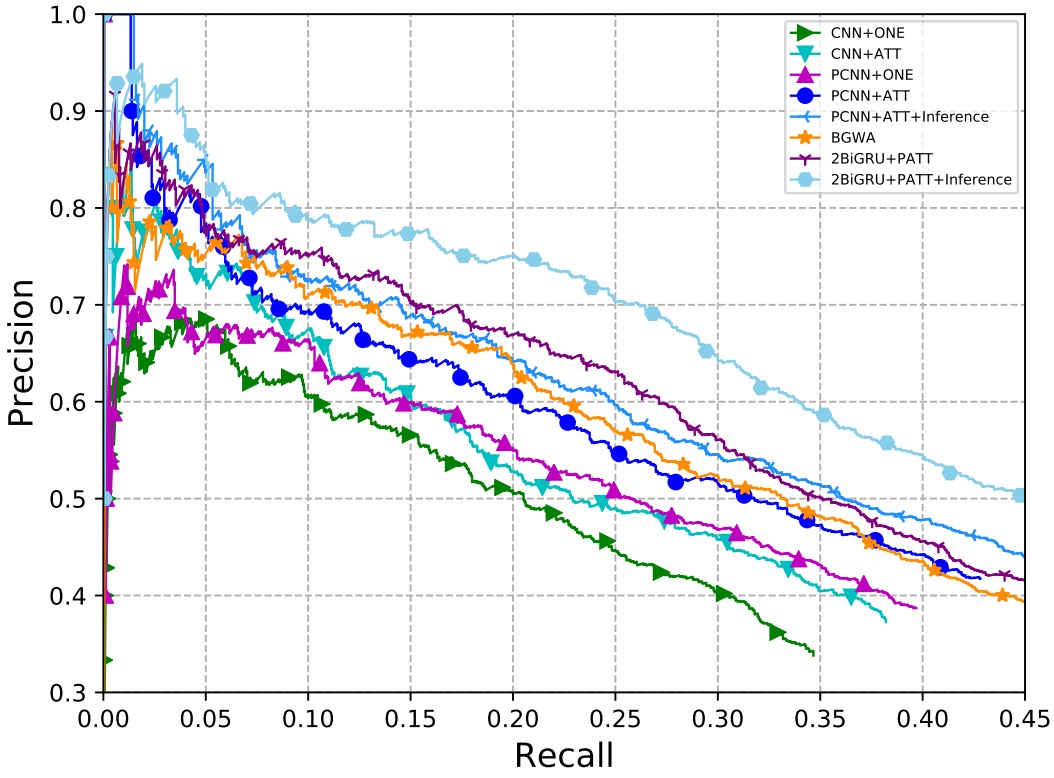

Figure 5.3: Performance comparison of the proposed model and state-of-the-art methods.

used in [48]. *PCNN+ATT* is one of the state-of-the-art neural-network-based relation extraction models reported in the Riedel dataset. The precision-recall curves of these models are presented in Fig. 5.3. The results show that our 2BiGRU+PATT model performs better than all CNN-based models to a significant extent, especially when compared to the state-of-the-art PCNN+ATT system. Our 2BiGRU+PATT+Inference model achieves the best performance among all of the methods. This demonstrates the effectiveness of our proposed models for the distantly supervised relation extraction task.

We also compare our models with BGWA, which is a recent single-layer BiGRU-based relation extraction model with word-level and the sentence-level attention [56]. From Fig. 5.3, we observe that the BGWA model achieves per-

formance that is comparable to that of the PCNN+ATT model. BGWA is considered a baseline for evaluating the effectiveness of our piecewise attention as BGWA and 2BiGRU+PATT employ similar hierarchical attention networks (word-level or piecewise attention combined with sentence-level attention). The results indicate that the precision value of our 2BiGRU+PATT model is higher than that of the BGWA model when the recall value changes. This demonstrates the effect of using piecewise attention instead of word-level attention. Our new attention mechanism helps the relation extraction models to focus on the right context in a given sentence and captures the directionality of nonsymmetric relations more efficiently.

Next, we compare the effects of integrating our bag-level contextual inference method into different systems. Our inference method boosts the performance of the PCNN+ATT system significantly and makes PCNN+ATT+Inference comparable to 2BiGRU+PATT. The inference method also enables the proposed 2BiGRU+PATT+Inference model to achieve a large improvement compared to the 2BiGRU+PATT model. All of these examples show the superiority of our method against the state-of-the-art methods.

Performance of Our Annotated Dataset

In the Riedel testing set, there are $172,448$ sentences, and $6,444$ of them are labeled as *positive* examples by the distant supervision assumption. We replace the labels of $6,444$ sentences in the Riedel testing set, which are checked by annotators, and refer to this as our annotated dataset. It means that we only changed the labels of false positive sentences to "*NA*" (i.e., true negative), and the total number of sentences is unchanged.

