國立臺灣大學電機資訊學院資訊工程學系
碩士論文
Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Taiwan University
Master Thesis

中文句子語意關係之抽取
Semantic Relation Extraction from Chinese Sentences

陳昱儒
Yu-Ju Chen

指導教授：許永真博士
Advisor: Jane Yung-jen Hsu, Ph.D.

中華民國 104 年 6 月
June, 2015

國立臺灣大學碩士學位論文
口試委員會審定書

中文句子語意關係之抽取
Semantic Relation Extraction from Chinese Sentences

本論文係陳昱儒君 (R01922049) 在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 104 年 6 月 30 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

_____

_____ _____

_____ _____

_____ _____

_____ _____

所　長：_____

誌謝

感謝...

摘要

Abstract

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Nowadays, people get used to retrieving infomation from internet by computer. As an educated human being, we could understand the meaning of sentences and articles. But for computer, without structuralizing, each word is only a character.

1.2 Problem Description

1.3 Proposed Solution

1.4 Thesis Organization

Chapter 2

Background

In this chapter, we provide the introduction of knowledge representation and extraction in the first section. In the second section, we surveyed the related work about relation extraction, including the methods of supervised learning, distant supervision, and multiple instance learning. As for last section, we focus on the related work of Chinese relation extraction.

2.1 knowledge Representation and Extraction

In real world, knowledge exists in different forms like text, picture, audio, video, or even in abstract form, such as someone's memory. When considering textual knowledge, human could reason the meaning and retrieve the knowledge when it is needed. As for computers, to understand the textual knowledge, structured format is required. How to extract knowledge from unstructured text to structured representation remains an important issue. In this section, we will discuss several methods of knowledge representation and extraction.

2.1.1 Knowledge Representation

2.1.2 Knowledge Extraction

Automatic Content Extraction (ACE)[5], as a track of Text Analysis Conference (TAC) after 2009, aims at developing novel methods to extract information from natual language text. In this program, entities, relation, and events are extracted. After 2003, the program includes multilingual tracks, including English, Arabic and Chinese. The released data is used for supervised learning and promoted much good work. Some work related to this thesis will be discussed in subsequent section.

2.1.3 Relation Extraction

2.2 Related Work of Relational Pair Extraction

2.2.1 Supervised Learning

By using supervised learning, a set of training data is required and the extraction problem is formulated as classification problem. When considering single relation, the problem could be viewed as binary classification problem and aims at decide whether the relation exists in given entity pairs.

Feature-Based Methods

Kernel-Based Methods

2.2.2 Distant Supervision

To prevent ineffieiently generating training data, distant supervision is used for reducing the labelling work. Distant supervision uses weakly labelled training data to predict huge testing data. For relation extraction with distant supervision, Mintz[6] used Wikipedia as corpus and Freebase as seed pool for training a classification model. A strong assumption of distant supervision is that if two entities participate in one relation, then any sentence contains the two entity can represent the relation.

2.2.3 Multiple Instance Learning

Given Distant supervision, the labelling effort is heavily reduced but cause the problem of noise when the data sentences are correlated with the seed. For example, considering 2 sentences, “Alice was born in Taipei” and “Alice went to Taipei on Saturday” both contains the two entities “Alice” and “Taipei”, but the relation in the former sentence is $WasBornIn$ while in the latter sentence is $WentTo$. Riedel[7] indicated that 31% alignments of Freebase and New York Time Corpus violate the distant supervision assumption while only 13% ones of Freebase and Wikipedia violate.

2.3 Relational Pair Extraction in Chinese

2.3.1 Characteristics in Chinese Relational Pair Extraction

2.3.2 Related Work in Chinese

Chapter 3

Methodology

This first section defines the problem of relational pair extraction. Then the framework for solving this problem is presented in the next section. The framework includs four key components: data labelling, feature extracting, model learning, and iteratively training. The 4 components are explained in the following four sections.

3.1 Problem Definition

Considering the scenario of relational pair extraction, given a set of entity pairs as seeds indicating a relation, we are going to extract new pairs representing such relations from a corpus. The details are described in this section and begin with defining the notations used in this thesis.

