Ph.D. DISSERTATION

Knowledge Extraction and Integration for Knowledge Base Construction using Machine Learning

지식베이스 구축을 위한
머신러닝 기반 지식 추출 및 통합

BY

Jung, Woohwan

Feburary 2021

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Ph.D. DISSERTATION

Knowledge Extraction and Integration for Knowledge Base Construction using Machine Learning

지식베이스 구축을 위한
머신러닝 기반 지식 추출 및 통합

BY

Jung, Woohwan

Feburary 2021

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Knowledge Extraction and Integration for Knowledge Base Construction using Machine Learning

지식베이스 구축을 위한 머신러닝 기반 지식 추출 및 통합

지도교수 심 규 석
이 논문을 공학박사 학위논문으로 제출함

2021년 2월

서울대학교 대학원

전기 컴퓨터 공학부

정 우 환

정우환의 공학박사 학위 논문을 인준함

2021년 2월

위 원 장:	김 형 주	(인)
부위원장:	심 규 석	(인)
위 원:	홍 성 수	(인)
위 원:	정 교 민	(인)
위 원:	김 영 훈	(인)

Abstract

Knowledge bases have been successfully applied to many real world applications such as question answering, recommender system and natural language understanding. However, building a large knowledge base using human annotations takes a lot of time, effort, and money. Moreover, it is almost impossible to manually update a large amount of newly created relational facts in a timely manner. Accordingly, automated knowledge base construction has attracted a lot of attention over the last decade.

Knowledge fusion is a method to automatically construct knowledge bases from the entire web. It first extracts information from many web pages by using multiple relation extractors. Since the collected information is usually noisy due to extraction errors, knowledge fusion next identifies the correct information by using truth discovery techniques and appends the new information to the knowledge base. In this dissertation, we focus on extending the coverage and improving the accuracy of the knowledge fusion process.

With the development of deep learning, many recent works on relation extraction make use of deep learning techniques to improve accuracy. Since neural relation extraction models require a large amount of training data, they usually use distant supervision which automatically generates training data by assuming that if a relation between a pair of entities exists in a knowledge base, all sentences that contain these two entities express this relation. However, distant supervision inevitably suffers from the wrong labeling problem which degrades the accuracy of relation extraction. We develop a method to effectively train relation extraction models by also using human annotated data to improve their accuracy.

To extend the coverage of relation extraction, we also investigate the problem of extracting information about the topic entity. The topic entity of a document is the entity that is mainly described in the document. Since the topic entity is often missing

from some sentences, existing relation extraction models often fail to find the relations with the omitted topic entity. To extract those relations, we propose a topic-aware relation extraction model.

After extracting the relations from web pages, a truth discovery algorithm resolves the conflicts in the extracted information and identifies the correct information. Existing works on truth discovery usually assumed that claimed values are mutually exclusive and only one among them is correct. However, many claimed values are not mutually exclusive due to their hierarchical structures and so we need to take account of the hierarchical structure to infer the truths. We propose a probabilistic model that infers the truth by considering the hierarchical structures for the claimed values. Nevertheless, if many relation extractors generate similar errors, some of the errors might not be corrected by unsupervised truth discovery algorithms. Thus, we take advantage of human cognitive abilities by crowdsourcing the refinement of extracted information. We present a task assignment algorithm to optimize accuracy improvement given the constraint of a fixed budget for crowdsourcing.

keywords: Knowledge base, knowledge fusion, relation extraction, truth discovery, crowdsourcing

student number: 2012-20862

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A knowledge base is an online database that consists of a set of entities and the relations between the entities. In knowledge bases, the information is stored in the form of $\langle head\_entity, relation, tail\_entity \rangle$ triples. For example, the fact that 'Statue of Liberty is located in New York' is represented by $\langle Statue\_of\_Liberty, located\_in, New\_York \rangle$. Knowledge bases have been successfully applied to many real word applications such as question answering [80, 69], recommender system [22, 84] and natural language understanding [50, 86]. However, manually building a large knowledge base takes a lot of time, effort, and money. Moreover, it is almost impossible to manually update a large amount of newly created relational facts in a timely manner. Accordingly, automated knowledge base construction has been attracted a lot of attention from both industry and academia over the last decade.

YAGO [62] and DBpedia [37] are popular knowledge bases which are automatically constructed. However, since they mainly extract triples from Wikipedia pages by exploiting website-specific information extractors, the coverage is limited. Knowledge fusion is a method to automatically construct knowledge bases from the entire web and has been extensively investigated in [15, 65, 42, 13, 17]. As shown in Figure 1.1, knowledge fusion collects the triples through two steps: relation extraction and truth discovery. In the first step, multiple relation extractors identify triples from

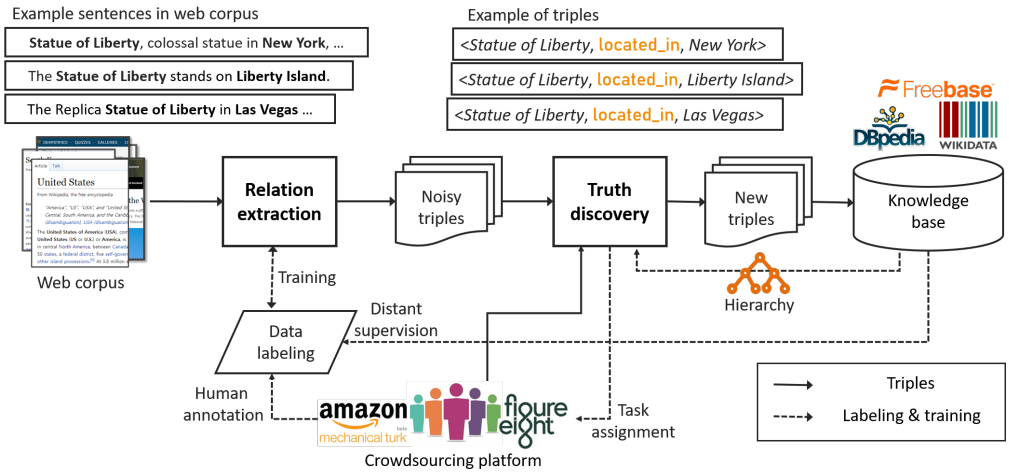


Figure 1.1: Knowledge fusion process

web pages. Due to extraction error or wrong information provided by web sources, the extracted triples can be erroneous. In the second step, a truth discovery algorithm finds the correct triples from the extracted triples with assessing and exploiting the reliabilities of information sources. For example, two conflicting triples $\langle Statue\_of\_Liberty, located\_in, New\_York\rangle$ and $\langle Statue\_of\_Liberty, located\_in, Las\_Vegas\rangle$ can be simultaneously extracted from different web pages or by different extractors. In this case, knowledge fusion identifies the correct triples by using truth discovery techniques and append the correct triples to the knowledge base.

In this dissertation, we focus on extending the coverage and improving the accuracy of the knowledge fusion process. Specifically, we present machine learning techniques for relation extraction and truth discovery in knowledge fusion.

With the development of deep learning, neural networks have become major tools for relation extraction [47, 83, 67, 6, 73]. The previous studies use deep neural networks to find relations without handcrafted features. In [83] and [82], convolution neural networks (CNN) [7] are used to encode the text and the entity pairs that we want to find the relations. Some of the previous works [6, 67] exploit recurrent neural networks (RNNs) [26]. More recently, [73] proposed a fine-tuned model of BERT for

document-level relation extraction. Unlike the previous works that propose standalone relation extraction models, we mainly focus on the knowledge fusion process which utilizes multiple relation extraction models. Thus, we first propose a method to effectively train existing relation extraction models regardless of the model architecture. In addition, we present a new relation extraction model to find the triples that can be easily missed by existing relation extraction models.

Since neural relation extraction models usually require a large amount of training data, distant supervision [56] is used to automatically generate labeled training data. Distant supervision assumes that if a text contains an entity pair which has a relation in a knowledge base, the text actually expresses the relation in the knowledge base. Thus, there exists a labeling bias in distant supervision. For example, many cities have their 'sister cities'. Distant supervision produces the label $\langle e_h, sister\_city, e_t \rangle$ when the two cities e_h and e_t appear in the same text. However, 'sister city' relation rarely appears in human-annotated labels because the relationship is not frequently expressed in the text data. In a real dataset DocRED [77], we observed that the 'sister city' relation is annotated 86 times more frequently by distant supervision compared to human annotators. It implies that most of the labels annotated by distantly supervision is incorrect for some relations. The labeling bias can substantially degrade the accuracy of relation extraction even though we utilize human annotations and distant supervision together. We propose a new method to effectively train relation extraction models with distant supervision and human annotation by considering the labeling bias.

To extend the coverage of relation extraction, we study the problem of extracting relation with the topic entity. The topic entity of a document is the entity that is mainly described in the document. In many sentences, the topic entity is replaced by a pronoun or even omitted because people can infer it from context. In this case, existing models often fail to extract the relation with the omitted topic entity. Consider the following example sentence in the wikipedia article for the United States:

With a population of over 328 million, it is the third most populous

country in the world.

Without considering the topic entity 'United States' and the other sentences, we cannot extract the triple $\langle United\_States, population, 328\ million \rangle$ from the above sentence. We propose a topic-aware relation extraction model to extract relations even when the topic entity is omitted.

After extracting the relations from web pages, a truth discovery algorithm resolves the conflicts in the extracted information and identifies the correct information. Since the information sources (web pages and relation extractors) have different accuracies, truth discovery algorithms assess the reliabilities of sources to find the correct triple among the conflicting triples. Hierarchical structures in entities can help us find the correct triples and accurately estimate the reliabilities of sources. Suppose that the relation extractors retrieved the three triples in Figure 1.1. Considering the fact that Liberty Island is an island in New York, 'New York' and 'Liberty Island' are not conflicting with each other. In this case, the geographical hierarchy provides strong evidence to support that the Statue of Liberty stands on Liberty Island, New York. Moreover, if we do not consider the hierarchical structure, we may underestimate the reliability of some sources because only one of 'New York' and 'Liberty Island' is regarded as a correct value for the location of the Statue of Liberty. In addition, we observed that information sources have different generalization tendencies as well as different accuracy. Some sources usually provide more generalized values (e.g., countries and continents) and other sources provide more specific values (e.g., cities or more specific locations). Although many entities are hierarchically structured, most of the existing works on truth discovery do not consider the hierarchical structure. We propose a novel truth discovery model that considers the different generalization tendencies of sources.

Knowledge fusion is an error-prone process because it extracts and integrates information from various types of web pages. According to [15], up to 96% of the false triples are made by extractors rather than by the information sources. Such extraction errors can be easily corrected by manually checking the information sources. To reduce

the noise in the triples to be appended to knowledge bases, we utilize human resources through crowdsourcing. Since the budget for crowdsourcing is limited, we study the problem of maximizing the accuracy improvement with a budget constraint. We propose a method to estimate the expected accuracy improvement from a task as well as an efficient algorithm to assign the tasks which are expected to increase accuracy the most.

1.1 Contributions of This Dissertation

Contributions of this dissertation are as follows:

- We first introduce and study a new problem named topic-aware relation extraction. We propose the *T-REX* model to find the relations with the topic entity even when the topic entity is omitted in some sentences.

- We present a new framework to effectively train relation extraction models with both human annotation and distant supervision. We analyze the labeling bias of the distant supervision which can significantly deteriorate the accuracy of relation extraction. To deal with the labeling bias, we propose a new structure for the output layer of relation extraction models that prevents overfitting to the noisy labels obtained from distant supervision. In addition, we propose a loss function based on the analysis of the labeling bias.

- We next investigate the problem of truth discovery in the presence of hierarchies. We point out that information sources on the web have different generalization tendencies as well as different reliabilities. We propose a truth discovery algorithm utilizing the hierarchical structures in the extracted values. To the best of our knowledge, it is the first work that assesses both the reliabilities and the generalization tendencies of the sources.

- We finally study the problem of assigning tasks in crowdsourcing platforms. To

assign a task that will most improve the accuracy, we develop an incremental EM algorithm to estimate the accuracy improvement for a task. We also devise an efficient task assignment algorithm to enable the interactive crowdsourcing with low latency.

Although the relation extraction and truth discovery are used in many applications other than the knowledge fusion, the techniques presented in this dissertation are mainly devised for the knoweldge fusion. For example, our topic-aware relation extraction model is not directly applicable to question answering, which is a well-known application of relation extraction, because it does not extract all relations in the document. However, it can be a useful tool for the knowledge fusion since it identifies many relations which can be easily missed by the other relation extraction models in the knowledge fusion process. Moreover, we utilize crowdsourcing for truth discovery since the extraction errors can be easily corrected by human workers. In this dissertation, we propose several techniques to effectively and efficiently utilize existing knowledge bases and human resources for knowledge fusion. In the relation extraction task, we propose a method to train relation extraction models with both human annotated data and distantly supervised data which is labeled by using knowledge bases. To improve the process of truth discovery, we utilize the hierarchy of entities obtained from knowledge bases. In addition, we devise a task assignment algorithm to efficiently utilize crowdsourcing.

Automated knowledge base construction is a complex process that requires many techniques such as semantic parsing, entity linking, relation extraction, truth discovery, entity resolution and schema alignment. In this dissertation, we focus on two main sub procedures to find new relational facts to be appended to knowledge bases: relation extraction and truth discovery. For example, in the relation extraction task, we assume that entity mentions in text are already annotated by entity recognition and linking techniques such as [24, 36, 52]. Note that most chapters of this dissertation have been published in peer-reviewed papers [28, 29, 30].

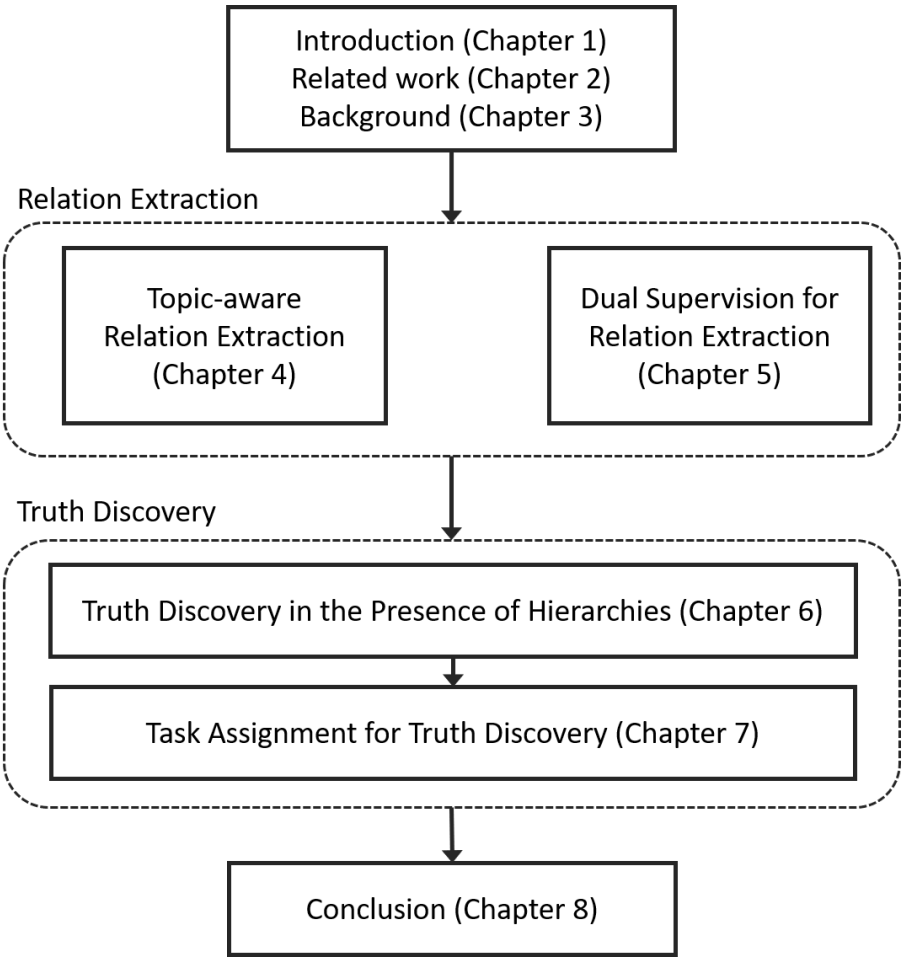


Figure 1.2: Overview of this dissertation

1.2 Overview of This Dissertation

The overview of this dissertation is summarized in Figure 1.2. In the next chapter, we review the related works on automated knowledge base construction. In Chapter 3, we provide important definitions and technical backgrounds. The remaining part of this dissertation is organized into two main parts: 1) relation extraction (Chapters 4-5), and 2) truth discovery (Chapters 6-7). In Chapter 4, we propose a topic-aware relation extraction model. Chapter 5 presents the dual supervision framework for training relation extraction models with distant supervision and human annotation. Chapter 6 proposes

a truth discovery algorithm which utilizes hierarchical structures in extracted values. In Chapter 7, we introduce an algorithm for the efficient assignment of tasks to workers in crowdsourcing platforms. We finally present our conclusions in Chapter 8.

Chapter 2

Related Work

2.1 Knowledge Base Construction

The researches on automated knowledge base construction can be classified into 2 groups. YAGO [62] and DBpedia [37] are popular knowledge bases which belongs to the first group. They mainly extract triples from Wikipedia pages using a predefined format. For example, they utilize infoboxes to extract the properties of entities. This approach provides accurate extraction results. However, the coverage is limited because other websites have different formats. The works [15, 65, 42, 13, 17] in the second group exploits relation extraction and truth discovery (data fusion) techniques to construct knowledge bases from the entire web. Our work also belongs to the second group. In the rest of this chapter, we review the related works on relation extraction and truth discovery.

2.2 Relation Extraction

We briefly survey the existing works for relation extraction (RE). [56] propose distant supervision to overcome the limitation of the quantity of human-annotated labels. They utilize lexical, syntactic and named entity tag features obtained by existing NLP tools

to extract relations. Other early works in [64, 27] also utilized hand-crafted features to find the relations in text. However, since such RE models take the input features from NLP tools, the errors generated by the NLP tools are propagated to the RE models. In order to deal with the error propagation, the works [47, 83, 82, 67, 73] use deep neural networks such as CNN, LSTM and BERT instead of handcrafted features to encode the text for finding the relations. Since many relational facts are expressed across multiple sentences, the recent works [77, 73] studied document-level RE. [77] provide a document-level RE dataset (DocRED) as well as compare the models adapted from the sentence-level RE models [83, 26, 6, 67]. Moreover, a fine tuned model [73] of BERT [12] for document-level RE achieved a higher F1 score than the baselines on DocRED. The works on document-level RE do not consider the topic entity since they want to extract all relations including the relations between non-topic entities. Relation extraction with the topic entity of the document has been addressed in previous works for semi-structured data such as HTML web pages[8, 51, 76]. However, since they utilize HTML tags and DOM tree structures, those works are not directly applicable to relation extraction from plain text.

The wrong labeling problem in distant supervision has been addressed in many previous works [82, 47, 79, 3]. Some of the works [82, 47, 79] build a bag-of-sentences for a pair of entities and extract relational facts from the bag-of-sentences with attention over the sentences. [3] propose a bag-of-sentences-level model which utilizes human annotation. However, they use the human annotated labels only to determine whether there exist a relationship or not since the labels are obtained from a different domain. The goal of these works is different from ours which is to find the relations *appearing in a given text* (e.g., a document). Thus, the bag-of-sentences-level models have a limitation to be used for some applications such as question answering.

The most relevant work to our dual supervision is [78]. This paper proposes the bias adjustment methods to utilize a small amount of human annotated data to improve RE models trained on distantly supervised data by considering the different distribu-

tion of human annotated labels and distantly supervised labels. However, they do not use human annotated data to train the models and use the HA data only to obtain a statistic to be used the determine the size of the bias adjustment. Thus, the bias adjustment methods cannot consider contextual information. In addition, since the models are trained only on distantly supervised data, the performance improvement is limited although many human annotated labels are available.

2.3 Truth Discovery

The problem of resolving conflicts from multiple sources (i.e., truth discovery) has been extensively studied in [11, 14, 16, 72, 61, 41, 43, 45, 81, 87, 89, 88, 46, 11, 75, 10, 92, 32]. Truth discovery for categorical data has been addressed in [11, 14, 16, 61, 41, 89, 43, 81]. According to a recent survey [90], LFC [61] and CRH [41] perform the best in an extensive experiment with the truth discovery algorithms [90, 40, 11, 75, 10, 92, 32]. There exist other interesting algorithms [14, 16, 43, 89] which are not evaluated together in [90]. Accu [14] and PopAccu [16] combine the conflicting values extracted from different sources for the knowledge fusion process in [15]. They consider the dependencies between data sources to penalize the copiers' claims. DOCS[89] utilizes the domain information to consider the different levels of worker expertises on various domains. MDC[43] is a truth discovery algorithm devised for crowdsourcing-based medical diagnosis. The works in [45, 72, 87] studied how to resolve conflicts in numerical data from multiple sources.

The truth discovery algorithms in [87, 88, 58, 89] are based on probabilistic models. Resolving the conflicts in numerical data is addressed in [87] and discovering multiple truths for an object is studied in [88]. Probabilistic models for finding a single truth for each object is proposed in [58, 89]. However, none of those algorithms exploit the hierarchical relationships of claimed values for truth discovery. A previous work ASUMS [4] adopts an existing algorithm to consider hierarchical relationships. To find

the true value for each object, it greedily traverses down the hierarchy tree from the root until the confidence on the node is higher than the given thresholds θ. However, it requires a threshold to control the granularity of the inferred truth. In addition, since ASUMS uses existing method to evaluate the quality of sources, it does consider the different generalization tendencies of information sources. On the contrary, our proposed truth discovery algorithm considers the generalization tendencies of sources and automatically finds the truth without any given threshold.

Task assignment algorithms [5, 25, 91, 54, 89, 20] for crowdsourcing have been studied widely in recent years. The works in [5, 91, 89] can be applied to our crowdsourced truth discovery. For task assignment, AskIt [5] selects the most uncertain object for a worker. Meanwhile, the task assignment algorithm in [89] selects the object which is expected to decrease the entropy of the confidence the most. QASCA [91] chooses an object which is likely to most increase the accuracy. Since QASCA outperforms AskIt in the experiments presented in [91, 89], we do not consider AskIt in our experiments. In [25], task assignment for binary classification was investigated but it is not applicable to our problem to find the correct value among multiple conflicting values. Meanwhile, the task assignment algorithm is proposed in [54] for the case when the required skills for each task and the skill set of every worker is available. However, it is not applicable to our problem. A task assignment algorithm proposed in [20] assigns every object to a fixed number of workers. However, since we already have claimed values from sources, we do not have to assign all objects to workers.

Chapter 3

Background

Knowledge base is a repository of information that contains a set of entities and the relations between the entities. In the rest of this dissertation, we study two main problems of automated knowledge base construction: relation extraction and truth discovery. In this chapter, we introduce the notations and problem definitions to be used in this dissertation.

3.1 Relation Extraction

Relation extraction is a task to identify the semantic relationships between entities from unstructured data (e.g., text) semi-structured data (e.g., html). In this dissertation, we focus on relation extraction from text data. We assume that the set of relation type R is given in advance and each text is annotated with entity mentions. Relation extraction from text is classified into two categories according to the type of text: sentence-level relation extraction and document-level relation extraction. We next formally define the problems of the sentence-level relation extraction and document-level relation extraction.

For a pair of entities, since a sentence usually describes a single relationship between them, the sentence-level relation extraction is generally regarded as a *multi-class*

classification problem.

Definition 1 (**Sentence-level relation extraction**). *For a pair of the head and tail entities e_h and e_t, a relation type set R and a sentence s annotated with entity mentions, we determine the relation $r \in R$ between e_h and e_t in the sentence. Note that R includes a special relation type* `NA` *which indicates that there does not exist any relation between e_h and e_t.*

Since multiple relationships between a pair of entities can be expressed in a document, document-level relation extraction is usually defined as a *multi-label* classification problem.

Definition 2 (**Document-level relation extraction**). *For a pair of the head and tail entities e_h and e_t, a relation type set R and a document d annotated with entity mentions, we find the set of all relations $R^* \subset R$ between e_h and e_t appearing in document d. Note that R does not include* `NA` *in this case since it can be represented by an empty set of R^*.*

We present the dual supervision framework which can improve both sentence-level and document-level relation extraction models in Chapter 5.

We also introduce a new problem named topic aware relation extraction which is a special case of the document-level relation extraction. The topic entity of a document is the entity which is mainly described in the document. For example, the title of each Wikipedia article usually is a topic entity of the article. The goal of topic-aware RE is to find all distinct relations between the topic entity and the other entities that expressed in a document. The problem of topic-aware RE is formally defined as follows:

Definition 3 (Topic-aware relation extraction). *Let R be a set of relation types and d be a document with annotated mentions of an entity set E. Given a document d and its the topic entity $e_{topic} \in E$, we find all triples $X = \{x_1, x_2, ..., x_{|X|}\}$ from the document where $x_i = \langle h_i, r_i, t_i \rangle$, $r_i \in R$, $s_i \in E$, $o_i \in E$, and either h_i or t_i is the topic entity e_{topic}.*

We provide the topic-aware relation extraction model in Chapter 4.

3.2 Truth Discovery

Truth discovery is an unsupervised learning problem to resolve conflicts of noisy values. We study truth discovery problem to resolve the conflicts between triples extracted from multiples sources by multiple relation extractors. As discussed in Chapter 1, we utilize a hierarchy and crowdsourcing to improve the performance of truth discovery. In this section, we provide the definitions and the problem formulation of *crowdsourced truth discovery in the presence of hierarchy*.

Definitions and notations. A triple describe a particular attribute value of an object (i.e., entity). For example, a triple $\langle United\_States, capital, Washington\_D.C.\rangle$ represent the capital of the United States. For a pair of triples that have the same head entity and relation, we say that the triples are conflicting if the tail entities are different (e.g., $\langle United_States, capital, New\_York\rangle$ and $\langle United\_States, capital, Washington\_D.C.\rangle$). Truth discovery algorithms find the correct value among the candidate attribute values 'New York' and 'Washington D.C.'. For the ease of presentation, we assume that we are interested in a single attribute of entities although our algorithms can be easily generalized to find the truths of multiple attributes. Thus, we use 'the target attribute value of an object' and 'the value of an object' interchangeably.

A *source* is a structured or unstructured database which contains the information on target attribute values for a set of objects. In knowledge fusion process, each (extractor, web page) pair or (extractor, website) pair is regarded as a source. In this dissertation, we call a human worker in crowdsourcing platforms as a *worker* to avoid the confusion with the web information sources. The information of an object provided by a source or a worker is called a *claimed value*.