Fig. 5.4 shows the performance of our annotated dataset for three models: PCNN+ATT, 2BiGRU+PATT, and 2BiGRU+PATT+Inference. The "*" symbols denote the evaluations of our annotated dataset. It is observed that there are

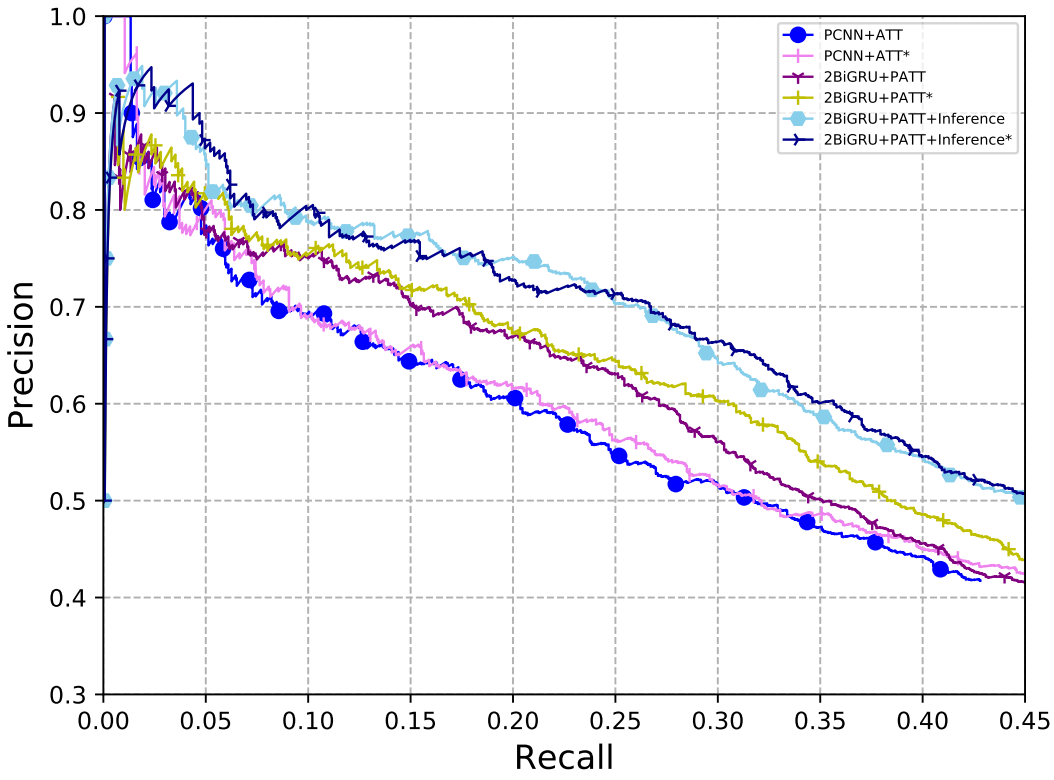

Figure 5.4: Performance of models on annotated dataset; * symbols denote evaluations of our annotated dataset.

slight changes when the results are reported on the original and our annotated dataset. However, all of the systems are robust, and our 2BiGRU+PATT model performs even better on the annotated dataset. Our bag-level contextual method still shows its benefits and does not require any external resources of knowledge bases. Furthermore, ours is the first work to report on the performance of various relation extraction models on an annotated dataset with a high number of testing examples ($5,863$) checked by humans.

Table 5.2: P@N for relation extraction in bags with different numbers of sentences; * symbols denote evaluations of our annotated dataset; One, Two, and All denote number of sentences randomly selected from a bag; best scores are in boldface.

Test Settings	**One**			
P@N(%)	100	200	300	Mean
PCNN+ATT	73.3	**69.2**	60.8	67.8
2BiGRU+PATT	**76.2**	66.7	**62.1**	**68.3**
PCNN+ATT*	70.0	63.0	58.7	63.9
2BiGRU+PATT*	**74.3**	**64.2**	**59.5**	**66.0**
	Two			
P@N(%)	100	200	300	Mean
PCNN+ATT	77.2	**71.6**	**66.1**	71.6
2BiGRU+PATT	**80.2**	69.2	65.8	**71.7**
PCNN+ATT*	76.0	**71.5**	65.0	70.8
2BiGRU+PATT*	**80.2**	68.7	**65.1**	**71.3**
	All			
P@N(%)	100	200	300	Mean
PCNN+ATT	76.2	**73.1**	67.4	72.2
2BiGRU+PATT	**83.2**	**73.1**	**69.8**	**75.4**
PCNN+ATT*	75.0	71.5	66.3	70.9
2BiGRU+PATT*	**83.2**	**72.6**	**69.1**	**75.0**

Effect of Sentence Number

Following previous works, we also evaluate our methods with different numbers of sentences in the bags with more than one sentence. In this setting, one,

two, or all sentences are (randomly) selected from each bag for comparison in the testing phase [48]. We then report the P@100, P@200, P@300, and their mean for each model. The results are listed in Table 5.2. In all settings, our 2BiGRU+PATT model obtains higher average precision than the PCNN+ATT model, which demonstrates the efficacy of our method. These improvements are observed on both datasets to an extent of 3.2% (using all sentences in the Riedel dataset) and 4.1% (using all sentences in our annotated dataset). Using all of the sentences helps the models achieve the best results. However, adding sentences might result in more noise, which can affect the performance. This is illustrated in the "One" and "Two" settings. The 2BiGRU+PATT model using two sentences does not produce a higher improvement than when using only one sentence: 71.6 to 71.7% and 67.8 to 68.3% on the Riedel dataset, respectively; and 70.8 to 71.3% and 63.9 to 66.0% on our annotated dataset, respectively.