3.1.1 Notations

First, we let C denote a corpus. Each $s \in C$ is a sentence, which is constructed by words. Given a corpus C, an entity set is defined as $E = \{e \mid e\ is\ a\ word\ in\ C\}$. Then we let R denote a relation set. Each $r \in R$ is a relation, corresponding to a seed set $S_r = \{(r, e_i, e_j) \mid r \in R;\ e_i, e_j \in E\}$. The tuple $(r, e_i, e_j) \in S_r$ indicates that 2 entities e_i and e_j are semantically connected with the relation r. In this problem, a new pair set is defined as $N_r = \{(r, e_i, e_j) \mid r \in R; e_i, e_j \in E; (r, e_i, e_j) \notin S_r\}$.

3.1.2 Problem Definition

Given a corpus C and a seed set S_r, the relational pair extraction system will create a new pair set N_r. The pairs in N_r are extracted from C and excluded from S_r.

- **Input**: a corpus C, a seed set $S_r = \{(r, e_i, e_j) \mid r \in R; e_i, e_j \in E\}$
- **Output**: a set of new pairs $N_r = \{(r, e_i, e_j) \mid r \in R; e_i, e_j \in E; (r, e_i, e_j) \notin S_r\}$

For example, to extract new pairs related to the relation $AtLocation$ in a corpus C, the corpus C and a seed set $S_{AtLocation}$ are given as following. The sentences in C are selected from Wikipedia[10].

- Seed set $S_{AtLocation}$

  (AtLocation, Taipei, Taiwan)

  (AtLocation, Tokyo, Japan)

- Corpus C

  Taipei City is the capital city and a special municipality of **Taiwan**.

  Tokyo is the capital and largest city of **Japan**.

  Seoul Special City is the capital and largest metropolis of **South Korea**.

  Beijing is the capital of the People's Republic of **China**.

The new seed pairs related to $AtLocation$ are extracted from C, as shown in $N_{AtLocation}$.

- New seed set $N_{AtLocation}$

  (AtLocation, Seoul, South Korea)

  (AtLocation, Beijing, China)

3.2 Framework

The overall framework of the relational pair extraction system is shown on Figure 3.1, and the process is defined as Algorithm 1. The framework are separated into 3 parts: **bag generator**, **relation predictor** and **pair evaluator**.

Algorithm 1 Overall process of relational pair extraction

Input: a set of seeds $S_r^{(1)}$, a corpus C, an set of entities E, maximal iteration number M
Output: a set of new pairs N_r

generate a unlabelled pair set $U = \{(e_i, e_j) \mid e_i, e_j \in E\}$ from C
for t = 1 to M **do**
    generate a training bag set $B_{train}^{(t)}$ from C and $S_r^{(t)}$ with **Bag Generator**
    generate a testing bag set $B_{test}^{(t)}$ from C and U with **Bag Generator**
    train a model **Relation Predictor** with $B_{train}^{(t)}$
    with the **Relation Predictor**, predict labels for all data in $B_{test}^{(t)}$
    select positive pairs from $B_{test}^{(t)}$ as $N_r^{(t)}$
    generate new seed set $S_r^{(t+1)}$ from $N_r^{(t)}$ by **Pair Evaluator**
end for
return $N_r^{(1)} \cup N_r^{(2)} \cup ... \cup N_r^{(M)}$