Definition 4. *A* record *is a data describing the information about an object from a source. A record on an object o from a source s is represented as a triple* (o, s, v_o^s)

Symbol	Description
s	A data source
w	A crowd worker
v_o^s	Claimed value from s about o
v_o^w	Claimed value from w about o
R	Set of all records collected from the set of sources S
A	Set of all answers collected from the set of workers W
V_o	Set of candidate values about o
S_o	Set of sources which post information about o
W_o	Set of workers who answered about o
O_s	Set of objects that source s provided a value
O_w	Set of objects that worker w answered to
$G_o(v)$	Set of values in V_o which are ancestors of a value v except the root in the hierarchy H
$D_o(v)$	Set of values in V_o which are descendants of v

Table 3.1: Notations

where v_o^s is the claimed value of an object o collected from s. Similarly, if a worker w answers that the truth on an object o is v_o^w, the answer *is represented as* (o, w, v_o^w).

Let S_o be the set of the sources which claimed a value on the object o and V_o be the set of candidate values collected from S_o. Each worker in W_o answers a question about the object o by selecting a value from V_o.

In our problem setting, we assume that we have a hierarchy tree H of the claimed values. If we are interested in an attribute related to locations (e.g., birthplace), H would be a geographical hierarchy with different levels of granularity (e.g., continent, country, city, etc.). We also assume that there is no answer with the value of the root

Object	Source	Claimed value
Statue of Liberty	UNESCO	NY
Statue of Liberty	Wikipedia	Liberty Island
Statue of Liberty	Arrangy	LA
Big Ben	Quora	Manchester
Big Ben	tripadvisor	London

Table 3.2: Locations of tourist attractions

in the hierarchy since it provides no information at all (e.g., Earth as a birthplace). We summarize the notations to be used present our truth discovery algorithm in Table 3.1.

Example 1. *Consider the records in Table 3.2. Since the source Wikipedia claims that the location of the Statue of Liberty is Liberty Island, it is represented by* $v_o^s =$*'Liberty Island' where* $o =$*'Statue of Liberty' and* $s =$*'Wikipedia'. If a human worker 'Emma Stone' answered Big Ben is in London, it is represented by* $v_o^w =$*'London' where* $o =$*'Big Ben' and* $w =$*'Emma Stone'.*

Problem definition. The proposed truth discovery algorithm repeatedly alternates the hierarchical truth discovery and task assignment until the budget of crowdsourcing runs out. Given a set of objects O and a hierarchy tree H, we define the two subproblems of the crowdsourced truth discovery in the presence of hierarchies.

Definition 5 (Hierarchical truth discovery problem)**.** *For a set of records* R *collected from the sources and a set of answers* A *from the workers, we find the most specific true value* v_o^* *of each object* $o \in O$ *among the candidate values in* V_o *by using the hierarchy* H*.*

Definition 6 (Task assignment problem)**.** *For each worker* w *in a set of workers* W*,*

we select the top-k objects from O which are likely to increase the overall accuracy of the inferred truths the most by using the hierarchy H.

We present a hierarchical truth discovery algorithm in Chapter 6 and a task assignment algorithm in Chapter 7.

Chapter 4

Topic-aware Relation Extraction

4.1 Motivation

Relation extraction (RE) is a task to identify the semantic relations between entities from text. Most of the existing works for relation extraction [56, 47, 83, 82] mainly focus on the sentence-level extraction which finds the relations between entities in a sentence. However, many relational facts are expressed across multiple sentences. In a large-scale document-level relation extraction (DocRED) dataset published in [77], 46.4% of the relation instances are associated with multiple sentences and 40.7% of the relational facts cannot be found by the sentence-level extraction [77]. Extending the relation extraction from sentence-level to document-level enables us to find the relational facts over multiple sentences. Thus, recent works [77, 73] proposed models for document-level relation extraction. However, those models still do not fully take account of the unique characteristics of document-level RE.

A lot of documents are each written to describe a *topic entity*. For example, there is a title for each Wikipedia article and the title usually represents an entity that is mainly described in the article. Besides, the product pages on Amazon and the IMDb movie pages also have their topic entities. When the subject of a sentence is the topic entity, it is often represented by a pronoun or even omitted since the subject is obvious in

Robert K. Huntington

[1] **[Robert Kingsbury Huntington: E1]** (**[13 March 1921: E2]** - **[5 June 1942: E3]**), was a naval aircrewman and member of Torpedo Squadron 8. [2] He was radioman/gunner to Ensign George Gay's TBD Devastator aircraft. [3] Along with his entire squadron, **[Huntington: E1]** was shot down during the Battle of Midway, on 4-5 June 1942. [4] Born in **[Los Angeles: E4]**, California, enlisted in the United States Navy 21 April 1941. [5] He served on board Lexington (CV-2) and was rated aviation radioman third class before being transferred to Torpedo Squadron 8 on board Hornet (CV-8). ...

Entities	Relations
E1: **Robert K. Huntington**	R1: ⟨**E1**, `date_of_birth`, **E2**⟩
E2: **13 March 1921**	R2: ⟨**E1**, `date_of_death`, **E3**⟩
E3: **5 June 1942**	R3: ⟨**E1**, `born_in`, **E4**⟩
E4: **Los Angeles**	

Figure 4.1: An example of relations in a document

many cases. Figure 4.1 shows an example of Wikipedia pages whose title is 'Robert K. Huntington'. While the relational fact that Robert K. Huntington was born in LA is presented in the 4th sentence, it is not explicitly stated in the text that the person born in LA is Robert K. Huntington. RE with the topic entity in semi-structured HTML web pages has been addressed in [8, 51, 76]. However, since they utilize HTML tags and DOM tree structures, those works are not directly applicable to plain text. In a large-scale document-level RE dataset [77], the title entities are involved in 28.8% of all relation instances. However, existing models [77, 73] for extracting relations from text documents do not distinguish the topic entity from the other entities. Thus, they often fail to find the relations with the implicit mentions of the topic entity. To find such relations, we propose a Topic-aware Relation EXtraction (*T-REX*) model which

is robust to the omitted mentions of topic entities. As far as we know, this is the first work that utilizes the topic entities to extract relations from plain text.

An entity tends to be mentioned multiple times in a document, whereas it is usually mentioned once in a sentence. Moreover, each mention of an entity can be involved in a different relationship with the topic entity. For example, if a singer-songwriter writes and performs a song, the two different relations (i.e., `performer` and `composer`) can be separately expressed in a document for the song. Previous works [77, 73] build a vector representation of each entity by averaging the corresponding mention vectors before determining the relationships between entities. However, the different meanings of the mentions of an entity may disappear while averaging the mention vectors. To tackle the problem, we first predict the relationship between the topic entity and each mention of other entities. Then we combine the results by using a smooth-maximum function which is an approximation of the maximum function. Thus, our model can extract the relationships by taking advantage of the subtle meaning of each mention.

As we discussed in Chapter 1, knowledge fusion employs multiple relation extractors and ensemble them to improve the accuracy. Thus, the improvement in ensemble accuracy is an important performance indicator as well as the accuracy of the standalone model. Since our proposed model is the only one that considers the topic entity of a document, it can find the relations which cannot be detected by existing models. Thus, the accuracy of the extracted relations with the topic entity can be significantly improved when existing algorithms are ensembled with *T-REX*.

In this chapter, we propose the *T-REX* model to find the relations with the topic entity even when the topic entity is omitted in the sentence. By conducting experiments on the DocRED dataset, we show that *T-REX* outperforms the existing models in extracting relations with topic entities. We also show that the accuracy of the extracted relations with the topic entity is significantly improved when existing algorithms are ensembled with *T-REX*.

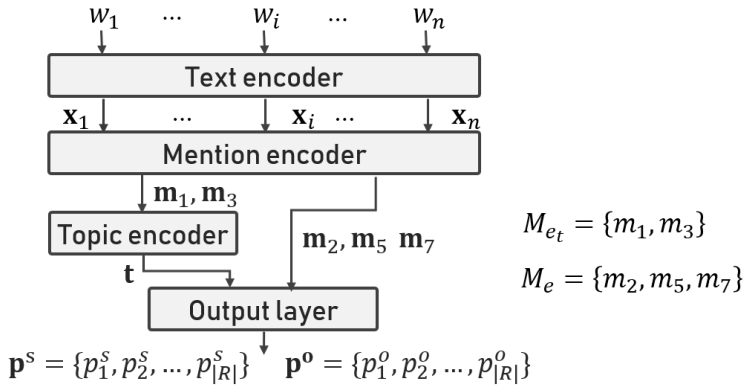


Figure 4.2: Architecture of the proposed model

4.2 Proposed Model

An entity mention is a string that refers to a real-world entity in a text. Since an entity can be mentioned multiple times in a document, there can be several entity mentions in a document for an entity. For instance, in Figure 4.1, 'Robert Kingsbury Huntington' and 'Huntington' in the input document are entity mentions of the entity 'Robert K. Huntington'. We assume that each entity mention is annotated in the text and the topic entity of each document is known in advance. Given an entity annotated document and its title entity, the goal of the task is to extract relations with the title entity of the document. The overall architecture of *T-REX* is shown in Figure 4.2. *T-REX* consists of three stacked encoders (*text encoder*, *mention encoder*, *topic encoder*) and an *output layer*. The text encoder receives a document and produces the embedding of each word in the document. For each entity mention in the document, the mention encoder generates the vector representation of the entity based on the embeddings of corresponding words. Title encoder produces the title vector from the mention vectors of the title entity. The output layer takes the title vector and the mention vectors of the other entity as input. Finally, the output layer predicts the relations between the title entity and the other entity. We next describe the encoders and the output layer in detail.

4.2.1 Encoders

The text encoder receives a document and produces the embedding of each word in the document. For each entity mention, the mention encoder generates a mention vector based on the embeddings of corresponding words. The topic encoder produces a topic vector from the mention vectors of the topic entity.

Text encoder. We use the pre-trained language model BERT [12] for the text encoder following the previous work [73]. BERT has a large model capacity that can be fine-tuned for many NLP tasks [68]. Thus, it enables us to use simple structures for the mention encoder and topic encoder. Let $\mathbf{x}_i$ denote the d_{bert}-dimensional embedding of the i-th token obtained from BERT. For each token in entity mentions, we use a linear layer to calculate a word vector $\hat{\mathbf{x}}_i$ to be given as an input to the mention encoder as follows:

$$\hat{\mathbf{x}}_i = \mathbf{W}\mathbf{x}_i + \mathbf{b}$$

where $\mathbf{W} \in \mathbb{R}^{d_{bert} \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ are trainable parameters.

Mention encoder. For each entity mention m_k spanning from the b_k-th token to the e_k-th token of the document, we compute the mention vector $\mathbf{m}_k$ by averaging the corresponding word vectors as

$$\mathbf{m}_k = \frac{1}{e_k - b_k + 1} \sum_{i=b_k}^{e_k} \hat{\mathbf{x}}_i.$$

Topic encoder. It computes the representation of the topic entity by averaging the mention vectors. Let M_{e_t} be the set of mentions of the topic entity e_t. Then, the topic entity vector $\mathbf{u}_{topic}$ is computed as

$$\mathbf{u}_{topic} = \frac{1}{|M_{e_t}|} \sum_{k \in M_{e_t}} \mathbf{m}_k.$$

Note that the previous works [77, 73] build a entity vector by averaging the mention vectors. However, we do not take the averages for the non-topic entities since different mentions can be involved in different relations.

4.2.2 Output Layer

The output layer takes the vector representation of the topic entity e_t and mention vectors of an entity e. Let p_r^h and p_r^t be the probabilities that the relations $\langle e_{topic}, r, e\rangle$ and $\langle e, r, e_{topic}\rangle$ are expressed in the document, respectively. The output layer estimates p_r^h and p_r^t for $r \in R$. As discussed in Section 4.1, we first make a prediction for each entity mention and combine the mention-wise result because different mentions of an entity can be involved in a different relationship with the topic entity. We first present how to compute p_r^h and briefly introduce how to compute p_r^t later.

Mention-wise prediction. Let $\mathbf{p}_k^h = \{\mathbf{p}_k^h(1), \mathbf{p}_k^h(2), ..., \mathbf{p}_k^h(|R|)\}$ be a $|R|$-dimensional vector where $\mathbf{p}_k^h(r)$ represents the probability that the relation $\langle e_t, r, e\rangle$ is expressed with the mention $m_k \in M_e$ where M_e is the set of mentions of entity e. For each mention $m_k \in M_e$ of an entity e, we compute $\mathbf{p}_k^h$ using a bilinear layer and a sigmoid activation function $\sigma(.)$ as:

$$\mathbf{v}_k^h = \mathbf{u}_{topic}{}^\top \mathbf{W}_p \mathbf{m}_k + \mathbf{b}_p,$$
$$\mathbf{p}_k^h = \sigma(\mathbf{v}_k^h)$$

where $\mathbf{W}_p \in \mathbb{R}^{d\times|R|\times d}$ and $\mathbf{b}_p \in \mathbb{R}^{|R|}$ are trainable parameters, and $\mathbf{u}_{topic}{}^\top \mathbf{W}_p \mathbf{m}_k$ is a bilinear tensor product which results in a $|R|$-dimensional vector.

Combining the mention-wise predictions. We want to output the relation if it is expressed at least once in the document. Thus, we combine the mention-wise probabilities via max-pooling as follows:

$$p_r^h = \max_{m_k \in M_e} \mathbf{p}_k^h(r).$$

Since the sigmoid function σ is monotonically increasing, we can interchange the max operation and σ as below.

$$p_r^h = \max_{m_k \in M_e} \mathbf{p}_k^h(r) = \max_{m_k \in M_e} \sigma(\mathbf{v}_k^h) = \sigma\Big(\max_{m_k \in M_e} \mathbf{v}_k^h(r)\Big)$$

Previous works [9, 71] report that a smooth maximum function *LogSumExp* (LSE) leads to faster training than the hard maximum function which has sparse gradients.

Thus, we also employ LSE to compute the document-level prediction as follows:

$$p_r^h = \sigma\Big(LSE(\mathbf{v}_k^h(r))\Big)$$

where $LSE(\mathbf{v}_k^h(r)) = \log \sum_{m_k \in M_e} exp(\mathbf{v}_k^h(r))$.

Computing p_r^t. To compute the p_r^t, we use the same network and interchange the inputs as

$$\mathbf{v}_k^t = \mathbf{m}_k^\top \mathbf{W}_p \mathbf{u}_{topic} + \mathbf{b}_p,$$
$$p_r^t = \sigma\Big(LSE(\mathbf{v}_k^t(r))\Big)$$

where $\mathbf{W}_p$ and $\mathbf{b}_p$ are the parameters used to compute p_r^h.

4.2.3 Training

Binary cross entropy function is used as the loss function. We used the Adam [35] optimizer with a learning rate 10^{-5}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Each training batch contains 8 documents. The hidden dimension size d is set to 128. We used the BERT-base model that provides 768-dimensional embeddings ($d_{bert} = 768$).

4.3 Experiments

We compared ***T-REX***[1] with five existing models: ***FTB***[73], ***CNN***[83], ***LSTM***[77], ***BiLSTM*** [6] and ***CA***[67]. *FTB* is a Fine-Tuned model of the BERT [12] for the document-level RE. The rest of them are originally proposed for sentence-level RE and adapted to document-level RE in [77]. All models are implemented in PyTorch and trained on a machine with a TITAN RTX GPU.

4.3.1 Experimental Settings

Dataset. We conducted experiments on the DocRED dataset [77], which is a large-scale dataset for document-level RE. The training set of DocRED contains 3,053 docu-

[1]The source code is available at `https://github.com/woohwanjung/T-REX`

Data	Docs.	# relations	Labeling
Train-HA	3,053	96	HA
Train-DS	101,873	96	DS
Dev	500	96	HA
Test	500	96	HA

Table 4.1: Statistics of DocRED dataset

ments with human annotated labels and 101,873 documents with distantly supervised labels. DocRED includes 1,000 documents with human annotated labels for testing. We remove 500 documents from the testing set to use them for validation. In supervised setting, we use the documents with human annotation for training. In weakly supervised setting, we use the documents with distantly supervised labels for training. To evaluate the models, we splitted the test set Table 4.1 shows the statistics of the dataset.

Note that we only use the relations with the topic entity to train our model and to test all models. The topic entity of each document is not given in DocRED. Thus, we annotate the topic entity of each document using string matching. More specifically, we choose the mention that has the maximum overlap of tokens with the title of the document and set the corresponding entity to the topic entity. Table 4.2 shows the number of all triples and topic-related triples in DocRED where a triple is a topic-related triple if its head entity or tail entity is the topic entity. Among 50,503 human annotated triples in DocRED, 14,544 triples are (28.8% of all triples) topic-related triples. The statistics indicate that a significant number of triples can be extracted by topic-aware relation extraction.

Evaluation measures. We use two widely used measures, F1 and AUC, to evaluate the models. As in [77], we also report the scores excluding the relations that exist in the training set. We refer to the scores as Ign F1 and Ign AUC, respectively. Note that

Data	Number of all triples	Number of topic-related triples	Proportion of topic-related triples
Train-HA	38,180	10,918	28.6%
Dev	6,316	1,788	28.3%
Test	6,007	1,838	30.6%
Total	50,503	14,544	28.8%

Table 4.2: Proportion of topic-related triples

we only use the relations with the topic entity to train our model and test all models. Thus, we compute the evaluation measures by using the relations where the subject or object is the topic entity.

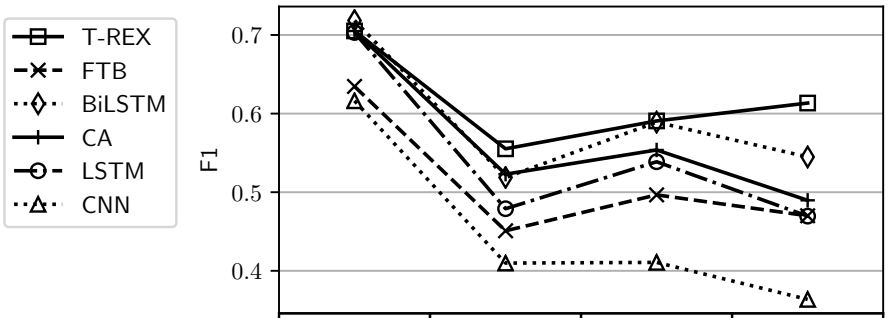
4.3.2 Experimental Results

Comparison with existing models. Table 4.3 shows the experimental results under the weakly supervised and supervised settings. *T-REX* outperforms all compared models in the weakly supervised setting. Besides, it shows the highest F1 and Ign F1 under supervised setting. *FTB* trained with human annotations has the highest AUC and Ign AUC among the models trained with human annotations. However, interestingly, *FTB* shows the worst performance under weakly supervised setting. The result implies that *FTB* is not robust to the false labels in weakly supervised setting.

Distance vs. F1 score. To validate the necessity of utilizing topic entities, we analyze the correlation between F1 score and the distance from the topic entity to other entity mentions in the document. Given a pair of entities, the distance is defined as the minimum number of sentences between the mentions of the two entities. For example, if the two entities are mentioned in the same sentence, the distance is 0. In Figure 4.3, we plotted F1 scores varying the distance d with weakly supervised setting. Every model shows the highest F1 score when the two entities are mentioned in the same sentence

	Model	F1	AUC	Ign F1	Ign AUC
Weakly supervised setting	*CNN* [83]	0.5461	0.5666	0.5271	0.5373
	LSTM [77]	0.6275	0.6463	0.5964	0.6055
	CA [67]	0.6397	0.6480	0.6168	0.6104
	BiLSTM [6]	0.6572	0.6704	0.6287	0.6288
	FTB [73]	0.5710	0.5650	0.5248	0.4965
	T-REX	**0.6624**	**0.6978**	**0.6364**	**0.6634**
Supervised setting	*CNN* [83]	0.5635	0.5564	0.5603	0.5512
	LSTM [77]	0.5797	0.5896	0.5746	0.5844
	CA [67]	0.5855	0.5969	0.5796	0.5886
	BiLSTM [6]	0.5774	0.5892	0.5739	0.5851
	FTB [73]	0.6491	**0.6631**	0.6458	**0.6577**
	T-REX	**0.6569**	0.6456	**0.6589**	0.6468

Table 4.3: Performance of the document-level RE models



	$d = 0$	$d = 1$	$d = 2$	$d \geq 3$
Num. Labels	1,237	339	113	149
Improvement	-1.8% $\sim$ 14.4%	6.2% $\sim$ 35.5%	0.3% $\sim$ 43.8%	12.6% $\sim$ 68.7%

Figure 4.3: F1 score by the distance from the topic entity

(i.e., $d = 0$). The performance gap between *T-REX* and other models widens considerably when the two entities are not mentioned in the same sentence. Especially, when we have $d \geq 3$, *T-REX* achieves 13 - 69% of performance improvement over the existing models. Since the distance becomes larger if the topic entity is omitted in some sentences, the result implies that *T-REX* finds the relations with the omitted topic entity better than the existing models. It shows that *T-REX* extracts many relations that are not detected by the existing models.

Performance of ensemble models. As we discussed in Section 4.1, improvement in ensemble accuracy is an important performance indicator as well as the accuracy of the stand-alone model. We conduct an experiment to see the additional accuracy improvement when *T-REX* is ensembled with the existing models. To ensemble a pair of RE models, we use a multilayer perceptron which has two hidden layers with 5 hidden units per each layer. For a triple $\langle e_t, r, e \rangle$, the input feature vector consists of the output probabilities obtained from the two models and the minimum distance between the mentions of e_t and e. The ensemble model is a 2-layer MLP with 5 hidden dimensions. The ensemble model is implemented by using scikit-learn [60] and trained with Adam optimizer. We report the F1 scores of all possible pair-wise ensemble results of the

Weakly supervised setting						
	T-REX	*BiLSTM*	*CA*	*FTB*	*LSTM*	*CNN*
T-REX	-	**0.7472**	0.7455	0.7206	0.7430	0.7367
BiLSTM	**0.7472**	-	0.7084	0.7139	0.7112	0.6931
CA	0.7455	0.7084	-	0.6914	0.7088	0.6895
FTB	0.7206	0.7139	0.6914	-	0.7050	0.6843
LSTM	0.7430	0.7112	0.7088	0.7050	-	0.6599
CNN	0.7367	0.6931	0.6895	0.6843	0.6599	-
Supervised setting						
	TEX	BiLSTM	CA	FTB	LSTM	CNN
TEX	-	0.6896	0.6898	**0.7122**	0.6867	0.6897
BiLSTM	0.6896	-	0.6248	0.6872	0.6041	0.6210
CA	0.6898	0.6248	-	0.6882	0.6367	0.6247
FTB	**0.7122**	0.6872	0.6882	-	0.6860	0.6886
LSTM	0.6867	0.6041	0.6367	0.6860	-	0.5914
CNN	0.6897	0.6210	0.6247	0.6886	0.5914	-

Table 4.4: Performance of pairwise ensemble models

implemented RE models in Table 4.4. For each row, we underlined the best ensemble result of the model. We found that every compared model shows the best performance when it is ensembled with *T-REX*. The result means that the accuracy of knowledge bases can be improved when *T-REX* is used to build knowledge bases with knowledge fusion.

Effectiveness of the output layer. Recall that the output layer of *T-REX* first make a prediction for each mention and combines the results by using the $LogSumExp$ function. We implemented three baselines to validate the effectiveness of the output

	Output layer	F1	AUC	F1 Ign	AUC Ign
Weakly supervised setting	Avg + Prediction [77, 73]	0.6450	0.6218	0.6168	0.5814
	Attention + Prediction	0.6489	0.6632	0.6225	0.6238
	Prediction + Max	0.6326	0.6855	0.6034	0.6525
	Prediction + LogSumExp	**0.6624**	**0.6978**	**0.6364**	**0.6634**
Supervised setting	Avg + Prediction [77, 73]	0.6480	0.5795	0.6494	0.5811
	Attention + Prediction	0.6454	0.6361	0.6489	0.6395
	Prediction + Max	**0.6594**	**0.6512**	**0.6609**	**0.6525**
	Prediction + LogSumExp	0.6569	0.6456	0.6589	0.6468

Table 4.5: Comparison with baseline output layers

layer. The first baseline (Avg+Prediction) is inspired by the models in [77, 73]. For a given entity e, it first computes the entity vector $\mathbf{e}$ by averaging its mention vectors. Then it feeds the entity vector $\mathbf{e}$ and the topic vector $\mathbf{t}$ to the output layer that consists of a Bilinear layer and the sigmoid activation function. The second baseline (Attention+Prediction) is obtained by substituting the simple averaging of the first baseline with the attention mechanism [2]. The last one (Prediction + Max) is obtained by replacing the $LogSumExp(LSE)$ function of our proposed model with the maximum function. Our proposed output layer always outperforms the first two baselines. The result shows that our approach, making a prediction for each mention and combining the results, is effective for the task we addressed. Moreover, LSE significantly outperforms Max in weakly supervised setting while Max shows slightly higher accuracy than LSE in the supervised setting. Since LSE considers all predictions while Max takes into account of a single prediction with the highest confidence, LSE generally exhibits better performance than Max.

Case study. Figure 4.1 shows a part of the document titled 'Robert K. Huntington' which is in the testing set of DocRED. The relations R1 and R2 are expressed in

the first sentence with explicit mentions of the subjects and objects. Thus, all models correctly identify the relations. Meanwhile, the relation R3 is represented in the fourth sentence and the subject of R3, which is the topic entity, is omitted in the sentence. Thus, all models except *T-REX* fail to find the relation in supervised setting while *T-REX* extracts the relation. With weakly supervised setting, *T-REX* and *LSTM* find the relation while the others cannot find the relation.

Chapter 5

Dual Supervision Framework for Relation Extraction

5.1 Motivation

As we discussed in Section 2.2, recent works [47, 83, 82, 67, 73] proposed deep neural networks for relation extraction to overcome the drawbacks of traditional works that utilize handcrafted features. To train a deep relation extraction model, we need a large volume of fully labeled training data in the form of text-triple pairs. Although human annotation provides high-quality labels to train the relation extraction models, it is difficult to produce a large-scale training data since manual labeling is expensive and time-consuming. Thus, distant supervision [56], which labels the relations between entities in a sentence using existing knowledge bases, is proposed to automatically produce a large labeled data. For a text with a head entity e_h and a tail entity e_t, when a triple $\langle e_h, r, e_t \rangle$ exists in the KB for any relation type r, distant supervision produces a label $\langle e_h, r, e_t \rangle$ even though the relationship is not expressed in the text. Thus, it suffers from the wrong labeling problem. For instance, if a triple $\langle UK, \texttt{capital}, London \rangle$ is in the KB, distant supervision labels the triple even for the sentence 'London is the largest city of the UK'.