P@N in All Bags

The P@N results for all bags are presented in Table 5.3. We can see that our proposed methods show their advantages and achieve notable performance for all values of P@100, P@200, P@300, and Mean. For the Riedel dataset, our 2BiGRU+PATT model performs better than the PCNN+ATT model when the average precision increases from 73.8% to 77.2%, and performs in a similar manner for the models that use our inference method (76.9% to 82.1%). For our annotated dataset, the scores also improved remarkably: 72.6 to 76.9% when using our novel BiGRU-based model, and 72.6 to 80.8% when incorporating the additional inference method. All of the proposed methods still show their robustness on both datasets.

Table 5.3: P@N for relation extraction in all bags; * symbols denote evaluations on our annotated dataset; best scores are in boldface.

Test Settings	All Bags			
P@N(%)	100	200	300	Mean
PCNN+ATT	81.0	71.0	69.3	73.8
PCNN+ATT+Inference	83.0	75.0	72.7	76.9
2BiGRU+PATT	82.2	75.6	73.8	77.2
2BiGRU+PATT+Inference	**87.1**	**81.1**	**78.1**	**82.1**
PCNN+ATT*	81.0	69.5	67.3	72.6
2BiGRU+PATT*	82.2	75.6	72.8	76.9
2BiGRU+PATT+Inference*	**86.1**	**79.6**	**76.7**	**80.8**

Parameter Tuning for Our Bag-level Contextual Inference Method

For our bag-level contextual inference method, we tune the top-k similar bags (this is shown in Algorithm 1) to find the best performance of two models: PCNN+ATT+Inference and 2BiGRU+PATT+Inference. The average P@N ($N = 100, 200, 300$) results for all bags are used for comparison. Table 5.4 lists the numbers of similar bags and inferred sentences that were generated by our inference method. When the number of similar bags increases, the number of inferred sentences is incremented accordingly in most cases. The maximum number of sentences is $1,807$, which corresponds to 28.04% of the positive examples in the original Riedel testing dataset. When the number of similar pairs >=15, the generated sentences are the same as for 14 since our method already generated all possible sentences for the bags with only one sentence.

The best average P@N score for each model is reported. The PCNN+ATT+Inference model reaches its best performance with top-$k = 2$, whereas our 2BiGRU+PATT+Inference model achieves the best result with top-$k = 9$. Compared to the original systems

Table 5.4: Parameter tuning for our bag-level context inference method; we only create data for bags with one sentence in testing set; maximum number of sentences added to each bag is five; when number of similar pairs >=15, generated sentences are same as for 14 since our method already generated all possible sentences for bags with one sentence; best score for each model is in boldface.

No. of Similar Bags (top-k)	1	2	3	4	5	6	7
No. of Inferred Sentences	756 (+11.73%)	1,302 (+20.20%)	1,526 (+23.68%)	1,656 (+25.70%)	1,713 (+26.58%)	1,748 (+27.13%)	1,769 (+27.45%)
PCNN+ATT+Inference - P@N (Mean)	76.7 ↑	**76.9** ↑	76.9 →	76.8 ↓	76.8 →	76.6 ↓	76.7 ↑
2BiGRU+PATT+Inference - P@N (Mean)	79.4 ↑	80.8 ↑	81.2 ↑	81.7 ↑	81.9 ↑	81.8 ↓	81.7 ↓
No. of Similar Bags (top-k)	8	9	10	11	12	13	14
No. of Inferred Sentences	1,786 (+27.72%)	1801 (+27.95%)	1801 (+27.95%)	1,802 (+27.96%)	1,803 (+27.98%)	1,803 (+27.98%)	1,807 (+28.04%)
PCNN+ATT+Inference - P@N (Mean)	76.7 →	76.7 →	76.7 →	76.7 →	76.7 →	76.7 →	76.7 →
2BiGRU+PATT+Inference - P@N (Mean)	82.0 ↑	**82.1** ↑	82.1 →	82.1 →	82.1 →	82.1 →	82.1 →

(which are listed in Table 5.3), the gap between 2BiGRU+PATT+Inference and 2BiGRU+PATT is higher than that of PCNN+ATT+Inference and PCNN+ATT: 82.1 to 77.2% compared with 76.9 to 73.8%, respectively. This is useful in practice because both models are beneficial when using the inference method to support the prediction. Our model shows its advantages and leverages the artificial data more efficiently.

5.5 Conclusion

In this chapter, we proposed novel neural relation extraction systems with two BiGRU layers and two attention modules: the piecewise and sentence-level attentions. We also presented a contextual inference method that can infer the most likely positive examples of an entity pair in bags with very limited contextual information without using any external knowledge bases or corpora. The experimental results showed that our proposed models offer significant improvements over state-of-the-art methods on our newly created dataset and the Riedel dataset.

Chapter 6

Conclusion

6.1 Discussion

6.1.1 Visualization of the Best Automatic Seed Selection Method in Bootstrapping Relation Extraction

In Figure 6.1, we run the K-means clustering algorithm, which achieved the best performance among our automatic seed selection methods, to partition all data points in our dataset into K=10 clusters. Each instance of the part-whole relation is represented by the embedding offset between its terms, e.g., the instance *(meal, toast)* corresponds to: vec("*meal*") - vec("*toast*"). As we can see, instances with yellow labels are clustered in the same group, e.g., belong to the *Portion-Of* subtype. It demonstrates the effectiveness of our automatic seed selection method.