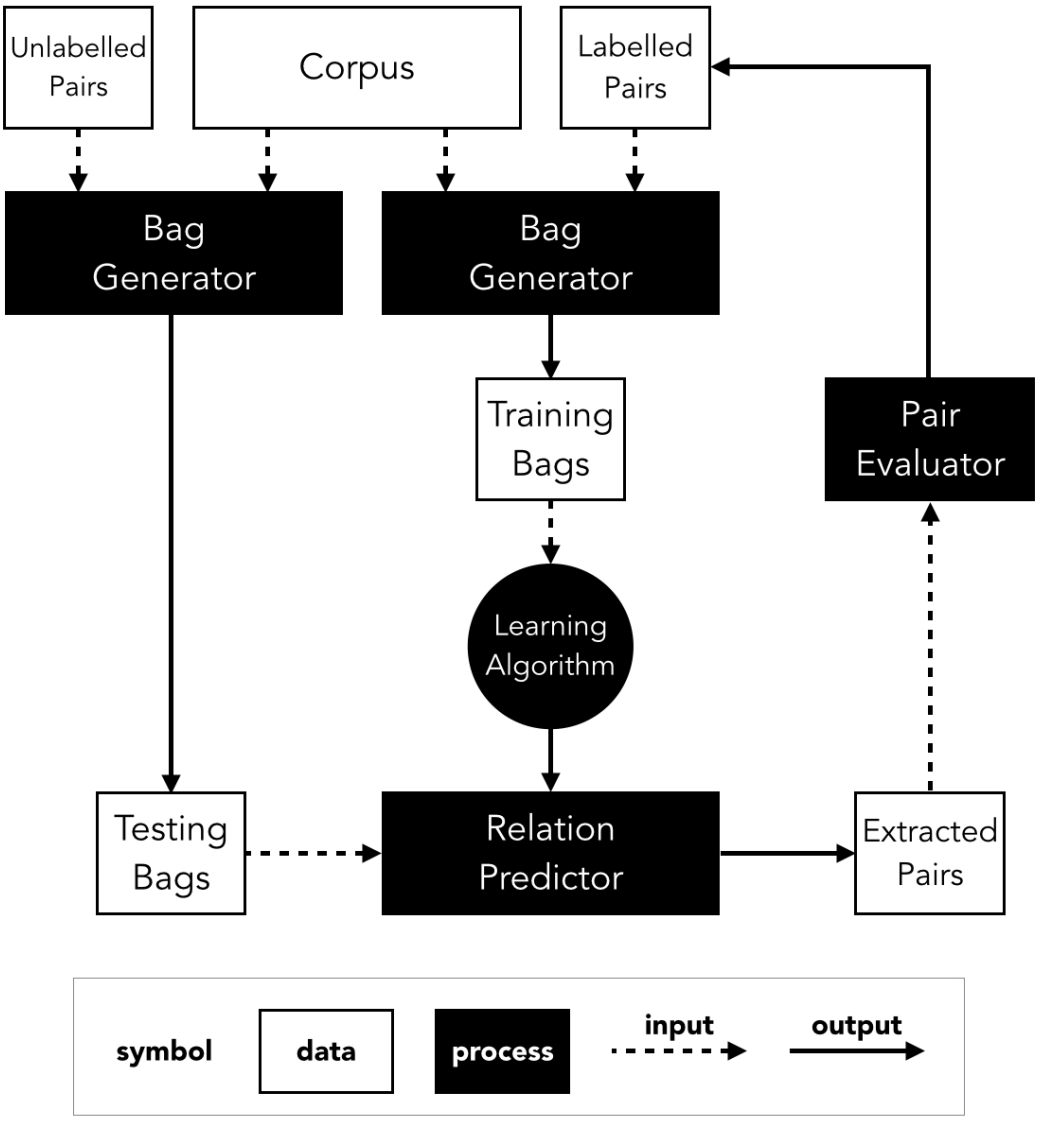

Figure 3.1: Framework of the relational pair extraction system

3.2.1 Bag Generator

The bag generator prepares training data and testing data for the relation predictor (Section 3.2.2). Each data is a bag, which collects sentences from a corpus and corresponds to an entity pair. For example, a bag of the pair ("**Taipei**", "**Taiwan**") consists of sentences mentioning "**Taipei**" and "**Taiwan**", such as "**Taipei** is the capital of **Taiwan**", "**Taipei** is the political center of **Taiwan**".

In Figure 3.1, training bags are generated from labelled pairs and testing bags are generated from unlabelled pairs. A labelled pair (r, e_i, e_j) corresponds to a labelled bag (y, b); an unlabelled data (e_i, e_j) corresponds to an unlabelled bag b. The bag generator aims at mapping pairs to bags. If the pairs are labelled, the label will be brought to the bag.

In this work, given a corpus C and a labelled pair set, also known as seed set $S_r = \{(r, e_i, e_j)\}$, a training bag set B_{train} is generated. For each $(y, b) \in B_{train}$, b is a bag of feature vectors and $y \in R$ refers to the label of this bag. The details of feature vector are described in Section 3.3.