Although each of the two labeling methods has a certain weakness, most of the existing works for RE utilize either human-annotated (HA) data or distantly supervised

(DS) data. To take advantage of the high accuracy of human annotation and the cheap cost of distant supervision, we propose to effectively utilize a large DS data as well as a small amount of HA data. Since DS data is likely to have *labeling bias*, simply combining the two types of data to train a RE model may decrease the prediction accuracy. To take a close look at the labeling bias, let the *inflation* of a relation type be the ratio of the average frequencies of the relation type per text in DS data and HA data, respectively. We say that a relation type is *unbiased* if the average frequency of the relation type in DS data is the same as that in HA data (i.e., the inflation of the relation is 1). By examining a document-level RE dataset (DocRED) [77] with 96 relation types, we found that the inflations of the relation types are from 0.48 to 85.9. It indicates that distant supervision tends to generate a large number of false labels for some relation types. Thus, although we train a model with additional true labels obtained by human-annotation, this bias produces a significant number of false positives since the number of false labels is much larger than that of true labels.

Recently, a domain adaptation approach to tackle the labeling bias problem for RE was proposed in [78]. It trains a RE model on DS data and adjusts the bias term of the output layer by using HA data. Although the bias adjustment achieves a meaningful accuracy improvement, it has a limitation. An underlying assumption of the method is that the labeling bias is static for every text since it adjusts the bias term only once after training and uses the same bias during the test time. However, the labeling bias varies depending on contextual information. For example, in DocRED dataset, most of the `capital` relation labeled by distant supervision are false positive. However, if the phrase 'is the capital city of' appears in the text, the label is likely to be a true label. Thus, we need to take account of contextual information to extract relations more accurately by considering the labeling bias.

To effectively utilize DS data and HA data for training RE models, we propose the *dual supervision framework* that can be applied to most existing RE models to achieve additional accuracy gain. Since the label distributions in HA data and DS data are quite

different, we cast the task of training RE models with both data as a multi-task learning problem. Thus, we employ the two separate output modules *HA-Net* and *DS-Net* to predict the labels by human annotation and distant supervision, respectively, while previous works utilize a single output module. This allows the different predictions of the labels for human annotation and distant supervision, and thus it prevents the degradation of accuracy by incorrect labels in DS data. If we simply separate the prediction networks to apply the multi-task learning, *HA-Net* cannot learn from distantly supervised labels. To enable *HA-Net* to learn from DS data, we propose an additional loss term called *disagreement penalty*. It models the ratio of the output probabilities from the prediction networks *HA-Net* and *DS-Net* by using maximum likelihood estimation with log-normal distributions to generate the calibrated gradient to update *HA-Net* to effectively reflect distantly supervised labels. Furthermore, our framework exploits two additional networks μ-Net and σ-Net to adaptively estimate the log-normal distribution by considering contextual information. Moreover, we theoretically show that the disagreement penalty enables *HA-Net* to effectively utilize the labels generated by distant supervision. Finally, we validate the effectiveness of the dual supervision framework on two types of tasks: sentence-level and document-level REs. The experimental results confirm that our dual supervision framework significantly improves the prediction accuracy of existing RE models. In addition, the dual supervision framework substantially outperforms the state-of-the-art method [78] in both sentence-level and document-level REs with the relative F1 score improvement of up to 32%.

5.2 Existing Works on Relation Extraction

The dual supervision framework provides a method to modify the structure of existing RE models to improve the accuracy. Thus, we first review existing models for relation extraction before presenting our dual supervision framework. A typical RE model consists of a feature encoder and a prediction network, as shown in Figure 5.1. The feature

encoder converts a text into the hidden representations of the head and tail entities. [6] and [73] exploit Bi-LSTM and BERT, respectively, to encode the text. On the other hand, [83] and [82] use CNN for the encoder. In addition, [83] propose the position embedding to consider the relative distance from each word to head and tail entities.

The prediction network outputs the probability distribution of the relations between the entities. Since sentence-level RE is a multi-class classification task, sentence-level RE models [6, 83, 82] utilize a *softmax classifier* as the prediction network and use categorical cross entropy as the loss function. On the other hand, document-level RE models [77, 73] use a *sigmoid classifier* and binary cross entropy as the prediction network and the loss function, respectively. Since the labels obtained from distant supervision are noisy and biased, with a single prediction network, it is hard to make accurate predictions for DS data and HA data together.

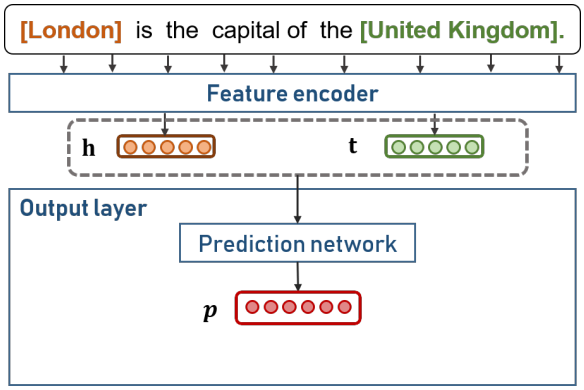


Figure 5.1: The overall architecture of existing RE models

5.3 Dual Supervision Framework

We first present an overview of the dual supervision framework which effectively utilizes both human-annotated (HA) data and distantly supervised (DS) data for training RE models. We next introduce the detailed structure of the output layer in our framework and propose our novel loss function with disagreement penalty that considers

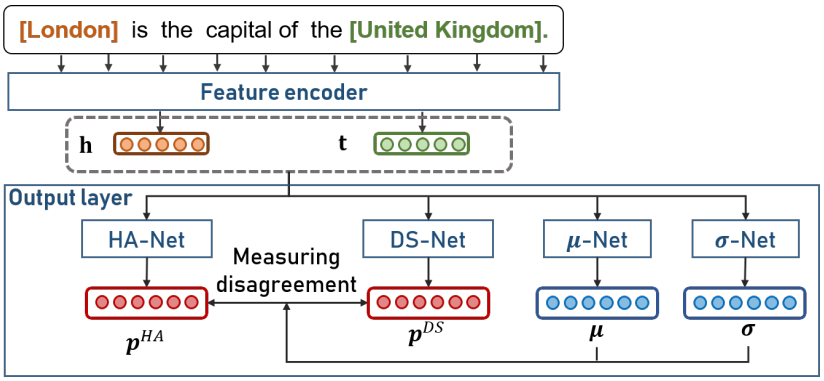


Figure 5.2: The overall model architecture of our dual supervision framework

the labeling bias of distant supervision. Then, we describe how to train the proposed model with both types of data as well as how to extract relations from the test data. Finally, we discuss how the disagreement penalty makes each prediction network learn from the labels for the other prediction network although we use separate prediction networks.

5.3.1 An Overview of the Dual Supervision Framework

As shown in Figure 5.2, our framework consists of a feature encoder and an output layer with 4 sub-networks. It is general enough to accommodate a variety of existing RE models to improve their accuracy. We can apply our framework to an existing RE model by using the feature encoder of the model and building the four sub-networks which exploit the structure of the original prediction network. Since our framework uses the feature encoder of the existing models, we briefly describe only the output layer here.

Unlike the previous works, to allow the difference in the predictions for human annotated labels and distantly supervised labels, we exploit multi-task learning by employing two separate prediction networks *HA-Net* and *DS-Net* to predict the labels in HA data and DS data, respectively. We also use *HA-Net* to extract relations from the test data. The separation of the prediction networks prevents the accuracy degradation

caused by incorrect labels from distant supervision. If we simply utilize two prediction networks to apply the multi-task learning, *HA-Net* cannot learn from distantly supervised labels although the prediction networks share the feature encoder. To enable *HA-Net* to learn from distantly supervised labels, we introduce an additional loss term called *disagreement penalty*. It models the disagreement between the outputs of *HA-Net* and *DS-Net* by using maximum likelihood estimation with log-normal distributions. Furthermore, to adaptively estimate the parameters of the log-normal distribution by considering contextual information, we exploit two parameter networks μ-Net and σ-Net.

For a label $\langle e_h, r, e_t \rangle$, let I_{HA} be an indicator variable that is 1 if the label is obtained by human annotation and 0 otherwise. The proposed framework uses the following loss function for a label $\langle e_h, r, e_t \rangle$

$$L_{h,t} = I_{HA} \cdot L_{h,t}^{HA} + (1 - I_{HA}) \cdot L_{h,t}^{DS} + \lambda \cdot L_{h,t}^{DS\text{-}HA} \tag{5.1}$$

where $L_{h,t}^{HA}$ and $L_{h,t}^{DS}$ denote the prediction loss of *HA-Net* and *DS-Net*, respectively, and $L_{h,t}^{DS\text{-}HA}$ is the disagreement penalty to capture the distance between the predictions by *HA-Net* and *DS-Net*. The hyper parameter λ controls the relative importance of the disagreement penalty to the prediction errors. By using a separate prediction network for each type of data and introducing the disagreement penalty, *HA-Net* learns from distantly supervised labels while reducing overfitting to noisy DS data.

5.3.2 Separate Prediction Networks

To alleviate the accuracy degradation from the noisy labels in DS data, we utilize two prediction networks. The network *HA-Net* is used to predict the human-annotated labels from the train data and to predict relations from the test data. The other prediction network *DS-Net* predicts the labels obtained by distant supervision. We use the prediction network of an existing model for both prediction networks of our framework without sharing the model parameters. The prediction networks *HA-Net* and *DS-Net* output the $|R|$-dimensional vectors $\mathbf{p}^{HA} = [p(r_1|e_h, e_t, HA), ..., p(r_{|R|}|e_h, e_t, HA)]$ and

$\mathbf{p}^{DS} = [p(r_1|e_h, e_t, DS), \ldots, p(r_{|R|}|e_h, e_t, DS)]$, respectively, where $p(r|e_h, e_t, HA)$ and $p(r|e_h, e_t, DS)$ are the probabilities that there exists a label $\langle e_h, r, e_t \rangle$, in HA data and DS data, respectively. We simply denote $p(r|e_h, e_t, HA)$ and $p(r|e_h, e_t, DS)$ by p_r^{HA} and p_r^{DS}, respectively.

5.3.3 Disagreement Penalty

Distant supervision labels are biased and the size of the bias varies depending on the type of relation. Moreover, the bias can vary depending on many other features such as the types of head and tail entities as well as the contents of a text. Thus, we propose to use an effective disagreement penalty to model the labeling bias depending on the context where the head and tail entities are located.

Distribution	**p-value**
Log-normal	0.008
Weibull	0.001
Chi-square	4.6×10^{-10}
Exponential	3.6×10^{-13}
Normal	1.2×10^{-15}

Table 5.1: The result of K-S test

Distribution of inflations. We measure the labeling bias by using the inflations of relations. Recall that the inflation of a relation type is the ratio of the average frequencies of the relation type per text in DS data and HA data, respectively. To investigate the distribution of inflations, we computed the inflations of 96 relation types in DocRED data[1]. Since Kolmogorov-Smirnov (K-S) test [53] is widely used to determine whether an observed data is drawn from a given probability distribution, we used it to find the best-fit distribution of the inflations. Since the range of the inflation is $[0, \infty)$, we eval-

[1]The inflations are shown in Appendix A.1

uated p-values of the four probability distributions supported on $[0, \infty)$: Log-normal, Weibull, chi-square and exponential distributions. In addition, we include the normal distribution as a baseline. Table 5.1 shows the result of K-S test for DocRED data. Note that a probability distribution has a high p-value if the probability distribution fits the data well. Since the log-normal distribution has the highest p-value, it is the best-fit distribution among the five probability distributions. Based on the observation, we model the disagreement penalty between the outputs of the two prediction networks.

Modeling the disagreement penalty. We develop the disagreement penalty based on the maximum likelihood estimation. Let X_r be the random variable which denotes the ratio of p_r^{DS} to p_r^{HA}. Since the inflation is the ratio of the number of labels in DS data and HA data, the ratio p_r^{DS}/p_r^{HA} represents the *conditional inflation* of the relation type r conditioned on the text with head and tail entities. Thus, we assume that X_r follows a log-normal distribution $L(\mu_r, \sigma_r^2)$ whose probability density function is

$$f(x) = \frac{1}{x\sigma_r\sqrt{2\pi}} exp\left(-\frac{(\log x - \mu_r)^2}{2\sigma_r^2}\right). \tag{5.2}$$

The disagreement penalty $L_{h,t}^{DS\text{-}HA}$ is defined as the negative log likelihood of the conditional inflation p_r^{DS}/p_r^{HA}, which is obtained by substituting p_r^{DS}/p_r^{HA} into Equation (5.2) as follows:

$$\begin{aligned} &-\log f\left(p_r^{DS}/p_r^{HA}\right) \\ &= \frac{1}{2}\left(\frac{\log p_r^{DS} - \log p_r^{HA} - \mu_r}{\sigma_r}\right)^2 + \log p_r^{DS} - \log p_r^{HA} + \log \sigma_r + \frac{\log 2\pi}{2}. \end{aligned} \tag{5.3}$$

Since $\frac{\log 2\pi}{2}$ is constant, we utilize the disagreement penalty in Equation (5.3) without the constant term.

If we set μ_r and σ_r to fixed values, we cannot effectively assess the conditional inflation since it can vary depending on the context. For example, although the inflation of the relation type `capital` is high, the conditional inflation should be lower if a particular phrase such as 'is the capital city of' appears in the text. To take account of the contextual information, we employ two additional networks μ-Net and σ-Net to estimate the μ_r and σ_r that are the parameters of log-normal distribution $L(\mu_r, \sigma_r^2)$.

5.3.4 Parameter Networks

The parameter networks μ-Net and σ-Net output the vectors $\mu = [\mu_1, ..., \mu_{|R|}]$ and $\sigma = [\sigma_1, ..., \sigma_{|R|}]$, respectively, which are the parameters of the log-normal distributions to represent the conditional inflation for $r \in R$. Both μ-Net and σ-Net have the same structure as those of the prediction networks except their output activation functions. For a log-normal distribution $L(\mu, \sigma)$, the parameter μ can be positive or negative, and σ is always positive. Thus, we use a hyperbolic tangent function and a softplus function [18] as the output activation functions of μ-Net and σ-Net, respectively.

For example, if the prediction network of the original RE model consists of a bilinear layer and an output activation function, the parameter vectors $\mu \in \mathbb{R}^{|R|}$ and $\sigma \in \mathbb{R}^{|R|}$ are computed from the head entity vector $\mathbf{h} \in \mathbb{R}^d$ and tail entity vector $\mathbf{t} \in \mathbb{R}^d$ as

$$\mu = tanh(\mathbf{h}^\top \mathbf{W}^\mu \mathbf{t} + \mathbf{b}^\mu),$$
$$\sigma = softplus(\mathbf{h}^\top \mathbf{W}^\sigma \mathbf{t} + \mathbf{b}^\sigma) + \varepsilon$$

where $softplus(x) = \log(1 + e^x)$ and ε is a sanity bound preventing extremely small values of σ_r from dominating the loss function, and $\mathbf{W}^\mu \in \mathbb{R}^{d \times |R| \times d}$, $\mathbf{W}^\sigma \in \mathbb{R}^{d \times |R| \times d}$, $\mathbf{b}^\mu \in \mathbb{R}^{|R|}$ and $\mathbf{b}^\sigma \in \mathbb{R}^{|R|}$ are learnable parameters. We set the sanity bound ε to 0.0001 in our experiment.

5.3.5 Loss Function

For sentence-level relation extraction, we use the categorical cross entropy loss as the prediction losses $L^{HA}_{h,t}$ and $L^{DS}_{h,t}$. For a label $\langle e_h, r, e_t \rangle$, we obtain the following loss function from Equations (5.1) and (5.3)

$$\begin{aligned} L_{h,t} =& I_{HA} \cdot L^{HA}_{h,t} + (1 - I_{HA}) \cdot L^{DS}_{h,t} + \lambda \cdot L^{DS\text{-}HA}_{h,t} \\ =& - I_{HA} \cdot \log p^{HA}_r - (1 - I_{HA}) \log p^{DS}_r + \lambda \left[\frac{1}{2} \left(\frac{\ell_r - \mu_r}{\sigma_r} \right)^2 + \ell_r + \log \sigma_r \right] \end{aligned} \tag{5.4}$$

where $\ell_r = \log p_r^{DS} - \log p_r^{HA}$, and I_{HA} is 1 if the label is from HA data and 0 otherwise.

5.3.6 Analysis of the Disagreement Penalty

Let $\mathbf{w}_{HA}$ be a learnable parameter of *HA-Net* which predicts relations in the test time. We investigate the effect of the disagreement penalty by comparing the gradients of loss functions with respect to $\mathbf{w}_{HA}$ for a human annotated label and a distantly supervised label.

For a label $\langle e_h, r, e_t \rangle$, let $\phi_r = (\log (p_r^{DS}/p_r^{HA}) - \mu_r)/\sigma_r^2$. If the label is human annotated, we obtain the following gradient of the loss $L_{h,t}$ with respect to $\mathbf{w}_{HA}$ from Equation (5.4)

$$\nabla L_{h,t} = \nabla L_{h,t}^{HA} + \mathbf{0} + \lambda \nabla L_{h,t}^{DS\text{-}HA} = -\left(1 + \lambda(1 + \phi_r)\right) \frac{1}{p_r^{HA}} \nabla p_r^{HA}. \tag{5.5}$$

On the other hand, if the label is annotated by distant supervision, the gradient becomes

$$\nabla L_{h,t} = \mathbf{0} + \mathbf{0} + \lambda \nabla L_{h,t}^{DS\text{-}HA} = -\lambda \left(1 + \phi_r\right) \frac{1}{p_r^{HA}} \nabla p_r^{HA}. \tag{5.6}$$

The two gradients in Equations (5.5) and (5.6) have the same direction of $-\nabla p_r^{HA}$. It implies that a human annotated label and a distantly supervised label have similar effects on training *HA-Net* except that the magnitudes of gradients are calibrated by $1+\lambda(1+\phi_r)$ and $\lambda(1+\phi_r)$, respectively. Thus, *HA-Net* can learn from not only human annotated labels but also distantly supervised labels by introducing the disagreement penalty. Recall that the log-normal distribution $L(\mu_r, \sigma_r)$ describes the conditional inflation for a given sentence with a head entity and a tail entity. If the median e^{μ_r} of $L(\mu_r, \sigma_r)$ has a high value, the distantly supervised label is likely to be a false label. Thus, we decrease the size of ϕ_r to reduce the effect of a distantly supervised label. On the other hand, as the median e^{μ_r} becomes lower, the size of ϕ_r increases to aggressively utilize the distantly supervised label.

5.3.7 Extension to Document-level Relation Extraction

For the document-level RE, we use the *binary* cross entropy as the prediction losses $L^{HA}_{h,t}$ and $L^{DS}_{h,t}$. For a pair of entities hand t, let $R_{h,t}$ be the set of relation types between the entities. In the train time, we use the following loss function for document relation extraction

$$\begin{aligned} L_{h,t} = &- I_{HA} \left(\sum_{r \in R_{h,t}} \log p_r^{HA} + \sum_{r \in R \setminus R_{h,t}} \log \left(1 - p_r^{HA}\right) \right) \\ &- (1 - I_{HA}) \left(\sum_{r \in R_{h,t}} \log p_r^{DS} + \sum_{r \in R \setminus R_{h,t}} \log \left(1 - p_r^{DS}\right) \right) \\ &+ \lambda \sum_{r \in R_{h,t}} \left[\frac{1}{2} \left(\frac{\ell_r - \mu_r}{\sigma_r} \right)^2 + \ell_r + \log \sigma_r \right]. \end{aligned}$$

where $\ell_r = \log p_r^{DS} / p_r^{HA}$, and I_{HA} is 1 if the labels are from HA data and 0 otherwise. We obtain the same property shown in Section 5.3.6 for the above loss function. In the test time, we regard that the model outputs the triple $\langle e_h, r, e_t \rangle$ if p_r^{HA} is greater than a threshold which is tuned on the development dataset.

5.4 Experiments

We conducted performance study for sentence-level and document-level REs by following the experimental settings of [78] and [77, 73], respectively. All models are implemented in PyTorch and trained on a V100 GPU. We initialized *HA-Net* and *DS-Net* to have the same initial parameters.

5.4.1 Experimental Settings

Compared methods. We compare our dual supervision framework, denoted by ***DUAL***, with the state-of-the-art methods ***BASet*** and ***BAFix*** in [78]. For sentence-level RE, we compare *DUAL* with two additional baselines ***MaxThres*** [63] and ***EntThres*** [49]

Data	Number of instances				# of relation types
	Train-HA	Train-DS	Dev	Test	
KBP	378	132,369	14,103	1,488	7
NYT	756	323,126	34,871	3,021	25
DocRED	38,269	1,508,320	12,332	12,842	96

Table 5.2: Statistics of datasets

which are only applicable to multi-class classification and cannot be used in document-level RE. *MaxThres* outputs `NA` if the maximum output probability is less than a threshold. Similarly, *EntThres* outputs `NA` if the entropy of the output probability distribution is greater than a threshold. While our dual supervision framework uses both types of data to train relation extraction models, the existing methods first train models on DS data and adjust the bias of the output layer or the output threshold by using HA data. We implemented an additional baseline named *DS+HA* which trains relation extraction models on both DS data and HA data with the original architecture of the models.

Dataset. KBP [48, 19] and NYT [64, 27] are datasets for sentence-level RE, and DocRED [77] is a dataset for document-level RE. The statistics of the datasets are summarized in Table 5.2. Since KBP and NYT do not have HA train data, we use 20% of the HA test data as the HA train data. In addition, we randomly split 10% of train data on KBP and NYT for the development (dev) data. Note that the ground truth of the test data in DocRED is not publicly available. However, we can get the F1 score of the result extracted from the test data by submitting the result to the DocRED competition hosted by CodaLab (available at https://competitions.codalab.org/competitions/20717). We report both the F1 scores computed from the dev data and the test data. Note that document-level RE data has much more training instances than the sentence-level relation extraction datasets.

Used relation extraction models. For *sentence-level RE*, we used the six models: ***BiGRU$_S$*** [85], ***PaLSTM$_S$*** [85], ***BiLSTM$_S$*** [85], ***CNN$_S$*** [83], ***PCNN$_S$*** [82], and ***BERT$_S$*** [73]. On the other hand, for *document-level RE*, we used the five models: ***BERT$_D$*** [73], ***CNN$_D$*** [83], ***LSTM$_D$*** [77], ***BiLSTM$_D$*** [6] and ***CA$_D$*** [67]. Note that *CNN$_D$*, *BiLSTM$_D$*, and *CA$_D$* are originally proposed for sentence-level RE and we used the adaptation of them to document-level RE provided in [77]. In addition, we adapt *BERT$_D$* to the sentence-level RE by changing the output activation function from sigmoid to softmax and denote it by *BERT$_S$*.

5.4.2 Implementation Details

Our implementation is available at `https://github.com/woohwanjung/dual`.

Sentence-level RE. We use the code which is made publicly available by [78] at `https://github.com/INK-USC/shifted-label-distribution`. All models except *BERT$_S$* are trained by stochastic gradient descent. Learning rate is initially set to 1.0, and decreased to 10% if there is no improvement on the dev data for 3 consecutive epochs. For the models, we set the hyperparameters λ and d to 10^{-3} and 200, respectively. To train *BERT$_S$* model, we used Adam optimizer with learning rate 10^{-5}. Moreover, the hyperparameters λ and d are set to 10^{-4} and 128, respectively. We alternately used an HA batch and a DS batch for dual supervision where an HA batch consists of training instances with human annotated labels and a DS batch consists of training instances with distantly supervised labels.

Document-level RE. For *BiLSTM$_D$*, *LSTM$_D$*, *CA$_D$* and *CNN$_D$*, we utilized the code which is available at `https://github.com/thunlp/DocRED` and implemented by [77]. In addition, we used the implementation of *BERT$_D$* that is available at `https://github.com/hongwang600/DocRed` and provided by [73]. We used Adam optimizer [35] to optimize the RE models. For the *BERT$_D$* model, we set the batch size to 12 and learning rate to 10^{-5}. For the other models, we followed the setting provided in [77]: batch size is 40, learning rate is 10^{-3}. We set the hyperparameters

RE models	*BiGRU$_S$*	*PaLSTM$_S$*	*BiLSTM$_S$*	*PCNN$_S$*	*CNN$_S$*	*BERT$_S$*
HA-Only	0.1984	0.1153	0.1787	0.3410	0.2586	0.1631
DS-Only	0.3909	0.3521	0.3519	0.2705	0.2810	0.3610
DS+HA	0.4375	0.4150	**0.4338**	**0.4067**	**0.3220**	0.3977
BASet	0.3972	0.4055	0.4053	0.2410	0.2400	0.3858
BAFix	0.4241	0.4027	0.3581	0.2931	0.2473	0.3383
MaxThres	0.4264	0.3630	0.4053	0.2815	0.2645	0.3751
EntThres	0.4470	0.4018	0.4248	0.2925	0.2826	0.3539
DUAL	**0.4749**	**0.4420**	0.4207	0.3872	0.2969	**0.4013**

Table 5.3: Sentence-level RE (KBP)

λ and d to 10^{-5} and 128, respectively. Each training batch has half of the instances with human-annotated labels and the other half of instances with distantly supervised labels.

5.4.3 Comparison with Existing Methods

We compare the dual supervision framework with the existing methods on the sentence-level and document-level RE datasets.