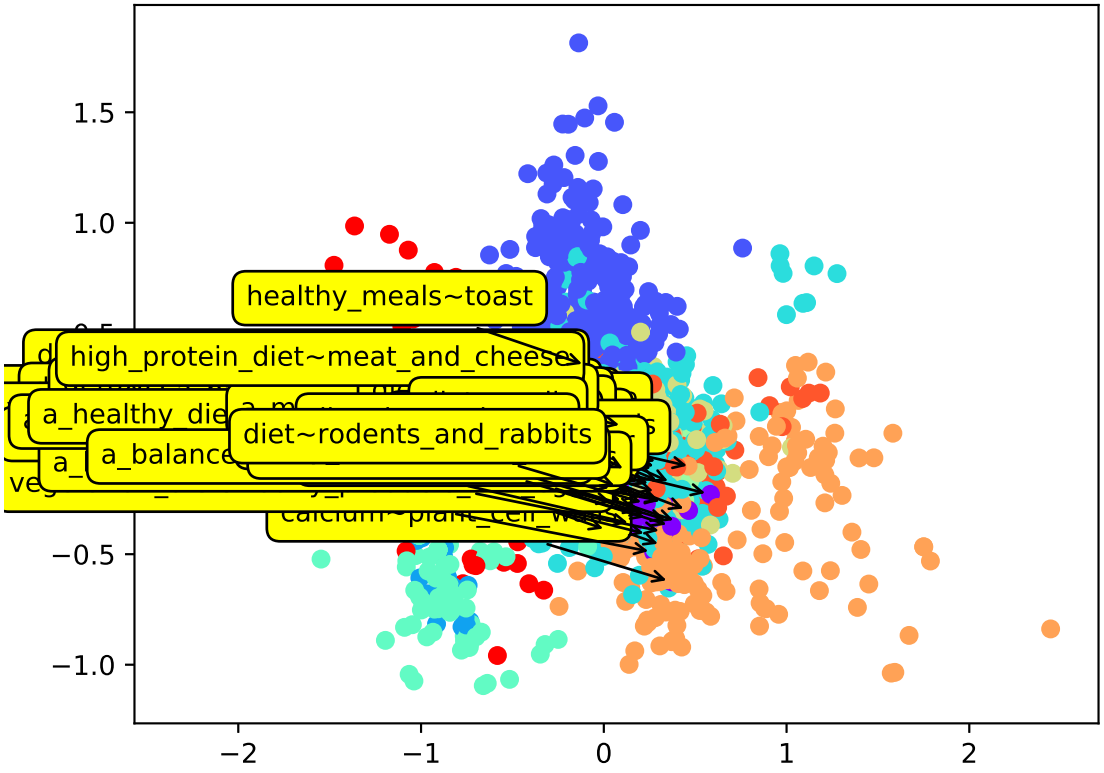

Figure 6.1: Illustration of K-means with number of clusters K=10 based on our dataset (in Section 4.4.1).

6.1.2 Case Study of Distantly Supervised Relation Extraction

Table 6.1 shows five randomly selected example results of our proposed models from the Riedel testing data. For each case, we show the gold labels and the top-3 predictions of our *2BiGRU+PATT* and *2BiGRU+PATT+Inference* models, respectively. The values appeared in parentheses represent their corresponding probabilities. The correct predictions are in boldface.

We can see that our two proposed models produce reasonable predictions in the analysis for our relation extraction task. For four of five cases (except the 4-th case), our proposed models give high probabilities to the correct predictions. The

Table 6.1: Some example results of our proposed models; correct predictions are in boldface.

A Sentence in Bag with Two Entities	Gold Labels	Top-3 Predictions of 2BiGRU+PATT (probability)	Top-3 Predictions of 2BiGRU+PATT+Inference (probability)	Unknown Words (unknown entity is in *italics)*
a 17th-century eyewitness account of the coronation of a shah , written by jean chardin , a french jeweler , is inscribed to **jean-baptiste__colbert** , then the finance minister of **france** .	/people/person/nationality	**/people/person/nationality (0.965)** NA (0.023) /people/person/place_of_birth (0.005)	**/people/person/nationality (0.978)** NA (0.011) /people/person/place_of_birth (0.007)	17th-century *jean-baptiste_colbert*
i am not apologetic about why the koran says this , said **seyyed__hossein__nasr** , an islamic scholar who teaches at **george__washington__university** .	/business/person/company	**/business/person/company (0.977)** NA (0.022) /people/person/religion (0.0004)	**/business/person/company (0.995)** NA (0.005) /people/person/religion (0.0001)	*seyyed_hossein_nasr*
on may 8 , representative marcy kaptur , an ohio democrat , and a dozen other legislators wrote to president felipe calderon of **mexico** and the governor of the state of **nuevo__leon** , of which monterrey is the capital , urging to thoroughly investigate the killing and provide protection for the rest of the mexico staff of the farm workers ' union .	/location/administrative _division/country	**/location/administrative __division/country (0.548)** NA (0.348) /people/person/nationality (0.083)	**/location/administrative __division/country (0.545)** NA (0.338) /people/person/nationality (0.095)	kaptur *nuevo_leon*
next year he is planning to publish the poetry of **aeronwy__thomas** , **dylan__thomas** 's daughter , and to bring her to the united states for a book tour along with the welsh poet and publisher peter thabit jones .	/people/person/children	NA (0.597) **/people/person/children (0.364)** /business/person/company (0.011)	NA (0.837) **/people/person/children (0.146)** /people/person/nationality (0.006)	*dylan_thomas* thabit
if they have a residence in **canada** , they can buy farmland in **saskatchewan** through the agriculture development corporation , a private company , for a minimum buy-in of $ 20,000 .	/location/location/contains & /location/country/ administrative_divisions	/**location/location/contains (0.790)** **/location/country/ administrative_divisions (0.194)** NA (0.016)	**/location/location/contains (0.658)** **/location/country/ administrative_divisions (0.336)** NA (0.005)	buy-in