Another bag set $B_{test} = \{b\}$ is used for testing. It is generated with a corpus C and an unlabeled pair set $U = \{(e_i, e_j) \mid e_i, e_j \in E\}$. The details of automatic labelling are explained in Section 3.4.

The input and output of a bag generator is defined as following:

- **Input**: a corpus C, a pair set S
- **Output**: a bag set B

3.2.2 Relation Predictor

The relation predictor is used for generating new pairs from the corpus as a standard machine learning process. With training bag set B_{train} and an algorithm $\mathcal{A}$, the predictor is created to predict the label of each bag $b_i \in B_{test}$.

In this work, tha algorithm $\mathcal{A}$ is a multiple instance learning algorithm due to the restriction of the problem. More details of the multiple instance learning process are de-

scribed in Section 3.5.

The input and output of the relation predictor is defined as following:

- **Input**: a training bag set B_{train}, a testing bag set B_{test}, a learning algorithm $\mathcal{A}$
- **Output**: a set of new pairs N

3.2.3 Pair Evaluator

To iteratively learn new pairs from the corpus, we update the seed set for each iteration. To avoid using the false positive pairs as seeds in the next iteration, the result should be evaluated by another mechanism. Here we use human intelligence as the evaluator.

Given the new pair set $N_r^{(t)}$ generated in the t^{th} iteration, we ask human to evaluate the correctness and generate another set $S_r^{(t+1)}$, which is the seed set in the next iteration. The details of this process are illustrated in Section 3.6.

The input and output of the pair evaluator is defined as following:

- **Input**: a set of candidate pairs $N_r^{(t)}$
- **Output**: a set of confident pairs $S_r^{(t+1)}$

3.3 Features of Data

When generating the training data, we transfer plain texts to features, including syntactic, semantic, and textual forms. Zhou(2005)[2] has addressed several NLP features for relation extraction, including bag of words, parse tree, entity type, and other features. Consider the example as following:

- Sentence

  想到台大學生在校門口為言論自由絕食靜坐

  (think of Students from National Taiwan University sitting before the school gate, having hunger strike and sit-in demonstration for freedom of speech)

Feature name	Explaination	Example
W_1	the first entity	學生
W_2	the second entity	校門
BW_1	bag of words before the first entity	想到, 台大
BW_2	bag of words after the second entity	口, 為, 言論, 自由, 絕食, 靜坐
BW_12	bag of words between the 2 entities	在
POS_1	pos tags before the first entity	D, VE, Nc
POS_2	pos tags after the second entity	Ncd, P, Na, Na, VA, VA
POS_12	pos tags between the 2 entities	P
DEP_12	dependency tree structure between the 2 entities	NN ← NP ← IP → VP → PP → LCP → NP → NN
ORDER	order of the object and location	object → location

Table 3.1: List of features, with the example related to sentence "想到台大學生在校門口為言論自由絕食靜坐" and entity pair ("學生", "校門")

想到	台大	學生	在	校門	口	為	言論	自由	絕食	靜坐
VE	Nc	**Na**	P	**Na**	Ncd	P	Na	Na	VA	VA

Table 3.2: part-of-speech example related to sentence "想到台大學生在校門口為言論自由絕食靜坐", the red words are entity pairs

- Entity pair

  學生 (student), 校門 (school gate)

The features used in this thesis are shown in Table 3.1. The part-of-speech tag[9] of this example is illustrated in Table 3.2 and dependency tree structure[1] is represented as Figure 3.2[8].

3.4 Assistant Labelling

Since the size of training bags used in this thesis is very large, it is difficult to label every data manually. In this section, we propose a process to automatically label the training bags. We follow the distant supervision assumption of relation extraction provided by Mintz(2009)[6]:

Assumption 1. *If two entities participate in a relation, all sentences that mention these two entities express that relation.*