Sentence-level RE. Table 5.3 and Table 5.4 show F1 scores for relation extraction on KBP and NYT, respectively. Note that *DS-Only* and *HA-Only* represent the original RE models trained only on distantly supervised and human-annotated labels, respectively. *DUAL* shows the highest F1 scores with all RE models except *BiLSTM$_S$*. Since KBP and NYT have a small number of human-annotated labels in train data, *HA-Only* shows worse F1 scores than *DS-Only*. Furthermore, *DUAL* achieves improvements of F1 score from 5% to 40% over *DS-Only* by additionally using the small amount of human annotated labels. On the other hand, the compared methods *BAFix*, *BASet*,

RE models	$BiGRU_S$	$PaLSTM_S$	$BiLSTM_S$	$PCNN_S$	CNN_S	$BERT_S$
HA-Only	0.0884	0.1259	0.1504	0.4463	0.3978	0.1953
DS-Only	0.4532	0.4429	0.4297	0.4177	0.4463	0.4625
DS+HA	0.5185	0.4662	**0.4764**	0.1387	0.2350	0.4027
BASet	0.4966	0.4555	0.4561	0.3584	0.4358	0.5081
BAFix	0.4613	0.4507	0.4707	0.4023	0.4532	0.5145
MaxThres	0.4531	0.4462	0.4350	0.4258	0.4655	0.4952
EntThres	0.4553	0.4472	0.4210	0.4154	0.4427	0.4940
DUAL	**0.5455**	**0.5210**	0.4524	**0.4986**	**0.4744**	**0.5300**

Table 5.4: Sentence-level RE (NYT)

MaxThres and *EntThres* often perform worse than *DS-Only* and *HA-Only*. Interestingly, with three RE models, the baseline *DS+HS* outperforms all other methods in KBP dataset which has a very few human annotated labels. This result implies that even with a few human annotated labels, it is more effective to use the labels to train relation extraction models than to use it to adjust the threshold or bias. In NYT dataset, the dual supervision framework outperforms all other methods with all relation extraction models except for *BiLSTM*. We will provide a detailed comparison of *DS+HS* and our dual supervision framework in Section 5.4.4 with varying the number of human annotated labels.

Document-level RE. We present F1 scores on DocRED in Table 5.5 and Table 5.6. *DUAL* outperforms all compared methods with all RE models. Especially, the F1 score of dual framework with $BERT_D$ shows more than 22% of improvement over *BASet* and *BAFix*. Since DocRED has a large human-annotated train data, *HA-Only* shows better performance than *DS-Only*. For $BERT_D$ and CNN_D, the existing methods show lower F1 scores compared to *HA-Only*. It shows that the accuracy can be degraded although

RE models	$BERT_D$	$BiLSTM_D$	CA_D	$LSTM_D$	CNN_D
HA-Only	0.5513	0.4992	0.4986	0.4817	0.4788
DS-Only	0.4683	0.4951	0.4890	0.4877	0.4166
DS+HA	0.5263	0.5389	0.5291	0.5313	0.4914
BASet	0.4807	0.5123	0.5024	0.5012	0.4349
BAFix	0.4802	0.5136	0.5070	0.5166	0.4365
DUAL	**0.5880**	**0.5510**	**0.5372**	**0.5392**	**0.4967**

Table 5.5: Document-level RE (DocRED: Dev)

we use additional DA data in addition to HA data due to the labeling bias. Meanwhile, we achieve a consistent and significant improvement by applying *DUAL*. In the rest of this section, we will provide detailed evaluation of performance on DocRED data which is the largest dataset in this experiment. For the test data of DocRED, the ground truth is not publicly available and only a F1 score can be obtained from the DocRED competition. Thus, we provide detailed evaluations of performance on the dev data only.

Through experiments with sentence-level relation extraction and document-level relation extraction tasks, we have found that our framework generally and significantly improves relation extraction performance. Therefore, we expect dramatic performance gains when applying our framework to relation extraction models in the knowledge fusion task.

Inflation vs. accuracy. To investigate the effect of the inflation to the accuracy of relation extraction, we split the relation types into 4 groups based on the inflation of the relation types. In Figure 5.3, we present the characteristics of each group and plot the F1 scores by groups for $BERT_D$ model and $BiLSTM_D$ model. All methods have the highest F1 scores when the inflation is close to 1 (at the 2nd group). Furthermore, the

RE models	$BERT_D$	$BiLSTM_D$	CA_D	$LSTM_D$	CNN_D
HA-Only	0.5478	0.4982	0.4992	0.4815	0.4681
DS-Only	0.4587	0.4809	0.4772	0.4713	0.4160
DS+HA	0.5244	0.5280	0.5203	0.5229	0.4771
BASet	0.4716	0.4949	0.4905	0.4905	0.4320
BAFix	0.4730	0.5061	0.4989	0.4977	0.4354
DUAL	**0.5774**	**0.5379**	**0.5306**	**0.5277**	**0.4909**

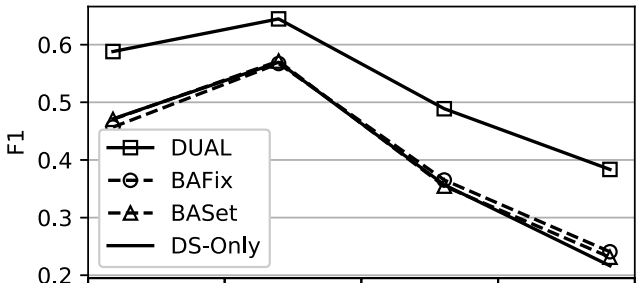
Table 5.6: Document-level RE (DocRED: Test)

improvement of F1 score by *DUAL* compared to the second best performer increases as the inflation moves away from 1. Thus, it confirms that our dual supervision framework effectively utilizes both human annotation and distant supervision by modeling the bias of the distant supervision. Since the other models CA_D, $LSTM_D$ and CNN_D show similar results with $BiLSTM_D$, we omit the result.

Precision-recall curves. The precision-recall curves of the compared methods are shown in Figure 5.4. As expected, *DUAL* consistently outperforms all compared methods. *BAFix* and *BASet* have similar precision-recall curves with *DS-Only*. Although *HA-Only* shows comparable precisions with *DUAL* when recall is low, the precision of *HA-Only* drops faster than that of *DUAL* with increasing recall. It implies that human annotated labels are not enough for training a model to extract a large number of relations. Meanwhile, *DUAL* extracts more relations from the document compared to existing models at the same precision level.

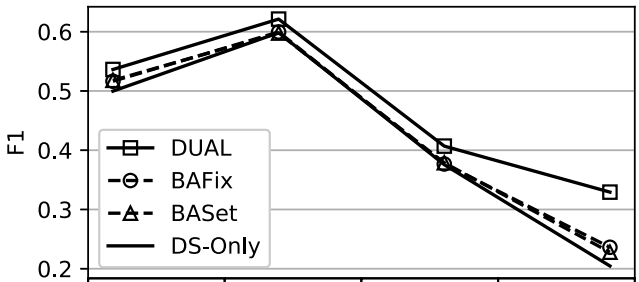
5.4.4 Ablation Study

We conducted an ablation study with the existing model $BERT_D$ on DocRED to validate the effectiveness of individual components of our framework. We compared



	Group #1	Group #2	Group #3	Group #4
Inflation	0.48~0.88	0.91~1.05	1.06~1.70	1.78~85.90
# Rel. types	24	24	24	24
# Instances	1,895	7,064	2,450	914
Improvement	+24.79%	+12.80%	+33.83%	+59.30%

(a) $BERT_D$



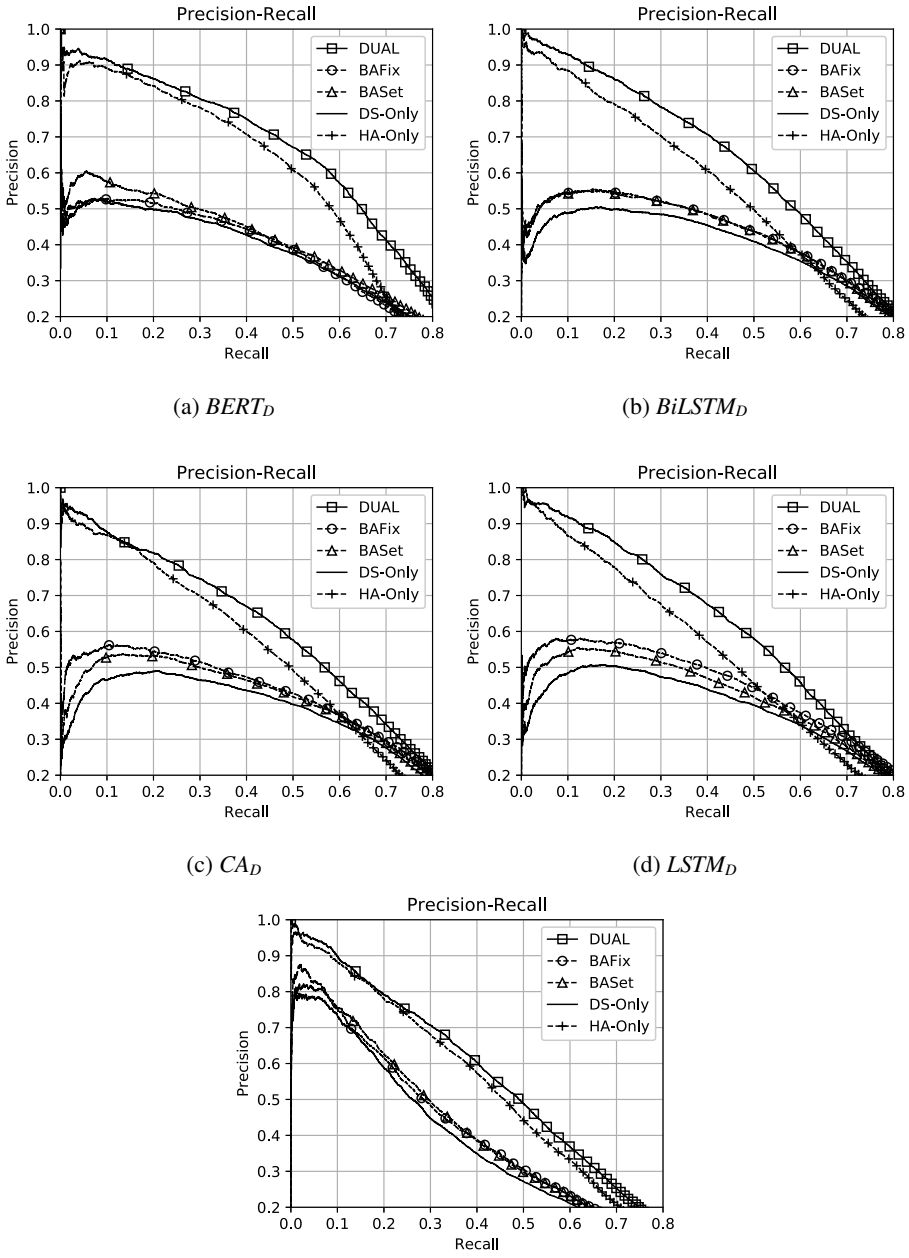
	Group #1	Group #2	Group #3	Group #4
Inflation	0.48~0.88	0.91~1.05	1.06~1.70	1.78~85.90
# Rel. types	24	24	24	24
# Instances	1,895	7,064	2,450	914
Improvement	+3.54%	+3.63%	+7.42%	+39.41%

(b) $BiLSTM_D$

Figure 5.3: F1 scores of different groups

DUAL (separate prediction networks + disagreement penalty) and two variations of our framework *Multitask* (separate prediction networks only) and *Single*. *Multitask* denotes a variation of *DUAL* which does not utilize the disagreement penalty while *Single* is the baseline *DS+HA* introduced in Section 5.4.1. Note that *Single* is also trained on both HA data and DS data together.

To show the effectiveness of the components depending on the size of HA data, we plotted the F1 scores with varying the number of human-annotated documents



(a) $BERT_D$ (b) $BiLSTM_D$

(c) CA_D (d) $LSTM_D$

(e) CNN_D

Figure 5.4: Precision-recall curves

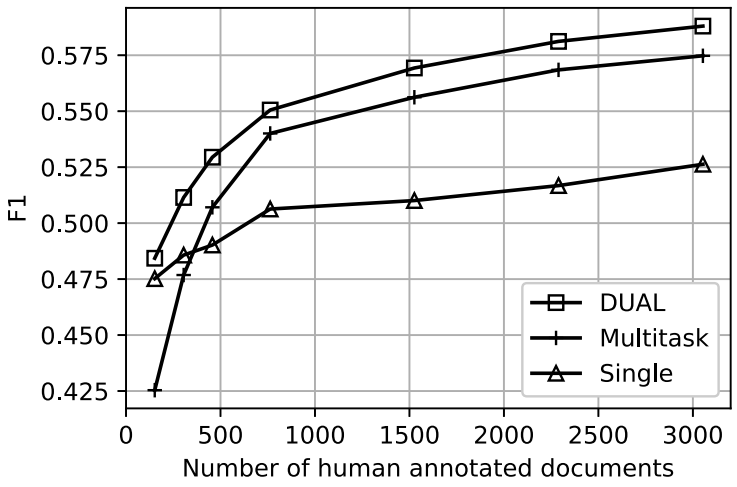


Figure 5.5: Varying the size of HA data

from 152 to 3,053 (i.e., from 5% to 100% of the documents in HA data) in Figure 5.5. As we expected, *DUAL* outperforms both variations in all settings. Furthermore, separation of the prediction networks significantly improves the accuracy when we have enough number of human-annotated labels. However, when we use less than 10% of the human annotated documents, *Multitask* suffers from the sparsity problem. By utilizing the disagreement penalty additionally, *DUAL* outperforms *Single* even when we use only 5% of the human-annotated documents for training the model. It implies that the disagreement penalty enables *HA-Net* to effectively learn from DS data as well as HA data.

5.4.5 Quality Comparison

To give an idea of what false relations are found by existing methods, we provide two example documents in the dev data of DocRED and the relations extracted by *DUAL*, *BAFix* and *DS-Only* with $BERT_D$ in Table 5.7. The relation $\langle Sweden, \texttt{capital}, Stockholm \rangle$ is expressed in the document titled 'Kungliga Hovkapellet' and all methods find the relation correctly. In the document titled 'Loopline Bride', the relation

Documents		
	Title: Kungliga Hovkapellet [1] Kungliga Hovkapellet is a Swedish orchestra, originally part of the Royal Court in **[Sweden]**'s capital **[Stockholm]**. [2] Its existence ...	**Title: Loopline Bridge** [1] The Loopline Bridge (or the Liffey Viaduct) is a railway bridge spanning the River Liffey and several streets in **[Dublin]**, **[Ireland]**. [2] It joins ...
Relations		
Ground truth	⟨**Sweden**, `capital`, **Stockholm**⟩	`NA`
DUAL	⟨**Sweden**, `capital`, **Stockholm**⟩	`NA`
BAFix	⟨**Sweden**, `capital`, **Stockholm**⟩	⟨**Ireland**, `capital`, **Dublin**⟩
DS-Only	⟨**Sweden**, `capital`, **Stockholm**⟩	⟨**Ireland**, `capital`, **Dublin**⟩

Table 5.7: Examples of documents and extracted relations

$\langle Ireland, \texttt{capital}, Dublin \rangle$ does not exist. However, *BAFix* and *DS-Only* output the incorrect relation. Since *DUAL* adaptively assess the labeling bias with μ-Net and σ-Net, *DUAL* does not output the false relation. In addition, since the RE models trained with *BAFix* and *DS-Only* fail to learn the text pattern corresponding to the relation type due to the labeling bias, they output many false labels such as $\langle Vietnam, \texttt{capital}, Taipei \rangle$ in many documents. It shows that the dual supervision framework effectively deal with the labeling bias of distant supervision by considering contextual information.

5.4.6 Topic-aware Relation Extraction

We apply the dual supervision framework to the T-REX model proposed in Chapter 4. Similar to document-level relation extraction models, we use bilinear layers for μ-Net

	F1	AUC
HA-Only	0.6569	0.6456
DS-Only	0.6624	0.6978
DUAL	**0.6930**	**0.7125**

Table 5.8: Dual supervision with T-REX model for topic-aware RE

and σ-Net. Table 5.8 shows the performance of T-REX model on DocRED dataset trained with *HA-Only*, *DS-Only* and *DUAL*. The result shows that our dual supervision framework is also effective in the topic-aware RE task.

5.4.7 Generalization Performance

Recall that our goal is to extract new triples to populate the knowledge base. To verify the utility of the extracted triples for knowledge base population, we manually examine the triples extracted from a wikipedia article 'Lark Force'. Table 5.9 shows 23 triples extracted by *DUAL* +*BERT*$_D$. Among the 23 extracted triples, 18 triples are correct. In addition, we manually check whether each triple exists in wikidata knowledge base. We found that 7 correct triples do not exist in the knowledge base. The newly discovered triples can be used to populate the knowledge base. Consider two conflicting triples $\langle HMAT\_Zealandia,\ country,\ Australia \rangle$ and $\langle HMAT\_Zealandia,\ country,\ Japan \rangle$. Since document-level relation extraction models independently extract triples from a document, conflicting triples can be extracted at the same time. Thus, we think that it would be an interesting research direction to develop a relation extraction model which considers the relationships between the triples.

To quantitatively evaluate the generalization performance of *DUAL*, we provide the F1, IgnF1 and their difference in Table 5.10 where IgnF1 is the F1 score computed without the triples that exist in the training data. We can see that the gap between F1 and IgnF1 is the smallest when we use the dual supervision framework. The result

Head entity	Relation	Tail entity	Correct	Exists in KB
Lark Force	operator	Australian Army	O	X
Lark Force	inception	”March 1941”	O	X
Lark Force	conflict	World War II	O	O
Lark Force	country	Australia	O	O
Lark Force	operator	Imperial Japanese Army	X	-
Australian Army	conflict	World War II	O	O
Australian Army	country	Australia	O	O
John Scanlan	military branch	Australian Army	O	X
John Scanlan	conflict	World War II	O	X
John Scanlan	country of citizenship	Australia	O	X
John Scanlan	military branch	Imperial Japanese Army	X	-
Australia	participant of	World War II	O	X
Australia	ethnic group	Japanese	X	-
MV Neptuna	country	Australia	O	O
HMAT Zealandia	country	Australia	O	O
HMAT Zealandia	country	Japan	X	-
Imperial Japanese Army	conflict	World War II	O	O
Imperial Japanese Army	country	Australia	X	-
Imperial Japanese Army	country	Japan	O	O
Imperial Japanese Army	country	Japanese	O	O
Japan	participant of	World War II	O	X
Japan	ethnic group	Japanese	O	O
USS Sturgeon	conflict	World War II	O	O

Table 5.9: Triples extracted from a wikipedia article ‘Lark Force’

implies that the generalization performance of the dual supervision is higher than those of the existing methods.

	F1	IgnF1	F1-IgnF1
HA-Only	0.5513	0.4949	0.0564
DS-Only	0.4683	0.3556	0.1127
BAFix	0.4802	0.3720	0.1082
BASet	0.4807	0.3622	0.1185
DUAL	**0.5880**	**0.5574**	**0.0306**

Table 5.10: Evaluation of the generalization performance

Chapter 6

Truth Discovery in the Presence of Hierarchies

6.1 Motivation

As pointed out in [13, 15, 44], the extracted values can be hierarchically structured. In this case, there may be multiple correct values in the hierarchy for an object even for functional predicates and we can utilize them to find the most specific correct value among the candidate values. For example, consider the three claimed values of 'NY', 'Liberty Island' and 'LA' about the location of the Statue of Liberty in Table 6.1. Because Liberty Island is an island in NY, 'NY' and 'Liberty Island' do not conflict with each other. Thus, we can conclude that the Statue of Liberty stands on Liberty Island in NY.

We also observed that many sources provide generalized values in the real-life. Figure 6.1 shows the graph of the generalized accuracy against the accuracy of the sources in the real-life datasets *BirthPlaces* and *Heritages* used for experiments in Section 6.3. The accuracy and the generalized accuracy of a source are the proportions of the exactly correct values and hierarchically-correct values among all claimed values, respectively. If a source claims the exactly correct values without generalization, it is located at the dotted diagonal line in the graph. This graph shows that many sources in real-life datasets claim with generalized values and each source has its own

Object	Source	Claimed value
Statue of Liberty	UNESCO	NY
Statue of Liberty	Wikipedia	Liberty Island
Statue of Liberty	Arrangy	LA
Big Ben	Quora	Manchester
Big Ben	tripadvisor	London

Table 6.1: Locations of tourist attractions

tendency of generalization when claiming values.

Most of the existing methods [88, 58, 89, 14, 16] simply regard the generalized values of a correct value as incorrect. Thus, it causes a problem in estimating the reliabilities of sources. According to [15], 35% of the false negatives in the data fusion task are produced by ignoring such hierarchical structures. Note that there are many publicly available hierarchies such as WordNet [70] and DBpedia [1]. Thus, a truth discovery algorithm to incorporate hierarchies is proposed in [4]. However, it does not consider the different tendencies of generalization and may lead to the degradation of the accuracy. Another drawback is that it needs a threshold to control the granularity of the estimated truth.

We propose a novel probabilistic model to capture the different generalization tendencies shown in Figure 6.1. Existing probabilistic models [58, 89, 14, 16] basically assume two interpretations of a claimed value (i.e., correct and incorrect). By introducing three interpretations of a claimed value (i.e., exactly correct, hierarchically correct, and incorrect), our proposed model represents the generalization tendency and reliability of the sources.

In this chapter, we propose a truth discovery algorithm utilizing the hierarchical structures in claimed values. To the best of our knowledge, it is the first work which

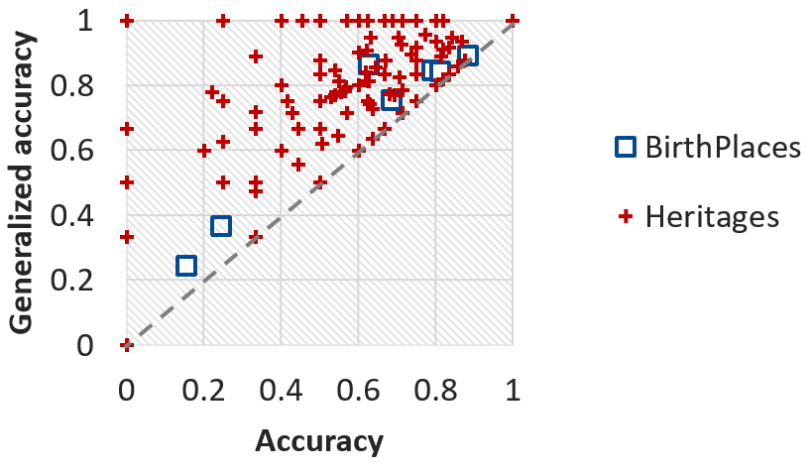


Figure 6.1: Generalization tendencies of the sources

considers both the reliabilities and the generalization tendencies of the sources. Note that our proposed truth discovery algorithm can also work without the answers obtained from workers. We empirically show that the proposed algorithm outperforms the existing works with experiments on real-life datasets.

6.2 Hierarchical Truth Discovery

For the hierarchical truth discovery, we first model the trustworthiness of sources and workers for a given hierarchy. Then, we propose a probabilistic model to describe the process of generating the set of records and the set of answers based on the trustworthiness modeling. We next develop an inference algorithm to estimate the model parameters and determine the truths.

6.2.1 Our Generative Model

Our probabilistic graphical model in Figure 6.2 expresses the conditional dependence (represented by edges) between random variables (represented by nodes). While the previous works [11, 75, 31, 61] assume that all sources and workers have their own

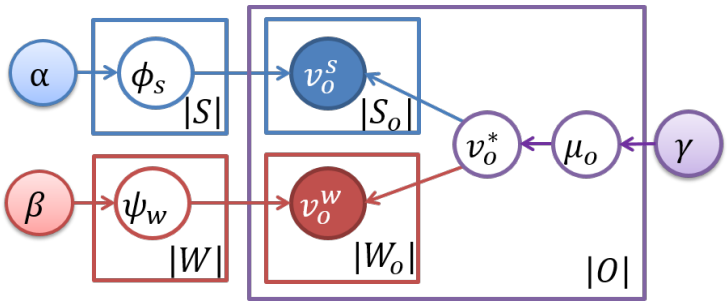


Figure 6.2: A graphical model for truth discovery

reliabilities only, we assume that each source or worker has its generalization tendency as well as reliability. We first describe how sources and workers generate the claimed values based on their trustworthiness. We next present the model for generating the true value. Finally, we provide the detailed generative process of our probabilistic model.

Model for source trustworthiness. For an object o, let v_o^* be the truth and v_o^s be the claimed value reported by a source s. Recall that V_o is the set of candidate values for an object o. Furthermore, we let $G_o(v)$ denote the set of candidate values which are ancestors of a value v except for the root in the hierarchy H.

There are three relationships between a claimed value v_o^s and the truth v_o^*: (1) $v_o^s = v_o^*$, (2) $v_o^s \in G_o(v_o^*)$ and (3) otherwise. Let $\phi_s = (\phi_{s,1}, \phi_{s,2}, \phi_{s,3})$ be the *trustworthiness distribution* of a source s where $\phi_{s,i}$ is the probability that a claimed value of the source s corresponds to the i-th relationship. In each relationship, a claimed value is generated as follows:

- **Case 1** ($v_o^s = v_o^*$)**:** The source s provides the exact true value with a probability $\phi_{s,1}$.
- **Case 2** ($v_o^s \in G_o(v_o^*)$)**:** The source s provides a *generalized true value* v_o^s with a probability $\phi_{s,2}$. In this case, the claimed value is an ancestor of the truth v_o^* in H. We assume that the claimed value is uniformly selected from $G_o(v_o^*)$.
- **Case 3** ($otherwise$)**:** The source s provides a wrong value v_o^s not even in $G_o(v_o^*)$.

The claimed value is uniformly selected among the rest of the candidate values in V_o.

The probability distribution ϕ_s is an initially-unknown model parameter to be estimated in our inference algorithm. Accordingly, the probability of selecting an answer v_o^s among the values in V_o for an object o is represented by

$$P(v_o^s|v_o^*, \phi_s) = \begin{cases} \phi_{s,1} & \text{if } v_o^s = v_o^*, \\ \frac{\phi_{s,2}}{|G_o(v_o^*)|} & \text{if } v_o^s \in G_o(v_o^*), \\ \frac{\phi_{s,3}}{|V_o|-|G_o(v_o^*)|-1} & \text{otherwise.} \end{cases} \tag{6.1}$$

For the prior of the distribution ϕ_s, we assume that it follows a Dirichlet distribution $Dir(\alpha)$, with a hyperparameter $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, which is the conjugate prior of categorical distributions.

Let O_H be the set of objects who have an ancestor-descendant relationship in their candidate set. In practice, there may exist some objects whose candidate values do not have an ancestor-descendant relationship. In this case, the probability of the second case (i.e., $\phi_{s,2}$) may be underestimated. Thus, if there is no ancestor-descendant relationship between the claimed values about o (i.e., $o \notin O_H$), we assume that a source generates its claimed value v_o^s with the following probability

$$P(v_o^s|v_o^*, \phi_s) = \begin{cases} \phi_{s,1} + \phi_{s,2} & \text{if } v_o^s = v_o^*, \\ \frac{\phi_{s,3}}{|V_o|-1} & \text{otherwise.} \end{cases} \tag{6.2}$$

Model for worker trustworthiness. Let v_o^w be the claimed value chosen by a worker w among the candidates in V_o for an object o. Similar to the model for source trustworthiness, we also assume the three relationships between a claimed value v_o^w and the truth v_o^*: (1) $v_o^w = v_o^*$, (2) $v_o^w \in G_o(v_o^*)$ and (3) otherwise. Each worker w has its *trustworthiness distribution* $\psi_w = (\psi_{w,1}, \psi_{w,2}, \psi_{w,3})$ where $\psi_{w,i}$ is the probability that an answer of the worker w corresponds to the i-th relationship. We assume

that the trustworthiness distribution is generated from $Dir(\beta)$ with a hyperparameter $\beta = (\beta_1, \beta_2, \beta_3)$.