contextual inference method can enhance the performance of our *2BiGRU+PATT* model with the help of supporting contexts and is useful in our task. Our *2BiGRU+PATT+Inference* model assigns comparable or higher scores to the correct predictions than the *2BiGRU+PATT* model.

In the last column of Table 6.1, we show the unknown words, which can not be found in our embedding matrix, in the corresponding sentence. The unknown entities are indicated in italics. An *unknown entity* affected significantly to the label of its bag for the short context, especially in the 4-th case. Since the is no meaningful text span between two entities *dylan_thomas* and *aeronwy_thomas*, and the 1^{st} entity's vector is missing from the embedding matrix, our models result in the second top-scoring predictions (i.e., */people/person/children*).

We checked the ratio of matched entities between the Riedel dataset and our embedding matrix. We use the word embeddings trained on the NYT corpus and keep the words which appear more than 100 times in the corpus as vocabulary. These word embeddings are similar to previous baselines [47, 48]. There are $69,040$ unique entities appeared in the Riedel dataset. However, we found that only $22,515$ of $69,040$ entities (32.61%) matched in our embedding matrix. It suggests that a larger text corpus should be used to cover the high number of entities appeared in the Riedel dataset and improve the performance of our proposed models. In addition, the vector embeddings of Wikipedia concepts and entities, such as a person's name, an organization or a place can be trained using the character embedding, which handles infrequent words better than the word embedding as the latter suffers from lack of enough training opportunity for out-of-vocabulary words.

Figure 6.2 shows similar entity pairs involved in our contextual inference method from both training and testing portions in the Riedel dataset. Recall that for each target bag (e_1, e_2) in a testing set, our contextual inference method selects top-k similar bags to (e_1, e_2) from the training set according to Eq. (21). We selected $1,000$ pairs between (e_1, e_2) and (x_1, x_2) that have highest similar-