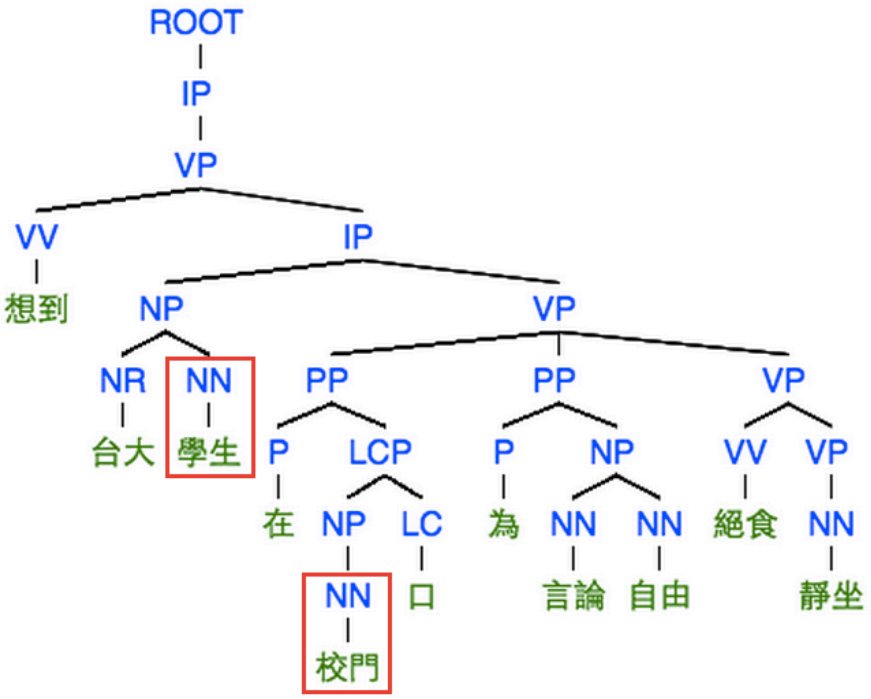

Figure 3.2: dependency tree example related to sentence "想到台大學生在校門口為言論自由絕食靜坐", the words in red circles are entity pairs

With Assumption 1, we could match any seed (r, e_i, e_j) to sentences mentioning e_i and e_j, and assign the sentences with corresponding label r. But this assumption is too strong because not all sentences containing (e_i, e_j) express the relation r. So Hoffmann(2011)[3] addressed a modified assumption by adapt the problem to a multiple instance learning problem.

Assumption 2. *If two entities participate in a relation, at least one sentence that mentions these two entities might express that relation.*

According to Assumption 2, given any seed (r, e_i, e_j) and a bag of sentences mentioning e_i and e_j, at least one sentences in the bag might express r. Instead of describing a relation with a $sentence$, here we use a bag to represent a relation. The Assumption 2 is midified as following:

Assumption 3. *If two entities participate in a relation, a bag of sentences that mention these two entities might express that relation.*

Any entity pair (r, e_i, e_j) in the seed set may correspond to a bag of sentences $b \subset C$,

where each sentence $s \in b$ contains the 2 entity e_i and e_j. Then we decide the label y for any b. When at least one sentence in b represents the relation r, $y = r$. Otherwise, if the given relation r does not shown in b, $y = null$. The process of automatic labelling for an entity pair (e_i, e_j) and corresponding set b is displayed in Algorithm 2.

Algorithm 2 Process of automatic labelling training data

Input: a corpus $C = s$, a seed set $S_r = \{(r, e_i, e_2) \mid r \in R;\ e_i,\ e_2 \in E\}$
Output: a labelled bag set $B_{train} = \{(y, b)\}$

```
initial a bag set B
for all (r, e_i, e_j) ∈ S_r do
    generate a sentence set: b_i = {s | s ∈ C; e_j, e_k are words in s}
    assign the label for b_i: y_i = r_i
    add (y_i, b_i) to B
end for
return B
```

3.5 Multiple Instance Learning

Since we adapt the assumption of multiple instance learning to the problem, we choose the algorithm of multiple instance classifier as our core process in the system.

3.5.1 Instance Space Learning

3.5.2 Bag Space Learning

3.5.3 Embedded Space Learning

3.6 Iteratively Learning Process

Chapter 4

Experiment and Result

This chapter shows the experiment for Chinese relational pair extraction. First, the data source and data generating process will be explained. Then the experimentl setting will be presented in the second section. In the last 2 sections, the result for testing the proposed framework with given data will be discussed.