Since it is difficult for the workers to be aware of the correct answer for every object, a worker can refer to web sites to answer the question. In such a case, if there is a widespread misinformation across multiple sources, the worker is also likely to respond with the incorrect information. Similar to [16, 58], we thus exploit the *popularity* of a value in Cases 2 and 3 to consider such dependency between sources and workers.

- **Case 1 ($v_o^w = v_o^*$):** The worker w provides the exact true value with a probability $\psi_{w,1}$.
- **Case 2 ($v_o^w \in G_o(v_o^*)$):** The worker w provides a generalized true value with a probability $\psi_{w,2}$. We assume that the claimed value v_o^w is selected according to the popularity $Pop_2(v_o^w|v_o^*) = \frac{|\{s|s \in S_o, v_o^s = v\}|}{|\{s|s \in S_o, v_o^s \in G_o(v_o^*)\}|}$ which is the proportion of the records whose claimed value is v_o^w out of the records with generalized values of v_o^*.
- **Case 3 ($otherwise$):** The claimed value is selected from the wrong values according to the popularity $Pop_3(v_o^w|v_o^*) = \frac{|\{s|s \in S_o, v_o^s = v\}|}{|\{s|s \in S_o, v_o^s \notin G_o(v_o^*), v_o^s \neq v_o^*\}|}$.

By the above model, the probability of selecting an answer v_o^w for the truth v_o^* of an object o is formulated as

$$P(v_o^w|v_o^*, \psi_w) = \begin{cases} \psi_{w,1} & \text{if } v_o^w = v_o^*, \\ \psi_{w,2} \cdot Pop_2(v_o^w|v_o^*) & \text{if } v_o^w \in G_o(v_o^*), \\ \psi_{w,3} \cdot Pop_3(v_o^w|v_o^*) & \text{otherwise.} \end{cases} \tag{6.3}$$

Similar to the model for source trustworthiness, if there is no ancestor-descendant relationship in the candidate values of an object o, the probability of selecting a claimed value v_o^w is

$$P(v_o^w|v_o^*,\psi_w) = \begin{cases} \psi_{w,1} + \psi_{w,2} & \text{if } v_o^w = v_o^*, \\ \psi_{w,3} \cdot Pop_3(v_o^w|v_o^*) & \text{otherwise.} \end{cases} \tag{6.4}$$

Model for truth. We introduce the probability distribution over the candidate answers to determine the truth, called *confidence distribution*. Each object o has a confidence distribution $\mu_o = \{\mu_{o,v}\}_{v \in V_o}$ where $\mu_{o,v}$ is the probability that the value $v \in V_o$ is the true answer for o. We also use a dirichlet prior $Dir(\gamma_o)$ for the confidence distribution μ_o where $\gamma_o = \{\gamma_{o,v}\}_{v \in V_o}$ is a hyperparameter.

Based on the above three models, the generative process of our model works as follows.

Generative process. Given a set of objects O, a set of sources S and a set of workers W, our proposed model assumes the following generative process for the set of records R and the set of answers A:

1. Draw $\phi_s \sim Dir(\alpha)$ for each source $s \in S$
2. Draw $\psi_w \sim Dir(\beta)$ for each worker $w \in W$
3. For each object $o \in O$
   (a) Draw $\mu_o \sim Dir(\gamma_o)$
   (b) Draw a true value $v_o^* \sim Categorical(\mu_o)$
   (c) For each source $s \in S_o$
      i. Draw a value v_o^s following $P(v_o^s|v_o^*,\phi_s)$
   (d) For each worker $w \in W_o$
      i. Draw a value v_o^w following $P(v_o^w|v_o^*,\psi_w)$

6.2.2 Estimation of Model Parameters

We now develop an inference algorithm for the generative model. Let $\Theta = \boldsymbol{\phi} \cup \boldsymbol{\psi} \cup \boldsymbol{\mu}$ be the set of all model parameters where $\boldsymbol{\phi} = \{\phi_s | s \in S\}$, $\boldsymbol{\psi} = \{\psi_w | w \in W\}$ and

$\boldsymbol{\mu} = \{\mu_o | o \in O\}$. We propose an EM algorithm to find the maximum a posteriori (MAP) estimate of the parameters in our model.

The maximum a posteriori (MAP) estimator. Recall that $R = \{(o, s, v_o^s)\}$ is the set of records from the sources and $A = \{(o, w, v_o^w)\}$ is the set of answers from the workers. For every object o, each source $s \in S_o$ and each worker $w \in W_o$ generates its claimed values independently. Then, the likelihood of R and A based on our generative model is

$$P(R, A|\Theta) = \prod_{o \in O} \prod_{s \in S_o} P(v_o^s|\phi_s, \mu_o) \cdot \prod_{o \in O} \prod_{w \in W_o} P(v_o^w|\psi_w, \mu_o)$$

where the probability of generating a claimed value by a source or a worker becomes

$$P(v_o^s|\phi_s, \mu_o) = \sum_{v \in V_o} P(v_o^s|\phi_s, v_o^* = v) \cdot \mu_{o,v} \tag{6.5}$$

$$P(v_o^w|\psi_w, \mu_o) = \sum_{v \in V_o} P(v_o^w|\psi_w, v_o^* = v) \cdot \mu_{o,v}. \tag{6.6}$$

Consequently, the MAP point estimator is obtained by maximizing the log-posterior as

$$\hat{\Theta} = \arg\max_{\Theta} \{\log P(R, A|\Theta) + \log P(\Theta)\} = \arg\max_{\Theta} \mathbb{F} \tag{6.7}$$

where the objective function $\mathbb{F}$ is

$$\begin{aligned} \mathbb{F} = & \sum_{o \in O} \sum_{s \in S_o} \log \sum_{v \in V_o} P(v_o^s|\phi_s, v_o^* = v) \cdot \mu_{o,v} \\ & + \sum_{o \in O} \sum_{w \in W_o} \log \sum_{v \in V_o} P(v_o^w|\psi_w, v_o^* = v) \cdot \mu_{o,v} \\ & + \sum_{s \in S} \log p(\phi_s|\alpha) + \sum_{w \in W} \log p(\psi_w|\beta) + \sum_{o \in O} \log p(\mu_o|\gamma_o). \end{aligned} \tag{6.8}$$

Note that although we assumed that each claimed value is generated independently according to its probability distribution defined in Eq. (6.5) and (6.6), the dependencies between sources and workers are already considered in $Pop_2(v_o^w|v_o^*)$ and $Pop_3(v_o^w|v_o^*)$.

The EM algorithm. We introduce a random variable C_v to represent the type of the relationship between the claimed value v and the truth v_o^*. It is defined as follows:

$$C_v = \begin{cases} 1 & \text{if } v = v_o^*, \\ 2 & \text{if } v \in G_o(v_o^*), \\ 3 & \text{otherwise.} \end{cases}$$

In the **E-step**, we compute the conditional distributions of the hidden variables $C_{v_o^s}$, $C_{v_o^w}$ and v_o^* under our current estimate of the parameters Θ. Let $f_{o,s}^v$, $f_{o,w}^v$, $g_{o,s}^t$ and $g_{o,w}^t$ denote the conditional probabilities $P(v_o^* = v|v_o^s, \mu_o, \phi_s)$, $P(v_o^* = v|v_o^w, \mu_o, \psi_w)$, $P(C_{v_o^s} = t|\mu_o, \phi_s)$ and $P(C_{v_o^w} = t|\ \mu_o, \psi_w)$, respectively. Using Bayes' rule, we can update the conditional probabilities as shown in Figure 6.3 where $D_o(v) = \{v'|v \in G_o(v') \wedge v' \in V_o\}$ is the set of descendants of v among the candidate values and $\neg D_o(v) = V_o - D_o(v) - \{v\}$ is the set of candidate values each of which is neither a descendant of the value v nor the v itself.

In the **M-step**, we find the model parameters Θ that maximize our objective function $\mathbb{F}$. We first add Lagrange multipliers to enforce the constraints of model parameters.

$$\mathbb{L} = \mathbb{F} + \sum_{s \in S} \lambda_{\phi,s} \left(1 - \sum_{t=1}^{3} \phi_{s,t}\right) + \sum_{w \in W} \lambda_{\psi,w} \left(1 - \sum_{t=1}^{3} \psi_{w,t}\right) + \sum_{o \in O} \lambda_{\mu,o} \left(1 - \sum_{v \in V_o} \mu_{o,v}\right)$$

We obtain the following equations for updating the model parameters Θ by taking the partial derivative of the Lagrangian $\mathbb{L}$ with respect to each model parameter and setting it to zero:

$$\mu_{o,v} = \frac{\sum_{s \in S_o} f_{o,s}^v + \sum_{w \in W_o} f_{o,w}^v + \gamma_{o,v} - 1}{|S_o| + |W_o| + \sum_{v' \in V_o} (\gamma_{o,v'} - 1)} \tag{6.9}$$

$$\phi_{s,t} = \frac{\sum_{o \in O_s} g_{o,s}^t + \alpha_t - 1}{|O_s| + \sum_{t'=1}^{3} (\alpha_{t'} - 1)} \tag{6.10}$$

$$\psi_{w,t} = \frac{\sum_{o \in O_w} g_{o,w}^t + \beta_t - 1}{|O_w| + \sum_{t'=1}^{3} (\beta_{t'} - 1)} \tag{6.11}$$

$$f^v_{o,s} = \frac{P(v^s_o|v^*_o = v, \phi_s)\cdot\mu_{o,v}}{\sum_{v'\in V_o} P(v^s_o|v^*_o = v', \phi_s)\cdot\mu_{o,v'}}$$
$$f^v_{o,w} = \frac{P(v^w_o|v^*_o = v, \psi_w)\cdot\mu_{o,v}}{\sum_{v'\in V_o} P(v^w_o|v^*_o = v', \psi_w)\cdot\mu_{o,v'}}$$

$$g^1_{o,s} = \frac{\phi_{s,1}\cdot\mu_{o,v^s_o}}{\sum_{v\in V_o} P(v^s_o|v^*_o = v, \phi_s)\cdot\mu_{o,v}}$$
$$g^2_{o,s} = \begin{cases} \frac{\sum_{v\in D_o(v^s_o)} \frac{\phi_{s,2}}{|G_o(v)|}\cdot\mu_{o,v}}{\sum_{v\in V_o} P(v^s_o|v^*_o=v,\phi_s)\cdot\mu_{o,v}} & \text{if } o \in O_H \\ \frac{\phi_{s,2}\cdot\mu_{o,v^s_o}}{\sum_{v\in V_o} P(v^s_o|v^*_o=v,\phi_s)\cdot\mu_{o,v}} & \text{otherwise} \end{cases}$$
$$g^3_{o,s} = \frac{\sum_{v\in\neg D_o(v^s_o)} \frac{\phi_{s,3}}{|V_o - G_o(v)|-1}\cdot\mu_{o,v}}{\sum_{v\in V_o} P(v^s_o|v^*_o = v, \phi_s)\cdot\mu_{o,v}}$$

$$g^1_{o,w} = \frac{\psi_{w,1}\cdot\mu_{o,v^w_o}}{\sum_{v\in V_o} P(v^w_o|v^*_o = v, \psi_w)\cdot\mu_{o,v}}$$
$$g^2_{o,w} = \begin{cases} \frac{\sum_{v\in D_o(v^w_o)} \psi_{w,2}\cdot Pop_2(v^w_o|v^*_o=v)\cdot\mu_{o,v}}{\sum_{v\in V_o} P(v^w_o|v^*_o=v,\psi_w)\cdot\mu_{o,v}} & \text{if } o \in O_H \\ \frac{\psi_{w,2}\cdot\mu_{o,v^w_o}}{\sum_{v\in V_o} P(v^w_o|v^*_o=v,\psi_w)\cdot\mu_{o,v}} & \text{otherwise} \end{cases}$$
$$g^3_{o,w} = \frac{\sum_{v\in\neg D_o(v^w_o)} \psi_{w,3}\cdot Pop_3(v^w_o|v^*_o = v)\cdot\mu_{o,v}}{\sum_{v\in V_o} P(v^w_o|v^*_o = v, \psi_w)\cdot\mu_{o,v}}$$

Figure 6.3: E-step for the proposed truth inference algorithm

where O_s and O_w are the sets of objects claimed by s and w, respectively. We infer the truth by choosing the value with the maximum confidence among the candidate values as

$$v_o^* = \arg\max_{v \in V_o} \mu_{o,v}. \tag{6.12}$$

6.2.3 Extension to Numerical Data

In the world wide web, numerical data also have an implicit hierarchy due to the significant digits which carry meaning contributing to its measurement resolution. For example, even though the area of Seoul is $605.196km^2$, different websites may represent the area in various forms depending on the significant figures (e.g., $605.2km^2$, $605km^2$). An existing algorithm [40] to handle numerical data utilizes a weighted sum of the claimed values to consider the distribution of the claimed values. However, such method is sensitive to outliers and thus need a proper preprocessing to remove the outliers. To overcome the drawbacks, we generate the underlying hierarchy in the numerical data by assuming that v_d is a descendant of v_a if a value v_a can be obtained by rounding off a value v_d. Then, we can use our TDH algorithm to find the truths in numerical data by taking into account the relationship between the values in the implicit hierarchy. Our algorithm is also robust to the outliers with extremely small or large value since we estimate the truth by selecting the most probable value from the candidate values rather than computing a weighted average of the claimed values.

6.3 Experiments

6.3.1 Test Environments

The experiments are conducted on a computer with Intel i5-7500 CPU and 16GB of main memory.[1] In this section, we only provide the experimental results without

[1] Code is available at `https://github.com/woohwanjung/truthdiscovery`

crowdsourcing. The empirical evaluation with crowdsourcing will be presented in the next chapter after proposing the task assignment algorithm for crowdsourcing.

Datasets. We collected the two real-life datasets and made it publicly available at `http://kdd.snu.ac.kr/home/datasets/tdh.php`. Statistic of the datasets are summarized in Table 6.2.

BirthPlaces: We selected 6,005 celebrities and crawled 13,510 records about the birthplaces of the celebrities from 7 websites (sources). For the gold standard data to evaluate the correctness of discovered birthplaces, we used IMDb biography which is available at *http://www.imdb.com*. Moreover, the hierarchy was created by integrating the birth information from the IMDb data. For example, if there is a person who was born in 'LA, California, USA', we assigned 'LA' as a child of 'California' and 'California' as a child of 'USA'. The hierarchy contains 4,999 nodes (e.g., countries, cities and etc.) and its height is 5.

Heritages: This is a dataset of the locations of World Heritage Sites provided by UNESCO World Heritage Centre, available at *http://whc.unesco.org*. We queried about the locations of 785 World Heritage Sites with Bing Search API and obtained 4,424 claimed values from 1,577 distinct websites. Since we searched through the search engine without specifying a pool of sources, this dataset contains far more sources than *BirthPlaces*. Instead, each source in this dataset has a few claimed values. Thus, we can evaluate the when the data is sparse to accurately evaluate reliabilities of sources. The hierarchy was created in the same way as we did for *BirthPlaces* and it has 1,027 nodes. The height of this hierarchy tree is 6.

Quality Measures. We use *Accuracy*, *GenAccuracy* and *AvgDistance* to evaluate the truth discovery algorithms. Let t_o be the truth of the object o in the gold standard and v_o^* be the estimated truth by an algorithm. Note that t_o may not exist in the set of candidate values. In this case, the most specific candidate value among the ancestors of the truth is assumed to be t_o. *Accuracy* is the proportion of objects that the algorithm discovers the truth exactly. It is widely used in [89, 17, 91, 90] to evaluate truth discovery

	$\|O\|$	$\|R\|$	$\|S\|$	$\|h_o\|$
BirthPlaces	6,005	13,510	7	4,999
Heritages	785	4,424	1,577	1,027

Table 6.2: Statistics of datasets

algorithms.

$$(\textit{Accuracy}) = \frac{\sum_{o \in O} I(v_o^* = t_o)}{|O|}$$

The ancestors of t_o are less informative but still correct values. Thus, we develop an evaluation measure named *GenAccuracy* which is the proportion of objects o whose estimated truth v_o^* is either the truth t_o or an ancestor of the truth.

$$(\textit{GenAccuracy}) = \frac{\sum_{o \in O} I(v_o^* \in G_H(t_o) \cup \{t_o\})}{|O|}$$

Ancestors of the truth have a different level of informativeness depending on the distance to the truth. For example, 'New York' is more informative than 'USA' as the location of the Statue of Liberty. Thus, we utilize another evaluation measure named *AvgDistance* which weights the estimated truth based on the distance from the ground truth. More specifically, it is the average number of edges $d(v_o^*, t_o)$ between the truth t_o and the estimated truth v_o^* in the hierarchy H.

$$(\textit{AvgDistance}) = \frac{\sum_{o \in O} d(v_o^*, t_o)}{|O|}$$

AvgDistance is robust to the case where the ground truth is less specific than the estimated truth. The estimated truth is regarded as a wrong value when we compute *Accuracy* and *GenAccuracy* even though the estimate truth is correct and more specific. Since the distance between the less specific ground truth and the estimated truth is generally small, *AvgDistance* compensates the drawback of *Accuracy* and *GenAccuracy*.

6.3.2 Implemented Algorithms

We implemented following 10 truth discovery algorithms in Python for comparative experiments.

- TDH: This is our algorithm proposed in Section 6.2. For the prior distribution $Dir(\alpha)$, we set the hyperparameter $\alpha = (3, 3, 2)$ since correct values are more frequent than wrong values for most of the sources. For the other hyperparameters β and γ, we set every dimension of β and γ to 2.
- ACCU: It is the algorithm proposed in [14] which considers the dependencies between sources to find the truths. The algorithm exploits Bayesian analysis to find the dependencies.
- POPACCU: This denotes the algorithm in [16] which extends ACCU. It computes the distribution of the false values from the records while ACCU assumes that it is uniform.
- LFC: This algorithm is proposed in [61] and utilizes a confusion matrix to model a source's quality.
- CRH: It is proposed in [41] to resolve conflicts in heterogeneous data containing categorical and numerical attributes.
- LCA: It is a probabilistic model proposed in [58]. We select GuessLCA to be compared in this paper which is one of the best performers among the 7 algorithms proposed in [58].
- ASUMS: This is proposed in [4] by adapting an existing method SUMS [57] to hierarchical truth discovery.
- MDC: This denotes the truth discovery method designed for medical diagnose from non-expert crowdsourcing in [43].
- DOCS: This is the state-of-the-art technique presented in [89] that suggests the domain-sensitive worker model.

	Dataset					
	BirthPlaces			*Heritages*		
Algorithm	*Accuracy*	*GenAccuracy*	*AvgDistance*	*Accuracy*	*GenAccuracy*	*AvgDistance*
TDH	**0.8913**	**0.8988**	**0.3151**	**0.7414**	0.8726	**0.5210**
VOTE	0.7900	0.8924	0.4961	0.6892	**0.8994**	0.6382
LCA	0.8834	0.8923	0.3414	0.6930	0.8866	0.6611
DOCS	0.8828	0.8916	0.3409	0.6904	0.8866	0.6599
ASUMS	0.8543	0.8571	0.4573	0.6229	0.7414	1.2000
MDC	0.8263	0.8432	0.5320	0.7254	0.8087	0.6869
ACCU	0.8137	0.8296	0.6063	0.5834	0.7656	1.0637
POPACCU	0.8133	0.8300	0.6070	0.6561	0.8586	0.7554
LFC	0.8085	0.8743	0.4669	0.6803	0.8076	0.8076
CRH	0.8083	0.8271	0.6120	0.6841	0.8828	0.6688

Table 6.3: Performance of truth discovery algorithms

- VOTE: This is a baseline that selects a value with the highest frequency in the claimed values.

6.3.3 Comparison with Existing Truth Discovery Algorithms

We first provide the performances of the truth discovery algorithms without using crowdsourcing in Table. 6.3.

***BirthPlaces*.** Our TDH outperforms all other algorithms in terms of all quality measures since TDH finds the exact truths by utilizing the hierarchical relationships. Since TDH estimates the reliabilities of the sources and workers by considering the hierarchies, it does not underestimate the reliabilities of the sources and workers. Thus, TDH also finds more correct values including the generalized truths. We will discuss the reliability estimation in detail at the end of this section by comparing TDH with

ASUMS. LCA is the second-best performer and VOTE shows the lowest *Accuracy* among all compared algorithms. However, in terms of *GenAccuracy*, VOTE performs the second-best. It is because many websites claim the generalized values rather than the most specific value. As truth inference algorithms estimate the truths more specifically, the differences between *Accuracy* and *GenAccuracy* become smaller. Thus, TDH and ASUMS, which utilize the hierarchy information, have smaller differences between *Accuracy* and *GenAccuracy* compared to the other algorithms.

***Heritages*.** In terms of *AvgDistance* and *Accuracy*, TDH performs the best among those of the compared algorithms. VOTE shows the highest *GenAccuracy* because many sources provide the generalized truths. In fact, a high *GenAccuracy* with low *Accuracy* and *AvgDistance* can be easily obtained by providing the most general values for the truths. However, such values usually are not informative. Since our algorithm shows much higher *Accuracy* and much lower *AvgDistance* than VOTE, we can see that the estimated truth by TDH is more accurate and precise than the result from VOTE. *Heritages* contains many sources and most of the sources have a few claims. Thus, it is very hard to estimate the reliability of each source accurately. Therefore, most of the compared algorithms show worse performance than VOTE in terms of *AvgDistance*. In particular, ACCU has the lowest *Accuracy*. The reason is that ACCU requires many shared objects between two sources in order to accurately determine the dependency between the sources. The average accuracy of the sources in *Heritages* is 58.0% while that of the sources in *BirthPlaces* is 72.1%. Thus, every algorithm shows a lower *Accuracy* in this dataset than in *BirthPlaces*.

Comparison with ASUMS. Since ASUMS [4] is the only existing algorithm which utilizes hierarchies for truth inference, we show the statistics related to the reliability distributions estimated by TDH and ASUMS for *BirthPlaces* dataset in Figure 6.4. *Accuracy* and *GenAccuracy* represent the actual reliabilities of each source computed from the ground truths. Recall that $\phi_{s,1}$ and $\phi_{s,2}$ are the estimated probabilities of providing a correct value and a generalized correct value respectively for a source s

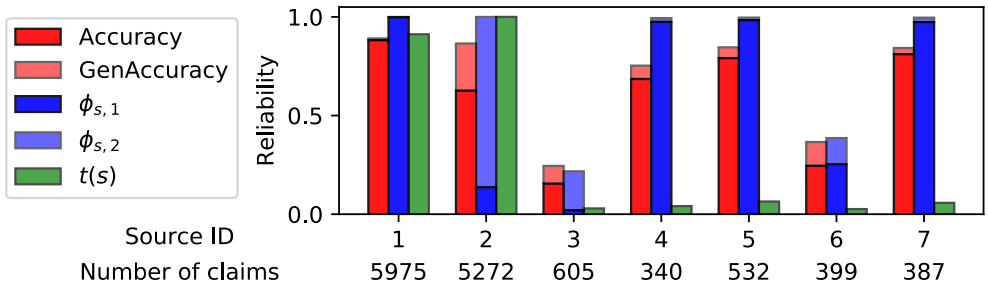


Figure 6.4: Source reliability distribution in *BirthPlaces*

by our TDH, as defined in Section 6.2. In addition, $t(s)$ is the estimated reliability of a source s by ASUMS which ignores the generalization level of each source. In each source s, the leftmost bar denotes accgen where the portion of *Accuracy* is also shown, the middle stacked bar shows $\phi_{s,1}$ and $\phi_{s,2}$ together, and the rightmost bar represents $t(s)$. The reliabilities of the sources 4, 5 and 7 computed by ASUMS (i.e. $t(s)$) are quite different from the actual reliabilities (i.e., *Accuracy*). As we discussed in Section 6.1, for a pair of sources that provide different claimed values with an ancestor-descendant relationship in a hierarchy, existing methods may assume that one of the claimed values is incorrect. Thus, the reliability of the source with the assumed wrong value tends to become lower by the existing methods. ASUMS suffers from the same problem and underestimates the reliabilities of the sources 4, 5 and 7 which provide a small number of claimed values. Meanwhile, our proposed algorithm TDH accurately estimates the reliabilities of the sources by introducing another class of the claimed values (generalized truth).

6.3.4 Comparison with Multi-truths Discovery Algorithms

Since there are multiple correct values including generalized values, The problem of truth dicovery in the presence of hierarchies can be regarded as a special case of the multi-truth discovery problem. We implement multi-truth discovery algorithms such as DART[46], LFC[61] and LTM[88] to compare with our TDH algorithm. Since

	Algorithm	Dataset					
		BirthPlaces			*Heritages*		
		Precision	Recall	F1	Precision	Recall	F1
Single truth	TDH	**0.899**	**0.921**	**0.910**	0.873	0.795	**0.832**
	VOTE	0.892	0.804	0.846	**0.899**	0.717	0.798
	LCA	0.892	0.913	0.903	0.878	0.711	0.786
	DOCS	0.892	0.913	0.902	0.887	0.722	0.796
	ASUMS	0.857	0.888	0.872	0.741	0.660	0.698
	POPACCU	0.847	0.858	0.852	0.859	0.694	0.768
	LFC	0.874	0.838	0.856	0.808	0.727	0.765
	MDC	0.844	0.853	0.848	0.807	0.792	0.800
	ACCU	0.830	0.842	0.836	0.766	0.631	0.692
	CRH	0.827	0.833	0.830	0.883	0.716	0.791
Multi -truths	LFC-MT	0.763	0.723	0.742	0.898	0.684	0.777
	DART	0.590	0.855	0.698	0.357	**0.994**	0.525
	LTM	0.780	0.472	0.588	0.871	0.672	0.759

Table 6.4: Comparison with multi-truth discovery algorithms

the multi-truths discovery algorithms independently generate the correct values, they may output the true values where there exist a pair of true values without ancestor-descendant relationship in the hierarchy. For example, from the given claimed values in Table 6.1, the multi-truth algorithms can answer that the 'Statue of Liberty' is located in LA and Liberty island. In this case, we cannot evaluate the result by our evaluation measures *Accuracy*, *GenAccuracy* and *AvgDistance*. Thus, to evaluate the performance of the tested algorithms, we utilize precision, recall and F1-score which are the evaluation measures typically used for multi-truths discovery. To use the multi-truths algorithms and the evaluation measures, we treat the ancestors of v and v itself

as the multi-truths of v. LFC can work as either a single truth algorithm or a multi-truths algorithm. We refer to the multi-truth version of LFC as LFC-MT to avoid the confusion.