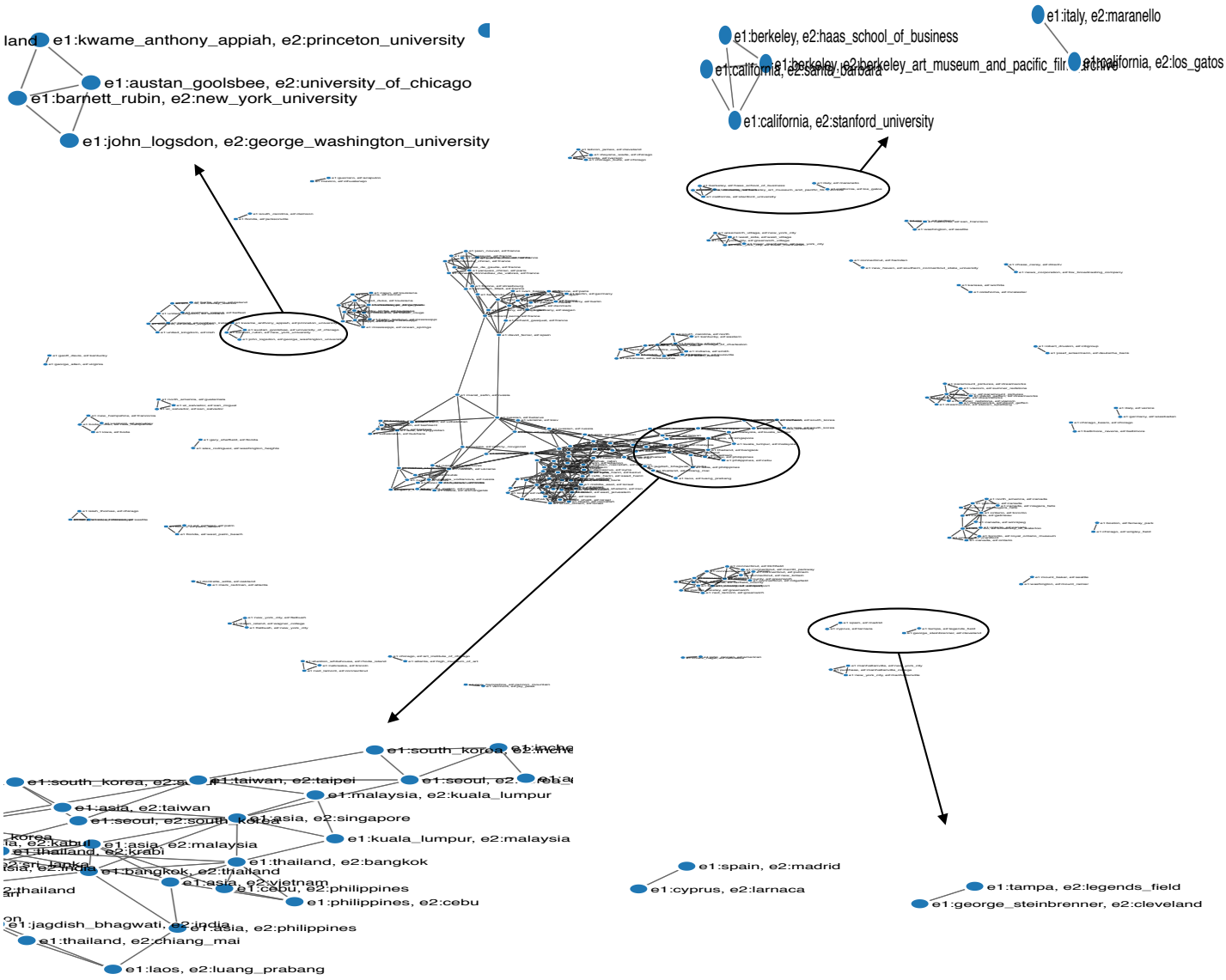

Figure 6.2: Some similar entity pairs involved in our contextual inference method; each node represents an entity pair, and similar entity pairs are linked by edges.

ity scores, and visualize these pairs using force-directed graph layout algorithms. Each entity pair (or a bag) is represented by a node, and similar entity pairs are linked by edges in the graph, which provides an overview of relationships among related bags.

In order to evaluate the quality of similar entity pairs chosen by our contextual inference method using the vector difference between entities' vectors, we randomly select 100 pairs between (e_1, e_2) and (x_1, x_2) (out of $1,000$ pairs above), and check whether these two pairs indeed have a similar semantic relationship. For example, *(atlanta, high_museum_of_art);(chicago, art_institute_of_chicago)* is assigned as correct since these two pairs are similar according to the */location/location/contains* relationship. In total, 83 out of 100 cases (83.0%) are judged as correct by two annotators. It demonstrated that using the vector difference between e_1 and e_2, and x_1 and x_2 in Eq. (21) is effective for calculating the similarity between bags. Without any external corpora or KBs, our inference method showed its advantages and leveraged the training data efficiently.

For better understanding the reason of the incorrect inference, we also analyzed each entity name in 17 incorrect cases (out of 100 cases above). For example, *(kentucky, centre_college);(mitch_mustain, arkansas)* is an incorrect example, where *mitch_mustain* is a person name, and others are locations or places. We found that 13 out of 17 incorrect cases (76.5%) contain at least one person name, while only 22 out of 83 correct cases (26.5%) have such entity type. It indicates that learning meaningful vector representations for person names is more difficult than for others. In the future work, we think that much efforts should be done to obtain better embeddings of rare entity names, such as the person names in the Riedel dataset.

Due to the diversity of relation types and limitations of model capabilities, we think that a small number of incorrect predictions are inevitable. In general, our proposed methods are very effective for improving the performance of the distantly supervised relation extraction systems.

6.1.3 Issue of Long-Tailed and Imbalanced Data

Extracting long-tail relations from the automatically labeled data, in which the number of training examples per class varies significantly from hundreds or thousands for head classes to as few as one for tail classes, is still a challenging problem even in big data [63]. We observe that 77.63% entity pairs have only one relation mention [64], and nearly 70% of the relations are long-tail in the Riedel dataset [17]. To overcome this issue, a relation extraction model needs to learn accurately for classes existing at the tail of the class distribution, for which only little data is available.

One possible solution for long-tail relation extraction is that, we can leverage the knowledge from data-rich classes at the head of the distribution to boost the performance of the data-poor classes at the tail [63, 65]. For example, we can transfer the knowledge from the head relation */people/deceased_person/place_of_death* to the long-tail relation */people/decease_person/place_of_burial*. These two relation are semantically similar, and can share common lexical or syntactic patterns in several relation mentions. However, this solution is domain-dependent and requires human effort to examine the characteristics of several target relations.

To deal with long tail relations, we presented a contextual inference method that infers the most likely positive examples of an entity pair in bags with limited contextual information without using any external resources for the distantly supervised relation extraction task. The experimental results showed that our proposed *2BiGRU+PATT+Inference* model achieved significant improvements over state-of-the-art methods.

6.1.4 Contribution of Two Annotated Datasets

The resources for the relation extraction task are limited. In this dissertation, we contribute two datasets, one for bootstrapping relation extraction, and the other for distantly supervised relation extraction.

First, we provided an annotated dataset of part-whole relations as a reliable resource for selecting seeds. To the best of our knowledge, there are no datasets available for all fine-grained subtypes of the part-whole relation, which is one of the most fundamental ontological relations, so far. Second, we provided an annotated dataset to guarantee the quality of the distant supervision testing data, and report on the actual performance of various relation extraction systems.