4.1 Dataset

Distant supervision learning, which is an semi-supervised framework, uses small size of labelled data and huge size of unlabelled data for training. In relation extraction work, the labelled data are usually from an existing dataset and the unlabelled data are from a huge corpus containing many sentences, such as news corpus or Wikipedia. In this thesis, we use Chinese entities pairs from ConceptNet as labeled data and sentences in Taiwan Sinica Corpus as unlabeled data.

4.1.1 ConceptNet

ConceptNet is developed by MIT Media Lab[4], containing only English data at beginning. It is a graphical structure, with node representing $concept$, and link representing $relation$, describing a commonsense knowledge base. A commonsense fact, formulated as two concepts and one relation bwtween them, is called $assertion$, which is stored as 2 nodes ($concepts$) and 1 link ($relation$) in the ConceptNet graph. The knowledge in

Relation Name	Number of Assertions
AtLocation	32816
CausesDesire	19408
HasProperty	6822
NotDesires	23930
UsedFor	13548
Causes	77336
HasSubevent	40655
PartOf	6159
Desires	21772
IsA	16094
HasFirstSubevent	12046
MadeOf	16357
CapableOf	27444
SymbolOf	4736
MotivatedByGoal	56636

Table 4.1: List of relations corresponding to Chinese knowledge in ConceptNet

ConceptNet is originally provided by human as crowesourcing. Afterwards, the authors imported knowledge from other datasets such as WordNet, Wikipedia, Wikitionary, DBPedia, and other sources. From ConceptNet 3, by collaborating with universities from different countries, ConceptNet absorbed data from other languages, becoming a multilingual commonsense dataset. Chinese assertions are collected from an online pet game, by feeding the pets new knowledge.

Before ConceptNet 4, the relations of ConceptNet is predefined. When collecting knowledge in Chinese, only 15 relations are considered (shown in Table 4.1). The relation $AtLocation$ represents the relationship between an object and a location. For example, $AtLocation(Taipei, Taiwen)$ indicates that "Taipei is located in Taiwan".

Each assertion is regarded as labelled data when training a distant supervision model. The experiment uses Chinese assertions related to the relation $AtLocation$. Since Chinese assertions in ConceptNet are generated from online users, the reliability is not guaranteed. Among all 32816 assertions, only about 37% of data are valid for training. The examples of seeds from ConceptNet are shown in Table 4.2.

label	entity (object)	entity (location)
O	教授 (professor)	研究所 (graduate school)
O	員工 (staff)	公司 (company)
O	人民 (people)	台灣 (Taiwan)
O	學生 (student)	學校 (school)
O	病人 (patient)	醫院 (hospital)
O	鯉魚 (carp)	池塘 (pond)
X	美國 (USA)	亞洲 (Asia)
X	程式 (program)	電腦 (computer)
X	太陽 (sun)	夏天 (summer)
X	觀眾 (audience)	電視 (television)

Table 4.2: Example of positive and negative entity pairs of Chineses knowledge in ConceptNet

4.1.2 Sinica Corpus

The full name of Sinica Corpus[9] is "Academia Sinica Balanced Corpus of Modern Chinese", developed from 1994. This corpus contains articles from 1981 to 2007, spread in different fields including philosophy, science, society, art, life, and literature. The source of articles is also diverse, including news, book, textbook, magazine, and so on, which shows the corpus contains different styles of articles. Each article is stored as sentences by segmenting the article with comma, period, semicolon, question mark and exclamation mark. In Chinese articles, the usage of comma is slightly different from English. Sometimes, the sentence before the comma and after the comma can be independent in grammar while they may be dependent in meaning. The corpus is separated as about 6 hundred thousand sentences and each sentence is segmented as words. Each word is annotated with a part-of-speech tag. Since the articles of Sinica Corpus are cross many years, the phrasing may be slightly diferent to the wording nowadays. The wording is more conscientious and careful because the sources are formal media. The example of sentences are shown on Table 4.3

The experiment uses the sentences in Sinica Corpus as unlabelled data. Each sentences are regarded as a list of words and each 2 words in the sentence may convey one or no relation. The method for generating training data from sentences is described in Section 3.4 on page 12.