Table 6.4 shows the performance of the truth discovery algorithms in terms of precision, recall and F1-score. For both datasets, the TDH algorithm is the best in terms of F1-score. Recall that the VOTE algorithm tends to find a generalized value of the exact truth. Since a generalized truth generates a small number of multi-truths, the VOTE algorithm shows the highest precision in *Heritages* dataset. However, since its recall is much lower than that of our TDH algorithm, the F1-score of the VOTE algorithm is lower than that of the TDH algorithm. Similarly, although the DART algorithm has the highest recall in *Heritages* dataset, the precision of the DART algorithm is the smallest among the precisions of all compared algorithms.

6.3.5 Performance Evaluation on a Numerical Dataset

To evaluate the extension to numerical data, we conducted an experiment on the stock datatset [42] which is trading data of 1000 stock symbols from 55 sources on every work day in July 2011. The detailed description of the data can be found in [42]. As we discussed at the end of Section 6.2.2, we can utilize our TDH algorithm for numeric dataset with implied hierarchy. We select three attributes 'change rate', 'open price' and 'EPS' of the dataset, and compared our TDH algorithm with the LCA, CRH, CATD[40], VOTE and MEAN algorithms. Note that CRH[41] and CATD[40] are designed to find the truth in numerical data. Recall that VOTE is a baseline algorithm which selects the candidate value collected from majority sources. We also implemented a baseline algorithm, called MEAN, which estimates the correct value as the average of the claimed numeric values.

Table 6.5 shows the mean squared error (MAE) and the relative error (R/E) of the tested algorithms. The TDH algorithm performs the best for every attribute. The MEAN and CATD algorithms show worse performance than the other algorithms.

	Change rate		Open price		EPS	
Algorithm	MAE	R/E	MAE	R/E	MAE	R/E
TDH	**0.0006**	**0.1011**	**0.0195**	**0.0354**	**0.0352**	**1.9513**
LCA	**0.0006**	**0.1011**	**0.0195**	**0.0354**	0.3831	16.2212
CRH	0.0020	1.6339	**0.0195**	**0.0354**	0.0610	1.9882
CATD	0.0104	2.3529	0.0211	0.0395	0.0803	3.2059
VOTE	**0.0006**	**0.1011**	**0.0195**	**0.0354**	0.0765	2.8402
MEAN	0.2837	30.8747	0.4047	0.5782	0.1762	7.3937

Table 6.5: Performance evaluation for numerical data

Since they utilize an average or a weighted average of the claimed values, they are sensitive to outliers. The result confirms that our TDH algorithm is effective even for numerical data.

Chapter 7

Task Assignment for Truth Discovery

7.1 Motivation

According to [15], upto 96% of the false triples are made by extraction errors rather than by the sources themselves. Human can easily correct the extraction errors by directly checking the information sources. Since crowdsourcing is an efficient way to utilize human intelligence with a low cost, it has been successfully applied in various areas of data integration such as schema matching [21], entity resolution [74], graph alignment [33] and truth discovery [91, 89]. Thus, we utilize crowdsourcing to improve the accuracy of the truth discovery.

It is essential in practice to minimize the cost of crowdsourcing by assigning proper tasks to workers. A popular approach for selecting queries in active learning is *uncertainty sampling* [38, 5, 34, 89]. It asks a query to reduce the uncertainty of the confidences on the candidate values the most. However, it considers only the uncertainty regardless of the accuracy improvement. QASCA algorithm [91] asks a query with the highest accuracy improvement, but measures the improvement without considering the number of collected claimed values. It can be inaccurate since an additional answer may be less informative for an object which already has many records and answers.

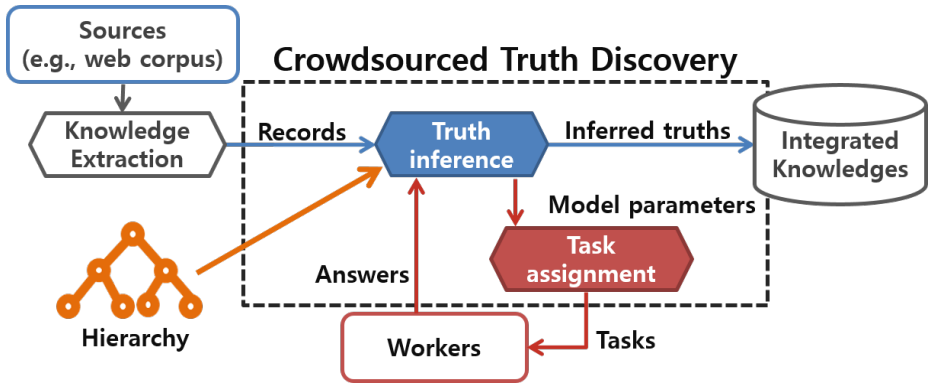


Figure 7.1: Crowdsourced truth discovery in KF

Assume that there are two candidate values of an object with equal confidences. If only a few sources provide the claimed values for the object, an additional answer from a crowd worker will significantly change the confidence distribution. Meanwhile, if hundreds of sources already provide the claimed values for the object, the influence of an additional answer is likely to be very little. Thus, we need to consider the number of collected answers as well as the current confidence distribution. Based on the observation, we develop a new method to estimate the increase of accuracy more precisely by considering the number of collected records and answers. We also present an incremental EM algorithm to quickly measure the accuracy improvement and propose a pruning technique to efficiently assign the tasks to workers.

As illustrated in Figure 7.1, our crowdsourced truth discovery for knowledge fusion consists of two components: *hierarchical truth discovery* and *task assignment*. The hierarchical truth discovery algorithm finds the correct values from the conflicting values, which are collected from different sources and crowd workers, using hierarchies. The task assignment algorithm distributes objects to the workers who are likely to increase the accuracy of the truth discovery the most. The proposed *crowdsourced truth discovery algorithm* repeatedly alternates the truth discovery and task assignment until the budget of crowdsourcing runs out. As discussed in [39], some workers answer slower than others and increase the latency. However, we do not investigate how

to reduce the latency in this work since we can utilize the techniques proposed in [23].

7.2 Task Assignment to Workers

In this section, we propose a task assignment method to select the best objects to be assigned to the workers in crowdsourcing systems. We first define a quality measure of tasks called *Expected Accuracy Increase (EAI)* and develop an incremental EM algorithm to quickly estimate the quality measure. Finally, we present an efficient algorithm for assigning the k questions to each worker w in a set of workers W based on the measure.

7.2.1 The Quality Measure

Given a worker w, our goal is to choose an object to be assigned to the worker w which is likely to increase the accuracy of the estimated truths the most. Thus, we define a quality measure for a pair of worker and an object based on the improvement of the accuracy. As discussed in [91], the improvement of the accuracy by a task can be estimated by using the difference between the highest confidence as follows:

$$(Accuracy\ improvement) = \frac{\max_v \mu_{o,v|w} - \max_v \mu_{o,v}}{|O|} \tag{7.1}$$

where $\mu_{o,v|w}$ is the estimated confidence on v if the worker w answers about an object o.

The quality measure used by QASCA. The QASCA[91] algorithm calculates the estimated confidence by using the current confidence distribution and the likelihood of the answer v_o^w given the truth $v_o^* = v$ as

$$\mu_{o,v|w} \propto \mu_{o,v} \cdot p(v_o^w = v' | v_o^* = v)$$

where v' is a sampled claimed value. There are two drawbacks in the quality measure of QASCA. First, since it computes the estimated confidence $\mu_{o,v|w}$ based on a sam-

pled answer $v_o^w = v$, the value of the quality measure is very sensitive to the sampled answer. In addition, QASCA does not consider the number of claimed values collected so far and the estimated confidence $\mu_{o,v|w}$ may not be accurate. For instance, assume that there exist two objects which have identical confidence distributions. If one of the objects already has many collected claimed values, an additional answer is not likely to change the confidence significantly. Thus, task assignment algorithms should select another object who has a smaller number of collected records and answers.

Our quality measure. To avoid the sensitiveness caused by sampling answers, we develop a new quality measure *Expected Accuracy Improvement (EAI)* which is obtained by taking the expectation to Eq. (7.1). That is,

$$EAI(w, o) = \frac{E[\max_v \mu_{o,v|w}] - \max_v \mu_{o,v}}{|O|}. \tag{7.2}$$

By the definition of expectation, $E[\max_v \mu_{o,v|w}]$ becomes

$$E[\max_v \mu_{o,v|w}] = \sum_{v' \in V_o} P(v_o^w = v'|\psi_w, \mu_o) \cdot \max_v \mu_{o,v|v_o^w=v'}. \tag{7.3}$$

where $\mu_{o,v|v_o^w=v'}$ is the conditional confidence when a worker w answers with v' about the object o.

Since $P(v_o^w = v'|\psi_w, \mu_o)$ can be computed by Eq. (6.6), to compute $E[\max_v \mu_{o,v|w}]$ by Eq. (7.3), we need the estimation of the conditional confidence $\mu_{o,v|v_o^w=v'}$ with an additional answer $v_o^w = v'$. Recall that the estimated confidence computed by QASCA may not be accurate because it does not consider the collected records and answers so far. To reduce the error, we use them to compute the conditional confidence $\mu_{o,v|v_o^w=v'}$. We can compute the conditional confidence $\mu_{o,v|v_o^w=v'}$ by applying the EM algorithm in Section 6.2.2 with the collected records and answers including $v_o^w = v'$. However, since it is computationally expensive, we next develop an *incremental EM algorithm*.

7.2.2 The Incremental EM Algorithm

Let $\mathbb{F}_{v_o^w=v'}$ be the objective function in Eq. (6.7) after obtaining an additional answer (o, w, v'). Then, we have

$$\mathbb{F}_{v_o^w=v'} = \mathbb{F} + \log \sum_{v \in V_o} P(v_o^w = v' | \psi_w, v_o^* = v) \cdot \mu_{o,v}$$

by adding the related term of the additional answer (log likelihood of the additional answer) to Eq. (6.8). Instead of running the iterative EM algorithm in Section 6.2.2, we incrementally perform a *single EM-step* to speed up for only the additional answer with the current model parameters and the above objective function.

E-step. Since we use the current model parameters, the probabilities of the hidden variables for collected records and answers are not changed. Thus, we only need to compute the conditional probabilities of the hidden variable given the additional answer as

$$f_{o,w|v_o^w=v'}^v = \frac{P(v_o^w = v' | v_o^* = v, \psi_w) \cdot \mu_{o,v}}{\sum_{v'' \in V_o} P(v_o^w = v' | v_o^* = v'', \psi_w) \cdot \mu_{o,v''}} \tag{7.4}$$

based on the equation for $f_{o,w}^v$ used at the E-step in Figure 6.3.

M-step. For the objective function $\mathbb{F}_{v_o^w=v'}$, we obtain the following equation of the M-step for the confidence distribution μ_o with the additional answer $v_o^w = v'$

$$\mu_{o,v|v_o^w=v'} = \frac{\sum_{s \in S_o} f_{o,s}^v + \sum_{w' \in W_o} f_{o,w'}^v + f_{o,w|v_o^w=v'}^v + \gamma_{o,v} - 1}{|S_o| + |W_o| + 1 + \sum_{v'' \in V_o} \left(\gamma_{o,v''} - 1\right)}$$

by adding the related terms $f_{o,w|v_o^w=v'}^v$ and 1 to the numerator and the denominator of the update equation in Eq. (6.9), respectively. Let $N_{o,v}$ and D_o be the numerator and the denominator in Eq. (6.9), respectively. Then, the above equation can be rewritten as

$$\mu_{o,v|v_o^w=v'} = \frac{N_{o,v} + f_{o,w|v_o^w=v'}^v}{D_o + 1}. \tag{7.5}$$

By substituting $f_{o,w|v_o^w=v'}^v$ in Eq. (7.5) with Eq. (7.4), the conditional confidence becomes

$$\mu_{o,v|v_o^w=v'} = \frac{N_{o,v} + \frac{P(v_o^w=v' | v_o^*=v, \psi_w) \cdot \mu_{o,v}}{\sum_{v'' \in V_o} P(v_o^w=v' | v_o^*=v'', \psi_w) \cdot \mu_{o,v''}}}{D_o + 1}. \tag{7.6}$$

Since $N_{o,v}$ and D_o are proportional to the number of the existing claimed values, the confidence will be changed very little if there are many claimed values already. Thus, we can overcome the second drawback of QASCA. Since $N_{o,v}$s and D_os are repeatedly used to compute $\mu_{o,v|v_o^w=v'}$, our truth discovery algorithm keeps $N_{o,v}$s and D_os in main memory to reduce the computation time.

Time complexity analysis. To calculate $E[\max_v \mu_{o,v|w}]$ by Eq. (7.3), $P(v_o^w = v'|\psi_w, \mu_o)$ is computed $|V_o|$ times and $\mu_{o,v|v_o^w=v'}$ is calculated for every pair of v and v' (i.e., $O(|V_o|^2)$ times). Moreover, computing $P(v_o^w = v'|\psi_w, \mu_o)$ and $\mu_{o,v|v_o^w=v'}$ take $O(|V_o|)$ time. Thus, it takes $O(|V_o|^3)$ time to compute $EAI(w, o)$ by Eq. (7.2). In reality, $|V_o|$ is very small compared to $|O|$,$|S|$ and $|W|$. In addition, by utilizing the pruning technique in the next section, we can significantly reduce the computation time. Therefore, the task assignment step can be performed within a short period of time compared to the EM steps of the truth discovery algorithm. The execution time for each step will be presented in the experiment section.

7.2.3 The Task Assignment Algorithm

To find the k objects to be assigned to each worker, we need to compute $EAI(w, o)$ for all pairs of w and o. To reduce the number of computing $EAI(w, o)$, we develop a pruning technique by utilizing an upper bound of $EAI(w, o)$. Since it takes $O(|W||O| \cdot (max_{o \in O}|V_o|)^3)$ time to compute $EAI(w, o)$ for all pairs of w and o, we first derive an upper bound of $EAI(w, o)$ and next propose an efficient task assignment algorithm by exploiting the upper bound to reduce the computation overhead.

An upper bound of EAI. We provide the following lemma which allows us to compute an upper bound $U_{EAI}(o)$.

Lemma 1. (Upper Bound of Expected Accuracy Increase) *For an object o and a worker w, we have*

$$EAI(w, o) \leq U_{EAI}(o) = \frac{1 - \max_v \mu_{o,v}}{|O| \cdot (D_o + 1)}. \tag{7.7}$$

Proof. From Eq. (7.6), since $\sum_{v'} P(v_o^w = v'|\psi_w, \mu_o) = 1$, we get

$$\begin{aligned} E[\max_v \mu_{o,v|w}] &= \sum_{v' \in V_o} P(v_o^w{=}v'|\psi_w, \mu_o) \cdot \max_v \mu_{o,v|v_o^w=v'} \\ &\leq \max_{v,v'} \mu_{o,v|v_o^w=v'} \cdot \sum_{v' \in V_o} P(v_o^w{=}v'|\psi_w, \mu_o) \\ &= \max_{v,v'} \mu_{o,v|v_o^w=v'}. \end{aligned} \tag{7.8}$$

Moreover, from Eq. (7.5), we obtain

$$\mu_{o,v|v_o^w=v'} = \frac{N_{o,v} + f^v_{o,w|v_o^w=v'}}{D_o + 1} \leq \frac{N_{o,v} + 1}{D_o + 1}. \tag{7.9}$$

By substituting Eq. (7.9) for $\mu_{o,v|v_o^w=v'}$ in Eq. (7.8), we derive

$$E[\max_v \mu_{o,v|w}] \leq \max_{v,v'} \mu_{o,v|v_o^w=v'} \leq \frac{\max_v N_{o,v} + 1}{D_o + 1}. \tag{7.10}$$

In addition, by applying Eq. (7.10) to Eq. (7.2), we get

$$EAI(w,o) \leq (\tfrac{\max_v N_{o,v}+1}{D_o+1} - \max_v \mu_{o,v})/|O|.$$

Since $\mu_{o,v} = \frac{N_{o,v}}{D_o}$, we finally obtain the upper bound of $EAI(w,o)$.

$$\begin{aligned} EAI(w,o) &\leq (\tfrac{\max_v N_{o,v}+1}{D_o+1} - \tfrac{\max_v N_{o,v}}{D_o})/|O| \\ &= \tfrac{1-\frac{\max_v N_{o,v}}{D_o}}{|O|\cdot(D_o+1)} = \tfrac{1-\max_v \mu_{o,v}}{|O|\cdot(D_o+1)} = U_{EAI}(o). \end{aligned}$$

□

We devise an algorithm to assign the best k objects to each available worker in crowdsourcing systems. Since a single answer is sufficient to find the correct value for some objects, we assign an object to only a single worker in each round. If the answer is not sufficient to find the correct value of the object, we assign the object to another worker in the next round.

Algorithm 1 Task Assignment

Input: set of workers W, number of questions k

Compute the upper bound $U_{EMCI}(o)$ for $o \in O$
$h_{UB} \leftarrow$ BuildMaxHeap($\{\langle U_{EAI}(o), o\rangle | o \in O\}$)
Sort workers in the decreasing order of $\psi_{w,1}$
(i.e., $\psi_{1,1} \geq \psi_{2,1} \geq \cdots \geq \psi_{|W|,1}$).
for $w = 1$ to $|W|$ **do**
    $h_{EAI}[w] \leftarrow$ BuildMinHeap($\{\}$)
while True **do**
    $\langle U_{EAI}(o), o\rangle \leftarrow h_{UB}$.extractMax()
    if $h_{EAI}[|W|].size = k$ **and** $h_{EAI}[w].min > U_{EAI}(o)$ for all w **then**
        break
    for $w = 1$ to $|W|$ **do**
        if w already answered on o **or** $h_{EAI}[w].min > U_{EAI}(o)$ **then**
            continue
        Compute $EAI(w, o)$
        $h_{EAI}[w]$.insert($\langle EAI(w, o), o\rangle$)
        if $h_{EAI}[w].size \leq k$ **then**
            break
        $o \leftarrow h_{EAI}[w]$.extractMin().value()

Our task assignment algorithm sequentially assigns each object to a worker by scanning the objects o with non-increasing order of the upper bound $U_{EAI}(o)$. To allocate an object to a worker, since $\psi_{w,1}$ is the probability of answering the truth, we consider the workers w with non-increasing order of $\psi_{w,1}$. After assigning an object to a worker w, if the number of assigned objects to the worker w exceeds k, we remove the object o with the minimum $EAI(w, o)$ and assign the deleted object to the next worker and perform the same step. While scanning the objects, we stop the as-

signment if the upperbound $U_{EAI}(o)$ is smaller than the minimum $EAI(w, o')$ among the $EAI(w, o')$s of all assigned objects and each worker has k assigned objects. The reason is that the $EAI(w, o)$ of the remaining objects o can be larger than that of any assigned object.

The pseudocode. It is shown in Algorithm 1. We first compute the upper bound $U_{EAI}(o)$ for every object $o \in O$ by Lemma 1 and build a maxheap h_{UB} of all objects by using $U_{EAI}(o)$ as the key to assign the objects to workers in the decreasing order of $U_{EAI}(o)$ (in lines 1-2). The workers are sorted in the decreasing order of $\psi_{w,1}$ to give a higher priority to reliable workers (in line 3). We next initialize a minheap $h_{EAI}[w]$ for each worker w to contain the k assigned objects (in lines 4-6). Then, we repeatedly extract an object from h_{UB} and assign the object to a worker in the sorted order of $\psi_{w,1}$ (in lines 12-18). Before assigning an object o, if the heaps $h_{EAI}[w]$s of all workers are full and the minimum value of $EAI(w, o')$ of the objects o' in all $h_{EAI}[w]$s is larger than the upper bound $U_{EAI}(o)$, we stop immediately.

7.3 Experiments

7.3.1 Test Environments

Basically, we conducted experiments with the same setting in Chapter 6 except for the followings. In this section, we utilized crowdsourcing to evaluate the task assignment algorithms. Since we cannot change the quality of workers, we first conducted experiments with simulated crowdsourcing. Next, we also verify the results with crowdsourcing with human annotators (workers).

Settings for simulated crowdsourcing. To evaluate the truth discovery algorithms with varying the quality of the answers from workers, we conducted experiments with simulated crowd workers. In our simulation, we assumed that each simulated worker answers a question correctly with its own probability p_w and randomly selects an answer from the candidate values with probability $1-p_w$. We sampled the probability

p_w from a uniform distribution ranging from $\pi_p - 0.05$ to $\pi_p + 0.05$ where the default value of π_p is 0.75. In the experiments, each of 10 worker answers 5 questions for each round.

Crowdsourcing with human annotators. We evaluated truth discovery algorithms and task assignment algorithms by crowdsourcing real human annotations in the interactive setting. In the experiment, 10 human annotators answered the assigned tasks for 20 rounds. However, this result in the interactive setting is less repeatable because other researchers cannot conduct the same experiment with the same workers. Therefore, we collected answers from 20 workers in Amazon Mechanical Turk (AMT) for all objects of *Heritages* dataset and made it publicly available. We also present the result with the answers obtained from AMT.

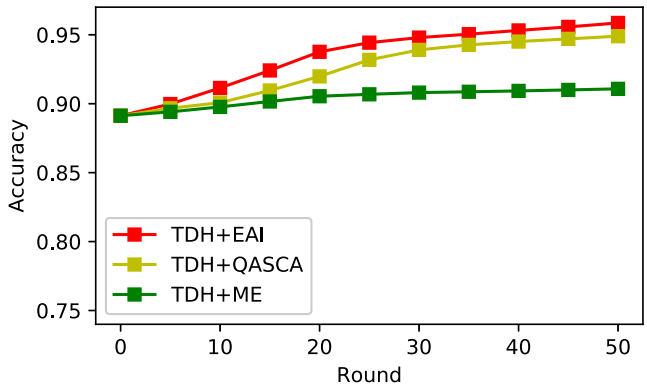
Implemented algorithms. We implemented the following task assignment algorithms.

- *EAI*: This is our proposed algorithm in Section 7.2.
- *MB*: It is the task assignment algorithm used by DOCS [89].
- *QASCA*: It is a task assignment algorithm proposed in [91].
- *ME*: This is our baseline algorithm which utilizes an uncertainty sampling. It selects an object o^* whose confidence distribution has the maximum entropy. (i.e., $o^* = argmax_{o \in O} \left(-\sum_{v \in V_o} \mu_{o,v} \cdot \log \mu_{o,v}\right)$)

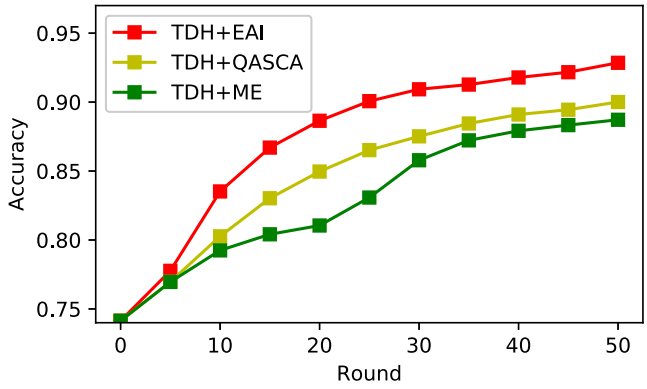
Note that *EAI* and *MB* are the task assignment algorithms specially designed to work with *TDH* and *DOCS*, respectively. *QASCA* can work with truth discovery algorithms based on probabilistic models such as *TDH*, *DOCS*, *LCA*, *ACCU* and *POPACCU*. All the truth discovery algorithms can be combined with *ME*.

7.3.2 Comparison of Task Assignment Algorithms

Before providing the full comparison of all possible combinations of truth inference algorithms and task assignment algorithms, we first evaluate the task assignment algorithms with our truth discovery algorithm TDH proposed in Chapter 6. We plotted



(a) *BirthPlaces*

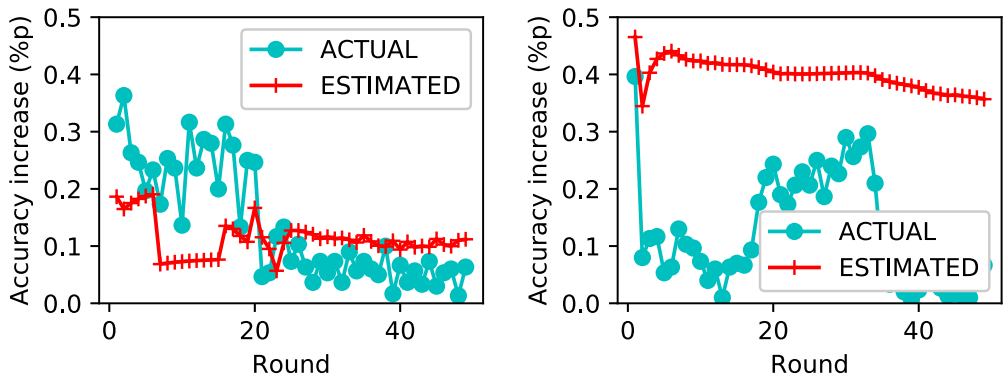


(b) *Heritages*

Figure 7.2: Evaluation of task assignment algorithms

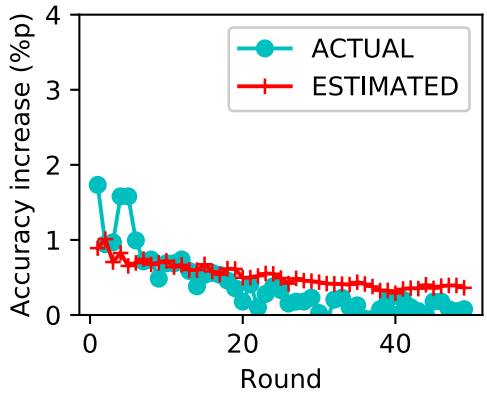
the average *Accuracy* of the truth discovery algorithms with different task assignment algorithms for every 5 round in Figure 7.2. The points at the 0-th round represent the *Accuracy* of the algorithms without crowdsourcing. All algorithms show the same *Accuracy* at the beginning since they use the same truth inference algorithm TDH. As the round progresses, the *Accuracy* of TDH+EAI increases faster than those of all other algorithms. The *Accuracy* of TDH+ME is the lowest since ME selects a task based only on the uncertainty without estimating the accuracy improvement by the task.