Both of the datasets above are made publicly available for other researchers to use as benchmarks in the relation extraction task. They can also be used for training end-to-end supervised relation extractions.

6.2 Conclusion

In this study, we focused on two weakly supervised approaches, namely *bootstrapping* and *distantly supervised* relation extraction methods, which significantly reduce the expensive cost for data labeling and human effort required, especially in supervised learning.

The first part of this dissertation addressed the subtasks of automatic seed selection for bootstrapping relation extraction, and noise reduction for distantly supervised relation extraction. We formulated the two subtasks as ranking problems, and proposed novel methods that can be applied for both of them. Our methods are inspired by ranking instances and patterns computed by the HITS algorithm, and selecting cluster centroids using K-means, latent semantic analy-

sis (LSA), or the non-negative matrix factorization (NMF) method. Experiments showed that our proposed methods achieved a better performance than the baseline systems.

The second part of this dissertation investigated distant supervision, and introduced a novel neural model that combines a bidirectional gated recurrent unit model with the piecewise attention, which significantly enhances the performance of the distantly supervised relation extraction task. In addition, we proposed a contextual inference method that can infer the most likely positive examples in bags with very limited contextual information. The experimental results showed that our proposed methods outperformed state-of-the-art baselines on benchmark datasets.

6.3 Future Work

In the future work, we plan to apply our *Espresso+Word2vec* bootstrapping system, as well as the automatic seed selection methods to the biomedical domain, for supporting the binary relation extraction task.

For the distantly supervised relation extraction task, we plan to develop more sophisticated methods for measuring the similarity between entity-pair bags, such as using the shortest dependency path between the two entities instead of the full sentence to infer similar examples from external text corpora, and apply our methods to other domains such as biomedical or scientific articles in order to further benefit this task. In addition, we will use our proposed models to explore other distant supervision settings, e.g., distant supervision with temporal reasoning (predict relations at any specific time spot) [66], or distant supervision in the document level [67, 68].

6.4 Closing Remark

Relation extraction is a challenging task in the information extraction field. It is hoped that the research in this dissertation will help in automating relation extraction systems in an effective and efficient way, leading to more precise, more broadly applicable and faster approaches in the future.

List of Major Publications

- Van-Thuy Phi, Joan Santoso, Masashi Shimbo, and Yuji Matsumoto. Ranking-Based Automatic Seed Selection and Noise Reduction for Weakly Supervised Relation Extraction. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), pp. 89-95, July 16th, 2018.

- Van-Thuy Phi, Joan Santoso, Van-Hien Tran, Hiroyuki Shindo, Masashi Shimbo, and Yuji Matsumoto. Distant Supervision for Relation Extraction via Piecewise Attention and Bag-Level Contextual Inference. IEEE Access 7 (2019). DOI: 10.1109/ACCESS.2019.2932041

- Van-Hien Tran, Van-Thuy Phi, Hiroyuki Shindo, and Yuji Matsumoto. Relation Classification Using Segment-Level Attention-based CNN and Dependency-based RNN. In Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), pp. 2793-2798, June 4th, 2019.

- Van-Thuy Phi, and Yuji Matsumoto. Integrating Word Embedding Offsets into the Espresso System for Part-Whole Relation Extraction. In Proceedings of The 30th Pacific Asia Conference on Language, Information and Computation (PACLIC 30), pp. 173-181, October 30th, 2016.

References

[1] Ellen Riloff, Rosie Jones, et al. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference (AAAI/IAAI)*, pages 474–479, 1999.

[2] Sergey Brin. Extracting patterns and relations from the world wide web. In *International Workshop on the World Wide Web and Databases*, pages 172–183. Springer, 1998.

[3] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM Conference on Digital Libraries*, pages 85–94. ACM, 2000.

[4] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011. Association for Computational Linguistics, 2009.

[5] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM, 2008.

[6] Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. Introducing wikidata to the linked data web. In *International Semantic Web Conference*, pages 50–65. Springer, 2014.

[7] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledge base. 2014.

[8] Sunita Sarawagi et al. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377, 2008.

[9] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 113–120. Association for Computational Linguistics, 2006.

[10] Ashwin Ittoo and Gosse Bouma. Minimally-supervised extraction of domain-specific part–whole relations using wikipedia as knowledge-base. *Data & Knowledge Engineering*, 85:57–79, 2013.

[11] Zornitsa Kozareva and Eduard Hovy. Not all seeds are equal: Measuring the quality of text mining seeds. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 618–626. Association for Computational Linguistics, 2010.

[12] Jason Eisner and Damianos Karakos. Bootstrapping without the boot. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.

[13] Tetsuo Kiso, Masashi Shimbo, Mamoru Komachi, and Yuji Matsumoto. HITS-based seed selection and stop list construction for bootstrapping. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 30–36. Association for Computational Linguistics, 2011.

[14] Dana Movshovitz-Attias and William W. Cohen. Bootstrapping biomedical ontologies for scientific text using nell. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 11–19. Association for Computational Linguistics, 2012.

[15] Van-Thuy Phi and Yuji Matsumoto. Integrating word embedding offsets into the espresso system for part-whole relation extraction. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pages 173–181, 2016.

[16] Patrick Pantel and Deepak Ravichandran. Automatically labeling semantic classes. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, 2004.