Sentence	Translation
民族學 (Na) 研究所 (Nc) 應 (D) 主持 (VC) 之 (DE) 「(PARENTHESISCATEGORY) 台灣 (Nc) 與 (Caa) 東南亞 (Nc) 土著 (Na) 文化 (Na) 與 (Caa) 血緣 (Na) 關係 (Na) 」(PARENTHESISCATEGORY) 主題 (Na) 研究 (Nv) 計劃 (Na) 之 (DE) 需要 (Na) ，(COMMACATEGORY)	According to the need for hosting the research project “aboriginal culture and blood relationship bwtween Taiwan and Southeast Asia”, Institute of Ethnology ...
邀請 (VC) 蘇聯 (Nc) 國家 (Na) 科學院 (Nc) 世界 (Nc) 文學 (Na) 研究所 (Nc) 研究員 (Na) Ｂｏｒｉｓ (FW) Ｐａｒｎｉｃｋｅｌ (FW) 教授 (Na) 於 (P) 六月 (Nd) 十九日 (Nd) 至 (P) 廿六日 (Nd) 來訪 (VA) ，(COMMACATEGORY)	... invited the researcher, Professor Boris Parnickel, who comes from Institute of World Literature of Russian Academy of Sciences, to visit from 19th to 26th June.
Ｂｏｒｉｓ (FW) 教授 (Na) 之 (DE) 專長 (Na) 為 (VG) 東南 (Ncd) （(PARENTHESISCATEGORY) 特別 (VH) 是 (SHI) 馬來亞 (Nc) 及 (Caa) 印尼 (Nc) ）(PARENTHESISCATEGORY) 的 (DE) 神話 (Na) 傳說 (Na) 及 (Caa) 民俗 (Na) ，(COMMACATEGORY)	Professor Boris specializes in the legend and folklore in Southeast Asia, especially in Malaysia and Indonesia, ...
在 (P) 此 (Nep) 一 (Neu) 領域 (Na) 已 (D) 有 (V_2) 傑出 (VH) 之 (DE) 研究 (Nv) 成果 (Na) 。(PERIODCATEGORY)	... having excellent achievements in this area.

Table 4.3: Example of sentences in Sinica Corpus

4.1.3 UDN News Dataset

UDN.com is a news website in Taiwan, reporting various types of online news. Different from Sinica Corpus, articles in UDN news are less formal, including many novel terms. Sentences in UDN news are used as unlabelled data. Articles are segmented to sentences by a period stop, which may conserve complete meaning in the sentence. The articles are raw texts without any label or processing. Before training, these sentences have to be segmented to words, labelled with part-of-speech tag, and other grammatically parsed with external NLP tools.

There are two sets of UDN data collected for experiment. One is recent news on the real time news column (from February to May 2015), and another is historical news in the past 10 years, including 3 different columns. The difference of these 2 dataset is the size and age of news.

4.2 Experiment setting

4.2.1 Experiment 1: Feature selection

4.2.2 Experiment 2: MIL Algorithm selection

4.2.3 Experiment 3: Iteration

4.2.4 Experiment 4: Multiple Relations

4.3 Evaluation

We use $Accuracy$ for evaluating the result of cross validation:

$$Accuracy = \frac{|accuate\ testing\ data|}{|total\ testing\ data|} \tag{4.1}$$

4.4 Result and Discussion

Chapter 5

Application

Chapter 6

Conclusion

Bibliography

[1] P.-C. Chang, H. Tseng, D. Jurafsky, and C. D. Manning. Discriminative reordering with Chinese grammatical relations features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*, Boulder, Colorado, June 2009.

[2] Z. GuoDong, S. Jian, Z. Jie, and Z. Min. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 427–434. Association for Computational Linguistics, 2005.

[3] 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.

[4] M. M. Lab. Conceptnet 5. http://conceptnet5.media.mit.edu/. Accessed: 2015-05-22.

[5] L. D. C. (LDC). Automatic content extraction (ace). https://www.ldc.upenn.edu/collaborations/past-projects/ace. Accessed: 2015-05-14.

[6] 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.

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

[8] M. Shang. Syntax tree generator. https://github.com/mshang/syntree, 2011.

[9] A. Sinica. Academia sinica balanced corpus of modern chinese. http://rocling.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm. Accessed: 2015-05-23.

[10] Wikipedia. Wikipedia, the free encyclopedia. http://en.wikipedia.org. Accessed: 2015-06-03.