As discussed in Section 7.2.1, our task assignment algorithm EAI estimates the

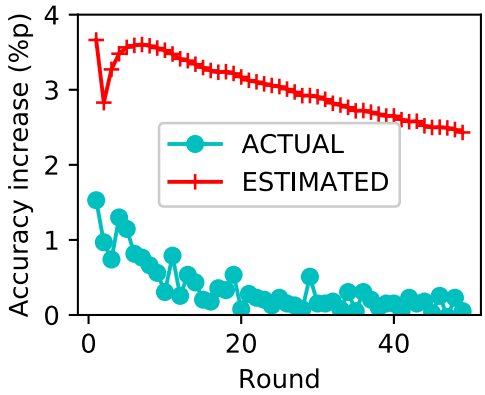


(a) *BirthPlaces*-EAI

(b) *BirthPlaces*-QASCA



(c) *Heritages*-EAI



(d) *Heritages*-QASCA

Figure 7.3: Actual and estimated accuracy improvement by EAI and QASCA

accuracy improvement by considering the number of existing claimed values and the confidence distribution whereas QASCA considers the confidence distribution only. We plotted the actual and estimated accuracy improvements by EAI and QASCA in Figure 7.3. The graphs show that the estimated accuracy improvement by EAI is similar to the actual accuracy improvement while QASCA overestimates the accuracy improvement at every round. On average, the absolute estimation errors from EAI are 0.08 and 0.26 percentage points (pps) while those errors from QASCA are 0.28

and 2.66 pps in *BirthPlaces* and *Heritages* datasets, respectively. This result confirms that EAI outperforms QASCA by effectively estimating the accuracy improvement. In terms of the other quality measures *GenAccuracy* and *AvgDistance*, our proposed EAI also outperforms the other task assignment algorithms in both datasets. Due to the lack of space, we omit the results with the other quality measures.

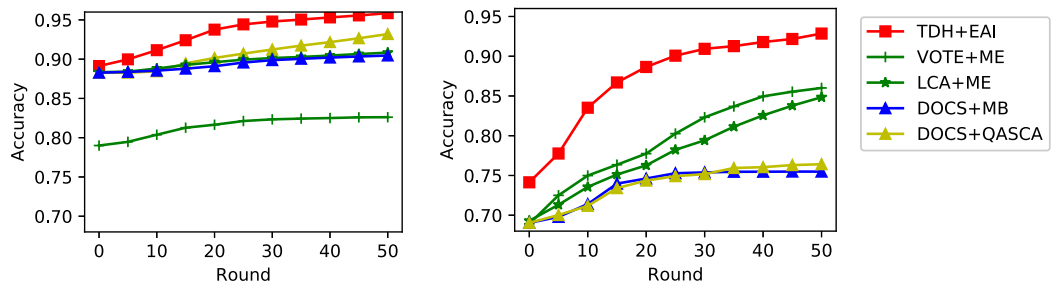
7.3.3 Simulated Crowdsourcing

We first evaluate the performance of crowdsourced truth discovery algorithms with the simulated crowdsourcing.

For all possible combinations of the implemented truth discovery and task assignment algorithms, we show the *Accuracy* after 50 rounds of crowdsourcing in Table 7.1 where the impossible combinations are denoted by '-'. As expected, TDH+EAI has the highest *Accuracy* in both datasets for all possible combinations. The result also shows that both TDH and EAI contribute to increasing *Accuracy*. The improvement obtained

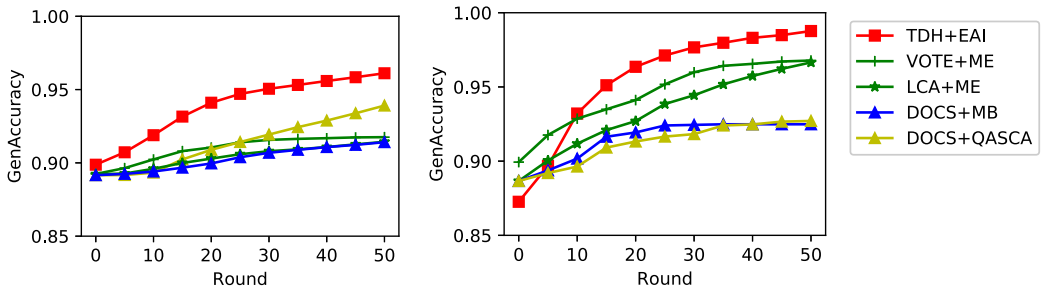
	BirthPlaces				*Heritages*			
	EAI	MB	QASCA	ME	EAI	MB	QASCA	ME
TDH	**0.9601**	-	**0.9500**	**0.9109**	**0.9304**	-	**0.8999**	**0.8884**
DOCS	-	**0.9052**	0.9341	0.8842	-	**0.7546**	0.7661	0.7631
LCA	-	-	0.8823	0.9089	-	-	0.7136	0.8507
POPACCU	-	-	0.9295	0.8987	-	-	0.7512	0.8336
ACCU	-	-	0.8468	0.8257	-	-	0.5796	0.5896
ASUMS	-	-	-	0.8700	-	-	-	0.7427
CRH	-	-	-	0.9000	-	-	-	0.8459
MDC	-	-	-	0.8254	-	-	-	0.7241
LFC	-	-	-	0.8287	-	-	-	0.7327
VOTE	-	-	-	0.8261	-	-	-	0.8634

Table 7.1: *Accuracy* of the algorithms after the 50th round



(a) *BirthPlaces* (b) *Heritages*

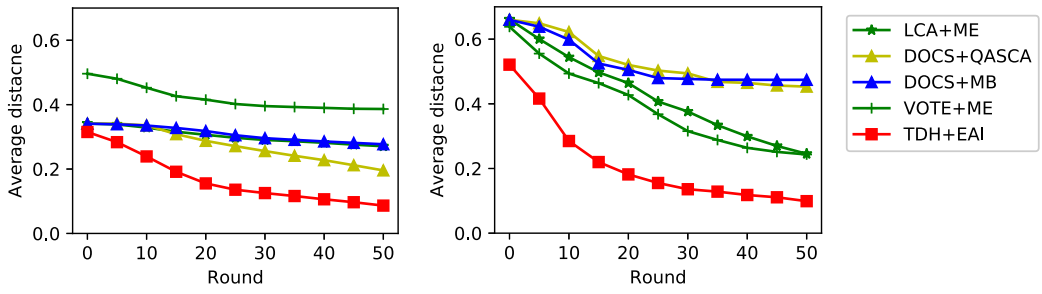
Figure 7.4: *Accuracy* with crowdsourced truth discovery



(a) *BirthPlaces* (b) *Heritages*

Figure 7.5: *GenAccuracy* with crowdsourced truth discovery

by EAI can be estimated by comparing the result of TDH+EAI to that of the second performer TDH+QASCA. The accuracies of TDH+EAI in *BirthPlaces* and *Heritages* datasets are 1 and 3 percentage points (pps) higher than those of TDH+QASCA, respectively. In addition, for each combined task assignment algorithm, the improvement by TDH can be inferred by comparing the results with those of other truth inference algorithms. In both datasets, TDH shows the highest *Accuracy* among the applicable truth inference algorithms for each task assignment algorithm. For example, TDH+QASCA shows 2.6 and 13 pps higher *Accuracy* in *BirthPlaces* and *Heritages* datasets, respectively, than the second performer DOCS+QASCA among the combi-



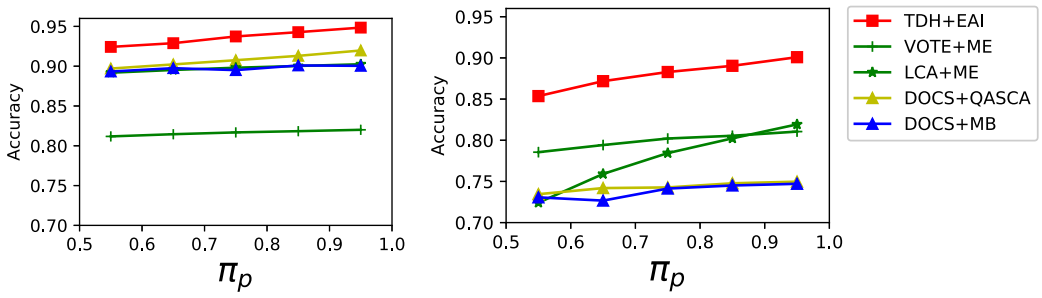
(a) *BirthPlaces* (b) *Heritages*

Figure 7.6: *AvgDistance* with crowdsourced truth discovery

nations with QASCA. In the rest of the paper, we report *Accuracy*, *GenAccuracy* and *AvgDistance* of TDH+EAI, DOCS+MB, DOCS+QASCA, LCA+ME and VOTE+ME only since these combinations are the best or the second-best for each task assignment algorithm.

Cost efficiency. We plotted the average *Accuracy* of the tested algorithms for every 5 rounds in Figure 7.4. TDH+EAI shows the highest *Accuracy* for every round in both datasets. For the *BirthPlaces* dataset, DOCS+QASCA was the next best performer which achieved 0.9341 of *Accuracy* at the 50-th round. Meanwhile, TDH+EAI only needs 17 rounds of crowdsourcing to achieve the same *Accuracy*. Thus, TDH+EAI saved 66% of crowdsourcing cost compared to the second-best performer DOCS+QASCA. Likewise, TDH+EAI reduced the crowdsourcing cost 74% in *Heritages* dataset compared to the next performer. In terms of *GenAccuracy* and *AvgDistance*, TDH+EAI also outperforms all the other algorithms as plotted in Figure 7.5 and Figure 7.6. The results confirm that TDH+EAI is the most efficient as it achieves the best qualities in terms of both *Accuracy* and *GenAccuracy*.

Varying π_p. We plotted the average *Accuracy* of all algorithms with varying the probability of correct answer π_p of simulated workers for *BirthPlaces* and *Heritages* datasets in Figure 7.7(a) and Figure 7.7(b), respectively. As we can easily expect, the accura-



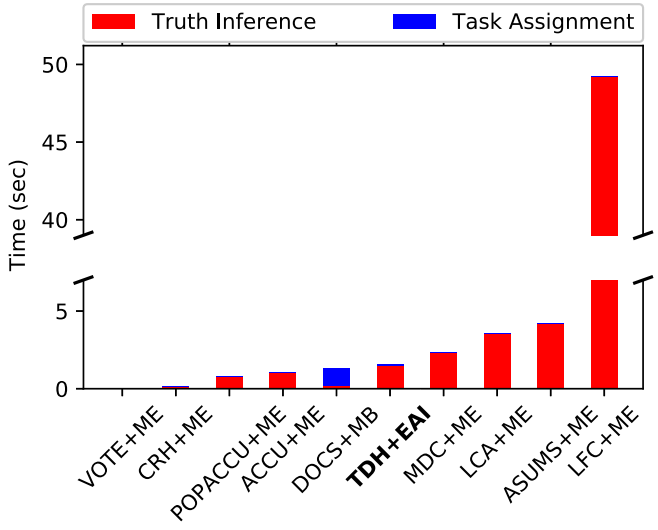
(a) *BirthPlaces* (b) *Heritages*

Figure 7.7: Varying π_p

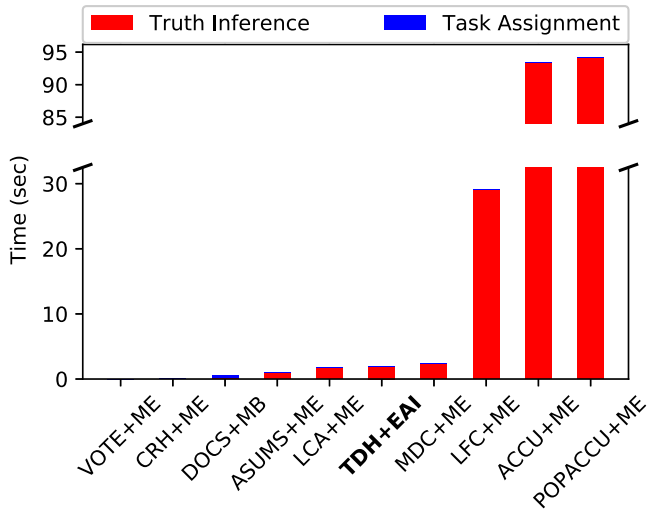
cies increase with growing π_p for most of the algorithms. For both datasets, TDH+EAI achieves the best accuracy with all values of π_p. In Heritages dataset, a source provided less than 10 claims on average and it makes difficult for truth discovery algorithms to estimate the reliabilities of sources. Therefore, the baseline VOTE+ME shows good performance on Heritages dataset. Meanwhile, the performance of the state-of-the-art DOCS is significantly degraded on the Heritages dataset.

Execution times. We plotted the average execution times of the tested algorithms over every round in Figure 7.8. VOTE, CRH+ME, DOCS+MB and TDH+EAI run in less than 2.0 seconds per round on average for both datasets. Other algorithms except for ACCU+ME, POPACCU+ME and LFC+ME also take less than 5 seconds, which is acceptable for crowdsourcing. Since LFC builds the confusion matrix whose size is the square of the number of candidate values, LFC is the slowest with *BirthPlaces* data. On the other hand, for *Heritages* dataset which is collected from much more sources than *BirthPlaces* dataset, ACCU and POPACCU take longer time for truth inference to calculate the dependencies between sources.

Effects of the filtering for task assignments. To test the scalability of our algorithm, we increase the size of both datasets by duplicating the data by upto 15 times. In Figure 7.9, with increasing data size, we plotted the execution times of our task



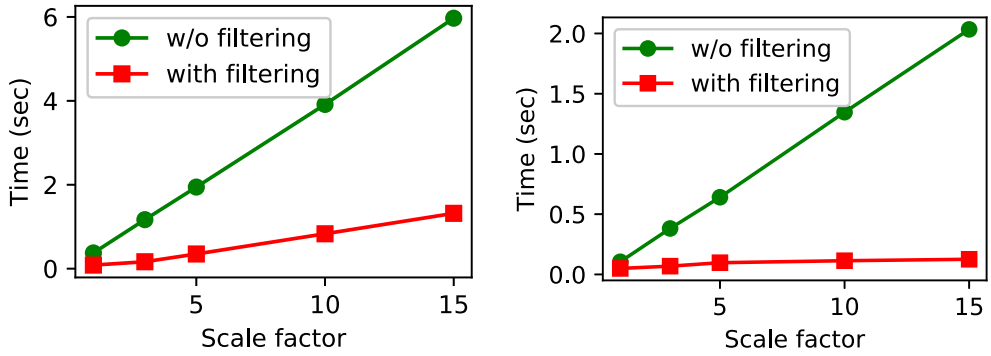
(a) *BirthPlaces*



(b) *Heritages*

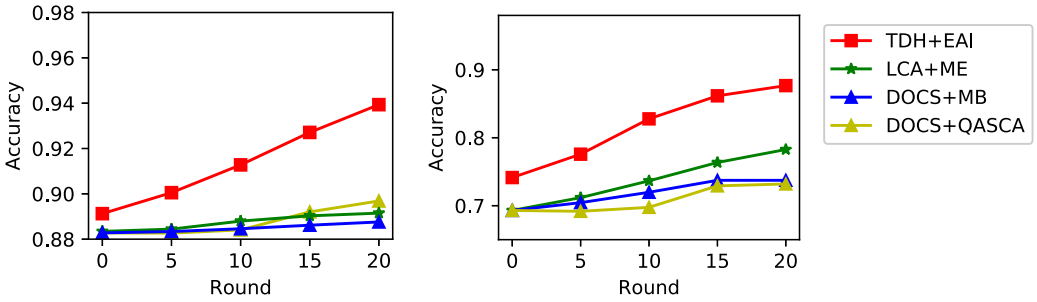
Figure 7.8: Execution time per round

assignment algorithm EAI with and without exploiting the upper bound proposed in Section 7.2.3. The filtering technique saved 78% and 94% of the computation time for the task assignment at the scale factor 15. The graphs show that the proposed upper



(a) *BirthPlaces* (b) *Heritages*

Figure 7.9: Execution time for task assignment per round



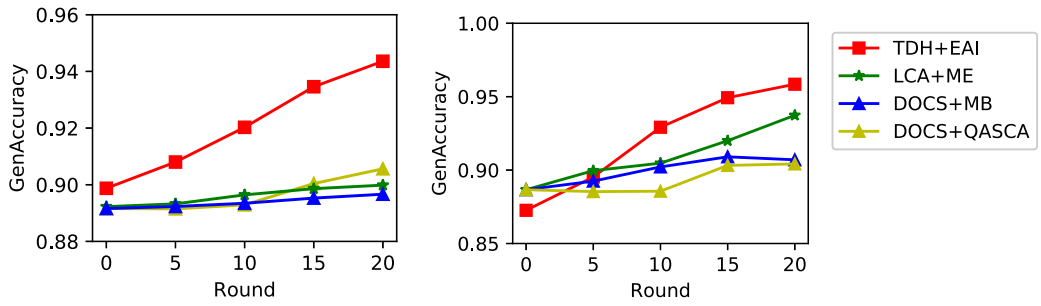
(a) *BirthPlaces* (b) *Heritages*

Figure 7.10: *Accuracy* with human annotations

bound enables us to scale for large data effectively. For the total execution time, including the truth inference, the filtering reduced 21% and 6% of the execution time on *BirthPlaces* and *Heritages* respectively at the scale factor 15.

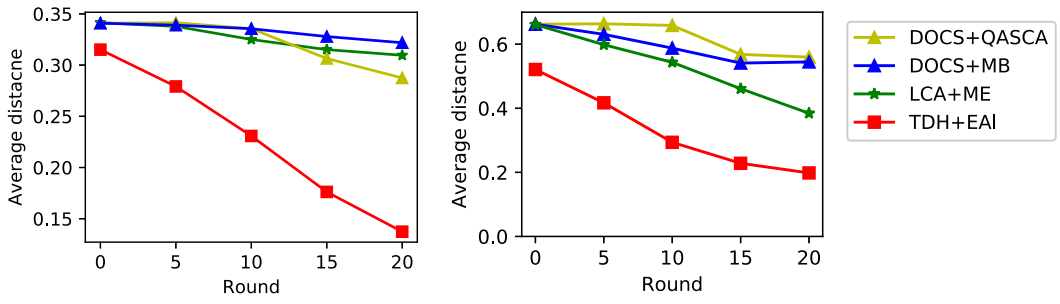
7.3.4 Crowdsourcing with Human Annotators

We evaluated the performance of the truth discovery algorithm by crowdsourcing real human annotations in the interactive setting. For this experiment, we selected DOCS+QASCA, DOCS+MB and LCA+ME for comparison with the proposed algo-



(a) *BirthPlaces* (b) *Heritages*

Figure 7.11: *GenAccuracy* with human annotations



(a) *BirthPlaces* (b) *Heritages*

Figure 7.12: *AvgDistance* with human annotations

rithm TDH+EAI. This is because they are the best existing algorithms for each task assignment algorithm. We conducted this experiment with 10 human annotators for 20 rounds on our own crowdsourcing system. For each worker, we assigned 5 tasks in each round. Figure 7.10,7.11 and 7.12 show the performances of the algorithms against the rounds. For both of the datasets, the results confirm that the proposed TDH+EAI algorithm outperforms the compared algorithms as in the previous simulations. Without crowdsourcing, the other algorithms show a higher *GenAccuracy* than TDH for *Heritages* dataset, because these algorithms tend to estimate the truths with more generalized form than TDH does. However, TDH+EAI shows the highest *GenAccuracy*

after the 3rd round because it correctly estimates the reliabilities and the generalization levels of the sources by using the hierarchy. For *BirthPlaces* dataset, *Accuracies* of the algorithms increase a little bit faster than those in the experiment with simulated crowdsourcing. However, for *Heritages* dataset, *Accuracies* of the algorithms increase much slower than in the experiment with simulated crowdsourcing. It seems that finding the locations of a world heritages is a quite harder task than finding the birthplaces of celebrities because the birthplaces are often big cities (such as LA), which are familiar to workers, but World Cultural Heritages and World Natural Heritages are often located in unfamiliar regions.

7.3.5 Crowdsourcing on AMT

To validate the results in previous experiments, we also evaluate the performances of TDH+EAI, DOCS+QASCA, DOCS+MB and LCA+ME based on the answers collected from Amazon Mechanical Turk (AMT). We collected answers for all objects in *Heritages* dataset from 20 workers in AMT. In addition, we made the collected answers available at `http://kdd.snu.ac.kr/home/datasets/tdh.php` to improve the reproducibility. To evaluate the algorithms based on the collected answers, we assign 5 tasks for each worker in a round. We plotted the performance of the algorithms in Figure 7.13. Since we use more workers than we did in Section 7.3.4, the performances improve a little bit faster, but the trends are very similar to those with 10 human annotators in the previous section. We observe that our TDH+EAI outperforms all compared algorithms even with a commercial crowdsourcing platform.

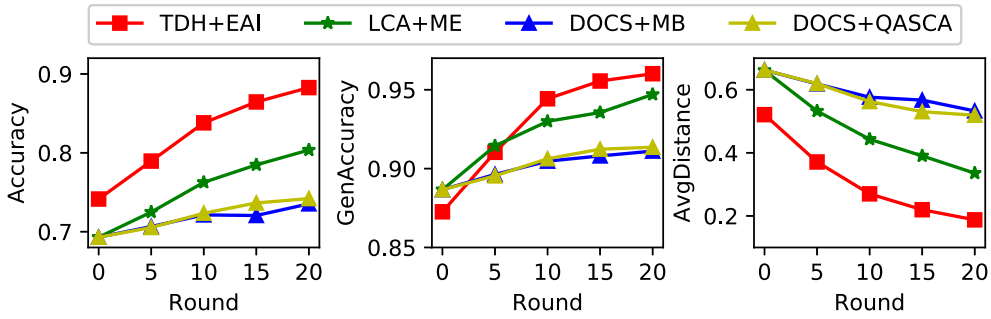


Figure 7.13: Crowdsourced truth discovery in *Heritages*

Chapter 8

Conclusion

Automated knowledge base construction has been studied extensively due to its importance in many downstream applications such as question answering and recommender systems. In this dissertation, we proposed four important techniques to improve the accuracy and coverage of the automated knowledge base construction.

First, we introduced a new problem named topic-aware relation extraction to extend the coverage of relation extraction. We proposed the *T-REX* which utilizes topic entities to extract relations from text. We empirically showed that *T-REX* outperforms existing models in extracting relations with topic entities. In addition, the experiment confirmed that *T-REX* extract many triples which are not detected by the existing models.

Second, we proposed the dual supervision framework to utilize human annotation and distant supervision based on the analysis of labeling bias in distant supervision. We devised a new structure for the output layer of RE models that consists of 4 sub networks. The new structure is robust to the noisy labeling of distant supervision since the labels obtained by human annotation and distant supervision are predicted by separate prediction networks *HA-Net* and *DS-Net*, respectively. In addition, we introduced an additional loss term called *disagreement penalty* which enables *HA-Net* to learn from distantly supervised labels. The parameter networks μ-Net and σ-Net adaptively

assess the labeling bias by considering contextual information. Moreover, we theoretically analyzed the effect of the disagreement penalty. Our experiments showed that the dual supervision framework significantly improves the performance of existing relation extraction models.

Third, we studied the problem of truth discovery in the presence of hierarchies. To utilize the hierarchical structures in claimed values, we proposed a probabilistic model and an inference algorithm for the model. To the best of our knowledge, this is the first truth discovery work which assess both reliabilities and generalization tendencies of sources. The performance study with real-life datasets confirmed the effectiveness of the proposed hierarchical truth discovery algorithm.

Finally, we examined the problem of task assignment to workers in crowdsourcing platforms. To assign a task that will most improve the accuracy, we develop an incremental EM algorithm to estimate the accuracy improvement for a task. We also proposed an efficient pruning technique to assign tasks to workers with a short latency. We conducted extensive experiments with simulated crowdsourcing, interactive crowdsourcing with human annotators, and crowdsourcing on a commercial platform AMT (amazon mechanical turk). The experimental results showed the effectiveness and efficiency of the task assignment algorithm.

We next discuss potential directions for future work on automated knowledge base construction and its downstream applications.

Rule mining in automatically constructed knowledge bases. Rule mining in knowledge bases has been extensively studied in many works such as [66, 55, 59]. Automatically constructed knowledge bases inevitably have much more errors compared to manually constructed knowledge bases. Thus, the results of rule mining in automatically constructed knowledge bases can be erroneous. Since truth discovery algorithms estimate the confidence on each relational fact and reliability of each source, incorporating such information to rule mining would be an interesting research direction.

Distant supervision with noisy knowledge bases. Recall that distant supervision

generates a lot of incorrect labels even when the knowledge base does not have an error. The wrong labeling problem in distant supervision has been addressed in many previous works [82, 47, 79, 3]. In addition, we also propose a method to alleviate the effect of the noisy labels obtained from distant supervision in chapter 5. Compared to distant supervision with manually constructed knowledge bases, distant supervision with automatically constructed knowledge bases may generate much more false labels since there can be many false relational facts in such knowledge bases. Thus, it would be a more challenging problem to train relation extraction models on distantly supervised data with noisy knowledge bases.

We believe that the techniques proposed in the dissertation enhance the downstream applications of knowledge bases such as question answering and recommender systems.

Bibliography

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[3] I. Beltagy, K. Lo, and W. Ammar. Combining distant and direct supervision for neural relation extraction. *arXiv preprint arXiv:1810.12956*, 2018.

[4] V. Beretta, S. Harispe, S. Ranwez, and I. Mougenot. How can ontologies give you clue for truth-discovery? an exploratory study. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*, page 15. ACM, 2016.

[5] R. Boim, O. Greenshpan, T. Milo, S. Novgorodov, N. Polyzotis, and W.-C. Tan. Asking the right questions in crowd data sourcing. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 1261–1264. IEEE, 2012.

[6] R. Cai, X. Zhang, and H. Wang. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 756–765, 2016.

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011.

[8] V. Crescenzi, G. Mecca, P. Merialdo, et al. Roadrunner: Towards automatic data extraction from large web sites. In *VLDB 2001, Proceedings of 27th International Conference on Very Large Data Bases, September 11-14, 2001, Roma, Italy*, pages 109–118. Morgan Kaufmann, 2001.

[9] R. Das, A. Neelakantan, D. Belanger, and A. McCallum. Chains of reasoning over entities, relations, and text using recurrent neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 132–141, 2017.

[10] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

[11] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*, pages 469–478. ACM, 2012.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610, 2014.