[17] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 Joint European Conference on Machine Learning and Principles of Knowledge Discovery in Databases (ECML PKDD)*, pages 148–163. Springer, 2010.

[18] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 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-Volume 1*, pages 541–550. Association for Computational Linguistics, 2011.

[19] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 455–465. Association for Computational Linguistics, 2012.

[20] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[21] Nguyen Bach and Sameer Badaskar. A survey on relation extraction. *Language Technologies Institute, Carnegie Mellon University*, 2007.

[22] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics*, volume 2, pages 539–545. Association for Computational Linguistics, 1992.

[23] Dmitry Davidov, Ari Rappoport, and Moshe Koppel. Fully unsupervised discovery of concept-specific relationships by web mining. 2007.

[24] Zornitsa Kozareva, Ellen Riloff, and Eduard H Hovy. Semantic class learning from the web with hyponym pattern linkage graphs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1048–1056, 2008.

[25] Mamoru Komachi, Taku Kudo, Masashi Shimbo, and Yuji Matsumoto. Graph-based analysis of semantic drift in Espresso-like bootstrapping algorithms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1011–1020. Association for Computational Linguistics, 2008.

[26] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 113–120. Association for Computational Linguistics, 2006.

[27] Mamoru Komachi, Taku Kudo, Masashi Shimbo, and Yuji Matsumoto. Graph-based analysis of semantic drift in espresso-like bootstrapping algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1011–1020. Association for Computational Linguistics, 2008.

[28] Partha Pratim Talukdar and Fernando Pereira. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1473–1481. Association for Computational Linguistics, 2010.

[29] Zornitsa Kozareva, Konstantin Voevodski, and Shang-Hua Teng. Class label enhancement via related instances. In *Proceedings of the conference on empirical methods in natural language processing*, pages 118–128. Association for Computational Linguistics, 2011.

[30] Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. Understanding seed selection in bootstrapping. In *Proceedings of the TextGraphs-8 Workshop*, pages 44–52, 2013.

[31] Ander Intxaurrondo, Mihai Surdeanu, Oier Lopez de Lacalle, and Eneko Agirre. Removing noisy mentions for distant supervision. *Procesamiento del lenguaje natural*, 51, 2013.

[32] Yang Xiang, Qingcai Chen, Xiaolong Wang, and Yang Qin. Distant supervision for relation extraction with ranking-based methods. *Entropy*, 18(6):204, 2016.

[33] Gang Li, Cathy Wu, and K. Vijay-Shanker. Noise reduction methods for distantly supervised biomedical relation extraction. In *SIGBioMed Workshop on Biomedical Natural Language Processing (BioNLP '17)*, pages 184–193. Association for Computational Linguistics, 2017.

[34] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics, 2011.

[35] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013.

[36] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[37] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391, 1990.

[38] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[39] Daniel D Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[40] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.

[41] Rafal Zdunek and Andrzej Cichocki. Fast nonnegative matrix factorization algorithms using projected gradient approaches for large-scale problems. *Computational Intelligence and Neuroscience*, 2008:3, 2008.

[42] Madelyn Anne Iris. Problems of the part-whole relation. In *Relational Models of the Lexicon*, pages 261–288. Cambridge University Press, 1989.

[43] Morton E. Winston, Roger Chaffin, and Douglas Herrmann. A taxonomy of part-whole relations. *Cognitive Science*, 11(4):417–444, 1987.

[44] Van-Thuy Phi and Yuji Matsumoto. Integrating word embedding offsets into the Espresso system for part-whole relation extraction. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pages 173–181, 2016.

[45] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[46] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics:*

Technical Papers, pages 2335–2344. Dublin City University and Association for Computational Linguistics, 2014.

[47] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762. Association for Computational Linguistics, 2015.

[48] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133. Association for Computational Linguistics, 2016.

[49] Mark Craven, Johan Kumlien, et al. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86, 1999.

[50] Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D Manning. Combining distant and partial supervision for relation extraction. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1556–1567, 2014.

[51] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, 2014.

[52] Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *AAAI*, pages 3060–3066, 2017.

[53] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks

for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 207–212, 2016.

[54] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics, 2009.

[55] Linyi Yang, Tin Lok James Ng, Catherine Mooney, and Ruihai Dong. Multi-level attention-based neural networks for distant supervised relation extraction. In *25th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland, 7-8 December 2017*. Insight Centre, 2017.

[56] Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. Improving distantly supervised relation extraction using word and entity based attention. *arXiv preprint arXiv:1804.06987*, 2018.

[57] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[58] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[59] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[60] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[61] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[62] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[63] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017.

[64] Kai Lei, Daoyuan Chen, Yaliang Li, Nan Du, Min Yang, Wei Fan, and Ying Shen. Cooperative denoising for distantly supervised relation extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 426–436, 2018.

[65] Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. *arXiv preprint arXiv:1903.01306*, 2019.

[66] Jianhao Yan, Lin He, Ruqin Huang, Jian Li, and Ying Liu. Relation extraction with temporal reasoning based on memory augmented distant supervision. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1019–1030, 2019.

[67] Chris Quirk and Hoifung Poon. Distant supervision for relation extraction beyond the sentence boundary. *arXiv preprint arXiv:1609.04873*, 2016.

[68] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. Docred: A large-scale

document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*, 2019.