[14] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment*, 2(1):550–561, 2009.

[15] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang. From data fusion to knowledge fusion. *Proceedings of the VLDB Endowment*, 7(10):881–892, 2014.

[16] X. L. Dong, B. Saha, and D. Srivastava. Less is more: Selecting sources wisely for integration. In *Proceedings of the VLDB Endowment*, volume 6, pages 37–48. VLDB Endowment, 2012.

[17] X. L. Dong and D. Srivastava. Knowledge curation and knowledge fusion: challenges, models and applications. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 2063–2066. ACM, 2015.

[18] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia. Incorporating second-order functional knowledge for better option pricing. In *Advances in neural information processing systems*, pages 472–478, 2001.

[19] J. Ellis. Linguistic resources for 2013 knowledge base population evaluations. In *Text Analysis Conference (TAC)*, 2012.

[20] J. Fan, G. Li, B. C. Ooi, K.-l. Tan, and J. Feng. icrowd: An adaptive crowdsourcing framework. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1015–1030. ACM, 2015.

[21] J. Fan, M. Lu, B. C. Ooi, W.-C. Tan, and M. Zhang. A hybrid machine-crowdsourcing system for matching web tables. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 976–987. IEEE, 2014.

[22] Y. Gao, Y.-F. Li, Y. Lin, H. Gao, and L. Khan. Deep learning on knowledge graph for recommender system: A survey. *arXiv preprint arXiv:2004.00387*, 2020.

[23] D. Haas, J. Wang, E. Wu, and M. J. Franklin. Clamshell: Speeding up crowds for low-latency data labeling. *Proceedings of the VLDB Endowment*, 9(4):372–383, 2015.

[24] X. Han, L. Sun, and J. Zhao. Collective entity linking in web text: a graph-based method. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 765–774, 2011.

[25] C.-J. Ho, S. Jabbari, and J. W. Vaughan. Adaptive task assignment for crowdsourced classification. In *International Conference on Machine Learning*, pages 534–542, 2013.

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

[27] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. 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.

[28] W. Jung, Y. Kim, and K. Shim. Crowdsourced truth discovery in the presence of hierarchies for knowledge fusion. In *Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019*, pages 205–216, 2019.

[29] W. Jung and K. Shim. Dual supervision framework for relation extraction with distant supervision and human annotation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6411–6423, 2020.

[30] W. Jung and K. Shim. T-rex: A topic-aware relation extraction model. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2073–2076, 2020.

[31] D. R. Karger, S. Oh, and D. Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pages 1953–1961, 2011.

[32] H.-C. Kim and Z. Ghahramani. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, pages 619–627, 2012.

[33] Y. Kim, W. Jung, and K. Shim. Integration of graphs from different data sources using crowdsourcing. *Information Sciences*, 385:438–456, 2017.

[34] Y. Kim, W. Kim, and K. Shim. Latent ranking analysis using pairwise comparisons in crowdsourcing platforms. *Inf. Syst.*, 65:7–21, 2017.

[35] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[36] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[37] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

[38] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[39] G. Li, J. Wang, Y. Zheng, and M. J. Franklin. Crowdsourced data management: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 28(9):2296–2319, 2016.

[40] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han. A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment*, 8(4):425–436, 2014.

[41] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1187–1198. ACM, 2014.

[42] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: Is the problem solved? In *Proceedings of the VLDB Endowment*, volume 6, pages 97–108. VLDB Endowment, 2012.

[43] Y. Li, N. Du, C. Liu, Y. Xie, W. Fan, Q. Li, J. Gao, and H. Sun. Reliable medical diagnosis from crowdsourcing: Discover trustworthy answers from non-experts. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 253–261. ACM, 2017.

[44] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. A survey on truth discovery. *SIGKDD Explorations Newsletter*, 17(2):1–16, Feb. 2016.

[45] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han. On the discovery of evolving truth. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 675–684. ACM, 2015.

[46] X. Lin and L. Chen. Domain-aware multi-truth discovery from conflicting sources. *Proceedings of the VLDB Endowment*, 11(5):635–647, 2018.

[47] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. 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, 2016.

[48] X. Ling and D. S. Weld. Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 94–100, 2012.

[49] L. Liu, X. Ren, Q. Zhu, S. Zhi, H. Gui, H. Ji, and J. Han. Heterogeneous supervision for relation extraction: A representation learning approach. *arXiv preprint arXiv:1707.00166*, 2017.

[50] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang. K-bert: Enabling language representation with knowledge graph.

[51] C. Lockard, X. L. Dong, A. Einolghozati, and P. Shiralkar. Ceres: distantly supervised relation extraction from the semi-structured web. *Proceedings of the VLDB Endowment*, 11(10):1084–1096, 2018.

[52] P. H. Martins, Z. Marinho, and A. F. Martins. Joint learning of named entity recognition and entity linking. *arXiv preprint arXiv:1907.08243*, 2019.

[53] F. J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[54] P. Mavridis, D. Gross-Amblard, and Z. Miklós. Using hierarchical skills for optimized task assignment in knowledge-intensive crowdsourcing. In *Proceedings of the 25th International Conference on World Wide Web*, pages 843–853, 2016.

[55] A. Melo and H. Paulheim. Detection of relation assertion errors in knowledge graphs. In *Proceedings of the Knowledge Capture Conference*, pages 1–8, 2017.

[56] M. Mintz, S. Bills, R. Snow, and D. 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: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[57] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 877–885. Association for Computational Linguistics, 2010.

[58] J. Pasternack and D. Roth. Latent credibility analysis. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1009–1020. ACM, 2013.

[59] H. Paulheim. Browsing linked open data with auto complete. 2012.

[60] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[61] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322, 2010.

[62] T. Rebele, F. Suchanek, J. Hoffart, J. Biega, E. Kuzey, and G. Weikum. Yago: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *International semantic web conference*, pages 177–185. Springer, 2016.

[63] X. Ren, Z. Wu, W. He, M. Qu, C. R. Voss, H. Ji, T. F. Abdelzaher, and J. Han. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1015–1024, 2017.

[64] S. Riedel, L. Yao, and A. McCallum. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010.

[65] M. Rodríguez, S. Goldberg, and D. Z. Wang. Sigmakb: multiple probabilistic knowledge base fusion. *Proceedings of the VLDB Endowment*, 9(13):1577–1580, 2016.

[66] A. Sadeghian, M. Armandpour, P. Ding, and D. Z. Wang. Drum: End-to-end differentiable rule mining on knowledge graphs. In *Advances in Neural Information Processing Systems*, pages 15347–15357, 2019.

[67] D. Sorokin and I. Gurevych. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1784–1789, 2017.

[68] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[69] H. Sun, B. Dhingra, M. Zaheer, K. Mazaitis, R. Salakhutdinov, and W. W. Cohen. Open domain question answering using early fusion of knowledge bases and text. *arXiv preprint arXiv:1809.00782*, 2018.

[70] P. University. About wordnet., 2010.

[71] P. Verga, E. Strubell, and A. McCallum. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884, 2018.

[72] M. Wan, X. Chen, L. Kaplan, J. Han, J. Gao, and B. Zhao. From truth discovery to trustworthy opinion discovery: An uncertainty-aware quantitative modeling approach. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1885–1894. ACM, 2016.

[73] H. Wang, C. Focke, R. Sylvester, N. Mishra, and W. Wang. Fine-tune bert for docred with two-step process. *arXiv preprint arXiv:1909.11898*, 2019.

[74] J. Wang, T. Kraska, M. J. Franklin, and J. Feng. Crowder: Crowdsourcing entity resolution. *Proceedings of the VLDB Endowment*, 5(11):1483–1494, 2012.

[75] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043, 2009.

[76] Y. Xia, J. Ma, Z. Zheng, and R. Liu. Web data extraction with seed samples, 2019.

[77] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, and M. Sun. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, 2019.

[78] Q. Ye, L. Liu, M. Zhang, and X. Ren. Looking beyond label noise: Shifted label distribution matters in distantly supervised relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3832–3841, 2019.

[79] Z.-X. Ye and Z.-H. Ling. Distant supervision relation extraction with intra-bag and inter-bag attentions. *arXiv preprint arXiv:1904.00143*, 2019.

[80] S. W.-t. Yih, M.-W. Chang, X. He, and J. Gao. Semantic parsing via staged query graph generation: Question answering with knowledge base. 2015.

[81] X. Yin, J. Han, and S. Y. Philip. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808, 2008.

[82] D. Zeng, K. Liu, Y. Chen, and J. 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, 2015.

[83] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. 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.

[84] F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W.-Y. Ma. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 353–362, 2016.

[85] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, 2017.

[86] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*, 2019.

[87] B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. In *Proceeding of the VLDB workshop on Quality in Databases (QDB'12)*, 2012.

[88] B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment*, 5(6):550–561, 2012.

[89] Y. Zheng, G. Li, and R. Cheng. Docs: a domain-aware crowdsourcing system using knowledge bases. *Proceedings of the VLDB Endowment*, 10(4):361–372, 2016.

[90] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng. Truth inference in crowdsourcing: is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552, 2017.

[91] Y. Zheng, J. Wang, G. Li, R. Cheng, and J. Feng. Qasca: A quality-aware task assignment system for crowdsourcing applications. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1031–1046. ACM, 2015.

[92] D. Zhou, S. Basu, Y. Mao, and J. C. Platt. Learning from the wisdom of crowds by minimax entropy. In *Advances in neural information processing systems*, pages 2195–2203, 2012.

Chapter A

Appendix

A.1 Inflation in DocRED dataset

The inflation of 96 relation types of DocRED is shown in Table A.1. As discussed in Section 5.3.3, we conducted Kolmogorov-Smirnov (K-S) test and observed that the log-normal distribution is the best-fit distribution of the inflations.

Relation type	# HA labels	# DS labels	$\frac{\text{\# HA labels}}{\text{\# HA docs.}}$	$\frac{\text{\# DS labels}}{\text{\# DS docs.}}$	Inflation $= \frac{\text{\# DS labels}/\text{\# DS docs.}}{\text{\# HA labels}/\text{\# HA docs.}}$
P17	8,921	313,961	2.9220	3.0819	1.055
P131	4,193	143,006	1.3734	1.4038	1.022
P27	2,689	126,360	0.8808	1.2404	1.408
P150	2,004	62,646	0.6564	0.6149	0.937
P577	1,142	37,538	0.3741	0.3685	0.985
P175	1,052	27,945	0.3446	0.2743	0.796
P569	1,044	33,998	0.3420	0.3337	0.976
P570	805	28,314	0.2637	0.2779	1.054
P161	621	21,139	0.2034	0.2075	1.020
P264	583	14,804	0.1910	0.1453	0.761
P527	632	22,596	0.2070	0.2218	1.071
P361	596	28,245	0.1952	0.2773	1.420

P495	539	36,029	0.1765	0.3537	2.003
P19	511	31,232	0.1674	0.3066	1.832
P571	475	26,699	0.1556	0.2621	1.684
P54	379	12,312	0.1241	0.1209	0.974
P102	406	11,582	0.1330	0.1137	0.855
P463	414	15,272	0.1356	0.1499	1.106
P3373	335	11,123	0.1097	0.1092	0.995
P40	360	11,831	0.1179	0.1161	0.985
P30	356	18,792	0.1166	0.1845	1.582
P50	320	8,856	0.1048	0.0869	0.829
P1441	299	6,763	0.0979	0.0664	0.678
P1001	298	9,945	0.0976	0.0976	1.000
P69	316	8,413	0.1035	0.0826	0.798
P26	303	9,723	0.0992	0.0954	0.962
P607	275	8,056	0.0901	0.0791	0.878
P57	246	9,865	0.0806	0.0968	1.202
P159	264	17,089	0.0865	0.1677	1.940
P22	273	9,065	0.0894	0.0890	0.995
P400	304	5,825	0.0996	0.0572	0.574
P1344	223	3,574	0.0730	0.0351	0.480
P206	194	6,585	0.0635	0.0646	1.017
P127	208	7,554	0.0681	0.0742	1.088
P170	231	6,036	0.0757	0.0593	0.783
P178	238	6,368	0.0780	0.0625	0.802
P20	203	24,937	0.0665	0.2448	3.681
P1412	155	6,313	0.0508	0.0620	1.221
P155	188	12,236	0.0616	0.1201	1.950
P118	185	6,024	0.0606	0.0591	0.976
P710	191	4,985	0.0626	0.0489	0.782
P6	210	6,859	0.0688	0.0673	0.979
P108	196	6,775	0.0642	0.0665	1.036
P276	172	6,654	0.0563	0.0653	1.159

P156	192	11,576	0.0629	0.1136	1.807
P674	163	3,447	0.0534	0.0338	0.634
P166	173	6,322	0.0567	0.0621	1.095
P194	166	2,989	0.0544	0.0293	0.540
P123	172	4,444	0.0563	0.0436	0.774
P140	144	5,143	0.0472	0.0505	1.070
P800	150	5,275	0.0491	0.0518	1.054
P449	152	4,237	0.0498	0.0416	0.835
P58	156	7,952	0.0511	0.0781	1.528
P35	140	4,257	0.0459	0.0418	0.911
P179	144	3,800	0.0472	0.0373	0.791
P706	137	5,063	0.0449	0.0497	1.108
P162	119	6,739	0.0390	0.0662	1.697
P37	119	6,562	0.0390	0.0644	1.653
P241	108	2,633	0.0354	0.0258	0.731
P31	103	5,561	0.0337	0.0546	1.618
P403	95	2,475	0.0311	0.0243	0.781
P580	110	6,549	0.0360	0.0643	1.784
P137	95	3,011	0.0311	0.0296	0.950
P585	96	2,920	0.0314	0.0287	0.912
P112	100	7,700	0.0328	0.0756	2.308
P86	79	4,249	0.0259	0.0417	1.612
P176	83	2,737	0.0272	0.0269	0.988
P749	92	3,335	0.0301	0.0327	1.086
P937	104	7,470	0.0341	0.0733	2.153
P36	85	34,047	0.0278	0.3342	12.004
P576	79	7,057	0.0259	0.0693	2.677
P355	92	2,436	0.0301	0.0239	0.794
P136	111	1,948	0.0364	0.0191	0.526
P364	66	2,274	0.0216	0.0223	1.033
P272	82	2,151	0.0269	0.0211	0.786
P172	79	7,563	0.0259	0.0742	2.869

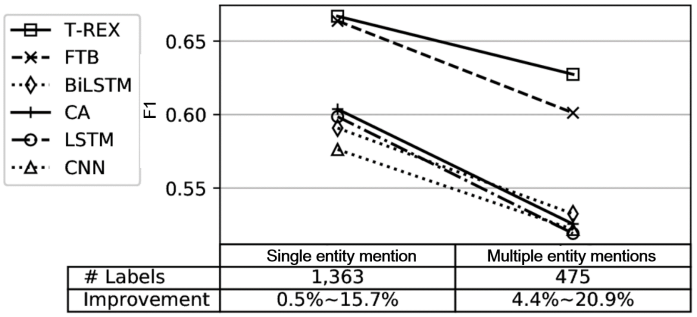
P205	85	3,299	0.0278	0.0324	1.163
P279	77	2,736	0.0252	0.0269	1.065
P1376	76	29,816	0.0249	0.2927	11.757
P171	75	2,167	0.0246	0.0213	0.866
P25	74	2,826	0.0242	0.0277	1.144
P488	63	2,216	0.0206	0.0218	1.054
P582	51	6,144	0.0167	0.0603	3.610
P740	62	4,531	0.0203	0.0445	2.190
P840	48	2,573	0.0157	0.0253	1.606
P1366	36	2,771	0.0118	0.0272	2.307
P676	36	2,415	0.0118	0.0237	2.010
P1336	33	1,600	0.0108	0.0157	1.453
P1056	36	624	0.0118	0.0061	0.519
P551	35	3,197	0.0115	0.0314	2.737
P39	23	1,692	0.0075	0.0166	2.204
P1365	18	1,811	0.0059	0.0178	3.015
P737	9	2,071	0.0029	0.0203	6.895
P190	4	11,471	0.0013	0.1126	85.900
P807	2	2,210	0.0007	0.0217	33.082
P1198	2	1,622	0.0007	0.0159	24.280

Table A.1: Inflations of relation types in DocRED dataset

A.2 An Additional Experiment with *T-REX*: Effect of the Number of Entity Mentions

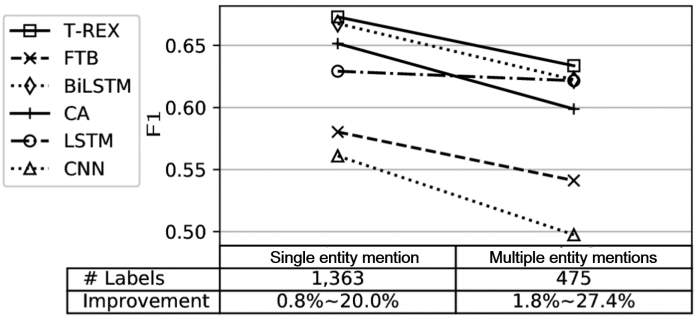
An entity tends to be mentioned multiple times in a document and each mention of an entity can be involved in a different relationship with the topic entity. To consider the subtle meaning of each mention, our topic-aware relation extraction model first predicts the relationship between the topic entity and each mention of other entities.

Then it combines the results by using a smooth-maximum function. To validate the effectiveness of our approach, we present the F1 scores of the triples with a single entity mention and multiple entity mentions, resepectively, in Figure A.1. Note that the performance improvement of *T-REX* is greater when there are multiple entity mentions. It confirms the effectiveness of our approach to consider the different meanings of entity mentions.



	Single entity mention	Multiple entity mentions
# Labels	1,363	475
Improvement	0.5%~15.7%	4.4%~20.9%

(a) Supervised setting



	Single entity mention	Multiple entity mentions
# Labels	1,363	475
Improvement	0.8%~20.0%	1.8%~27.4%

(b) Weakly supervised setting

Figure A.1: F1 score by the number of entity mentions

초 록

지식베이스는 질의응답시스템과 추천시스템, 자연어 이해 등 많은 분야에서 성공적으로 이용되고 있다. 그러나 사람이 직접 대용량의 지식베이스를 구축하는 것은 많은 시간과 노력, 금전적 비용을 초래한다. 게다가 끊임없이 새롭게 생성되는 많은 사실들을 사람이 직접 즉각적으로 업데이트하는 것은 불가능에 가깝다. 이에 따라 지난 10년간 지식베이스를 자동으로 구축하는 연구는 많은 관심을 끌어왔다.

지식융합 (knowledge fusion)은 도메인에 제한이 없는 전체 웹에서 데이터를 수집해 지식베이스를 구축하거나 확장하기 위한 대표적인 방법이다. 지식 융합은 먼저 여러 관계추출 (relation extraction) 기술을 이용해 많은 웹페이지에서 정보를 추출한다. 이때 관계추출기의 정확도 문제 등으로 인해 틀린 사실이 추출되는 경우가 자주 발생한다. 지식융합에서는 사실탐지 (truth discovery) 기술을 활용해 추출된 정보 중 정확한 정보를 찾아 지식베이스에 추가하는 일을 수행한다. 이처럼 지식융합이 관계추출과 사실탐지 두단계로 이루어져 있기 때문에 관계추출과 사실탐지의 성능이 지식베이스의 범위와 정확도를 결정한다고 볼 수 있다. 본 학위논문에서는 지식베이스의 범위와 정확도 향상을 위해 관계추출과 사실탐지의 정확도를 높이고 관계추출의 범위를 확장하는 연구를 수행한다.

딥 러닝 기술의 발전으로 많은 최근 연구들은 딥 러닝 기술을 관계추출에 활용하고 있다. 딥 러닝 학습에는 많은 양의 학습데이터가 필요하므로 주로 원격지도를 통해 자동으로 학습데이터를 생성하여 사용한다. 그러나 원격지도를 통해 생성된 데이터는 필연적으로 잘못된 레이블을 생성하는 경우가 많아 관계추출 정확도를 떨어트리는 요인으로 작용한다. 우리는 사람이 레이블을 붙인 적은 양의 학습데이터를

추가로 사용하여 관계추출의 정확도를 높이는 방법을 제안한다. 관계추출의 범위를 넓히기 위해 우리는 토픽 엔티티와 관련된 정보를 추출하는 연구를 수행한다. 토픽 엔티티는 문서에서 주로 서술되는 엔티티이다. 토픽 엔티티는 몇몇 문장에서는 대명사로 대체되거나 생략되는 경우가 많은데 이 경우 기존의 모델들은 해당 정보를 추출하지 못하는 경우가 많다. 이러한 정보를 추출하기 위해 우리는 토픽 엔티티를 고려한 관계추출 모델을 제안한다.

지식융합과정에서 사실탐지기술은 추출된 정보에서 상충되는 부분을 제거하고 정확한 정보를 찾아내는 역할을 수행한다. 사실탐지에 대한 기존 연구들에서는 하나의 대상에 대해 서로 다른 값들은 서로 배타적이어서 이 중 하나의 값만 사실이라고 가정하였다. 그러나 이러한 값 들에는 계층관계가 존재하기 때문에 서로 배타적이지 않은 경우가 많다. 따라서 우리는 계층관계를 고려하여 사실을 찾아야 한다. 우리는 계층구조를 고려한 확률모델과 이에 따른 추론 알고리즘을 제안한다. 그럼에도 불구하고, 많은 관계추출기가 비슷한 오류를 내는 경우 비지도학습인 사실탐지 기술로는 잘못된 정보를 수정하기 어렵다. 사람의 인지능력의 도움을 받아 이러한 오류를 수정하기 위해 우리는 크라우드소싱을 이용하였다. 우리는 크라우드소싱 비용이 제한된 상황에서 최대한의 정확도 증가를 얻기 위한 태스크 할당 알고리즘을 소개한다. 추가로, 태스크 할당 시 지연시간을 줄이기 위한 효율적인 필터링 기술도 제시한다.

주요어: 지식 베이스, 지식 통합, 관계 추출, 사실 탐지, 크라우드 소싱
학번: 2012-20862

감사의 글

관악에서 보낸 긴 시간만큼이나 많은 분들에게 큰 도움을 받았고 그 도움 덕분에 제가 여러 어려움과 방황을 극복하고 박사과정을 무사히 마칠 수 있었습니다. 제 마음속 감사함을 모두 담기에는 역부족이겠지만 짧은 글을 통해 감사한 마음을 전하고자 합니다.

먼저 지도교수님이신 심규석 교수님께 감사의 인사를 드립니다. 제가 아는 가장 열정적인 연구자이신 심규석 교수님은 저에게 있어 가장 큰 힘이 되는 공저자이기도 했습니다. 제가 놓쳤던 디테일한 부분까지 분석하고 바른 방향으로 이끌어 주신 교수님의 지도 덕분에 저의 연구를 더욱 발전시킬 수 있었고 연구자로써 한단계 성장할 수 있었습니다. 앞으로도 연구에 임하는 교수님의 진지하고 적극적인 자세를 마음속에 새겨두고 본받도록 하겠습니다.

그리고 바쁘신 와중에도 저의 학위논문 심사를 맡아 주신 김형주, 홍성수, 정교민, 김영훈 교수님께 감사드립니다. 교수님들께서 해 주신 조언과 격려들 덕분에 졸업논문을 더욱 좋은 내용으로 채울 수 있었을 뿐 아니라 저의 연구에 대한 흐름을 다시 한번 조망해 볼 수 있어 좋았습니다. 특히, 저에게 큰 의미가 있는 논문들을 함께 작업했던 김영훈 교수님께 다시 한번 감사말씀 드립니다.

KDDLAB에서 훌륭한 선후배님들을 만난 덕분에 지금의 성과를 얻을 수 있었습니다. KDDLAB 사람들과 토론하던 시간은 논문을 읽을 때보다 저에게 훨씬 많은 영감을 주었습니다. 특히, 제가 다양한 분야의 연구를 할 수 있었던 것은 함께 공부하며 언제나 저를 도와주던 동기 광호 덕분입니다. 함께 논문을 썼던 후배들인 수용이와 대영이에게도 고맙다는 말을 전합니다. 연구하는 도중에 만났던 크고 작

은 난관들을 이 후배들 덕분에 잘 헤쳐 나갈 수 있었습니다. 연구실 맏형 역할을 잘 해 주셨던 언제나 친절한 진현이형, 처음 데이터마이닝 공부를 할 때 많은 도움을 주었던 윤재에게도 감사의 인사를 전합니다. 지금도 열심히 연구를 하고 있는 성웅, 한준, 장혁, 영준에게는 함께하는 오랜 시간동안 도움만 받고 잘해주지 못한것 같아 고마운 마음과 미안한 마음이 동시에 듭니다. 지금도 잘하고 있고 앞으로는 더 잘할것이라 믿어 의심치 않는 후배들에게 격려와 응원의 마음을 함께 전합니다. 먼저 사회로 나가 계신 정훈이형, 정민이형, 진만이형, 희준이형, 주호형, 희원이형, 우열, 창형, 우인, 종민, 재희, 동희, 성욱, 영빈에게도 감사의 인사를 전합니다. 좋은 분들과 함께한 시간이었기에 저의 긴 대학원 생활이 힘들기만 하지는 않았습니다. 이 글에 다 싣지 못한 감사의 인사는 직접 뵙고 드리겠습니다. 앞으로도 좋은 인연을 이어갈 수 있기를 바랍니다.

학부동기인 정우, 성대, 우인, 추, 용희를 빼놓고는 20살 이후의 제 삶을 이야기 할 수 없을 것 같습니다. 함께 보낸 즐거운 시간들은 제 삶의 활력소가 되었고, 누구보다 저를 잘 이해하는 친구들의 응원은 제게 큰 힘이 되었습니다. 그리고 멀리 있어도 멀어지지 않는 나의 오랜 친구들 형준, 현수, 인준에게도 고맙다는 말을 전합니다.

마지막으로 제 졸업을 누구보다도 기뻐해준 가족들에게 가장 큰 감사의 인사를 전합니다. 지금까지 제가 걸어온 모든 한걸음 한걸음에는 언제나 부모님의 무한한 사랑과 도움이 있었습니다. 앞으로는 지금까지 보내주신 사랑에 보답하는 믿음직한 아들이 되겠습니다. 형이라는 말이 너무 잘 어울리는 나의 듬직한 형과 언제나 응원을 보내주시는 형수님에게 감사하다는 말씀을 드립니다. 그리고 곧 새로운 가족이 될 수민이의 변함없는 지지와 사랑에 고맙다는 말을 전합니다.

2021년 1월 12일 정우환

이 논문을 하늘에 계신 할머니와 외할아버지에게 바칩니다.