

國立臺灣大學管理學院資訊管理學研究所

碩士論文

Department of Information Management

College of Management

National Taiwan University

Master Thesis

主題文件內人際互動關係擷取之研究

FISER: An Effective Recognizer for Detecting

Topic-dependent Interactive Relation



Pi-Hua Chuang

指導教授：陳建錦 博士

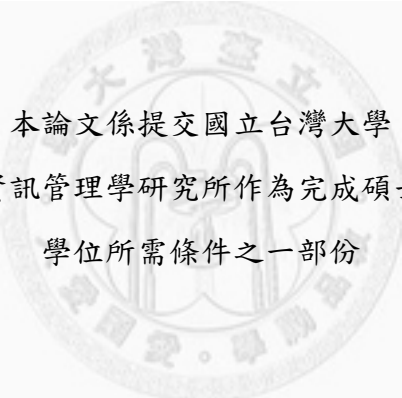
Advisor: Chien Chin Chen, Ph.D.

中華民國 101 年 7 月

July, 2012

主題文件內人際互動關係擷取之研究

FISER: An Effective Recognizer for Detecting  
Topic-dependent Interactive Relation



本論文係提交國立台灣大學  
資訊管理學研究所作為完成碩士  
學位所需條件之一部份

研究生：莊璧華 撰

中華民國 101 年 7 月

國立臺灣大學碩士學位論文  
口試委員會審定書

主題文件內人際互動關係擷取之研究

FISER: An Effective Recognizer for Detecting  
Topic-dependent Interactive Relation

本論文係莊璧華君（學號 R99725008）在國立臺灣大學資訊管理學系、所完成之碩士學位論文，於民國 101 年 7 月 9 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：

陳建勳

戴碧如

陳昌勳

李正帆

所長：

李瑞庭

## 致謝

還記得剛從大學畢業、到台北熟悉全新環境的那些日子，當初帶著不安又雀躍的心情來到這個的城市。兩年過去了，又見鳳凰花開的季節，我即將要繼續前往下一個人生的階段、面對全新的目標。回首過往的日子，就像陳之藩說的，得之於人者的太多了，只好感謝上天讓我遇見你們。生命裡面處處有貴人，所以我才能夠順利地從研究所畢業。

首先，要謝謝我的指導教授—陳建錦老師。在研究學習的過程中，老師給了我許多寶貴的建議並且有耐心的解說和指導，在這兩年之內，從老師身上學到了很多，不管是學術領域還是待人處事方面。除了課業方面，老師也時常關心學生的生活，時時關心我們英文畢業門檻考試通過了沒、修課的狀況、論文和實驗的進度，還有求職順不順利…等等。可以在求學生涯當中遇見這麼不可多得的好老師真是我的福氣！建錦老師，真的非常謝謝您兩年來的指導。

再來，也要謝謝我親愛的家人。謝謝爸爸、媽媽從小到大對我的栽培與付出，讓我可以求學的生涯中無後顧之憂的專心於課業，還時常打電話叮嚀要注意自己的身體、好好照顧自己…等；謝謝姑婆這兩年來的照顧，常常幫我準備很多水果，讓外食族的我也有均衡的飲養；謝謝阿筆，每天陪我講電話，聽我分享或是抱怨日常生活的小事，還有在面對許多決定時提供我寶貴的意見，而且也帶我出去玩樂、拍照留下很多好的回憶，順便抒發壓力。謝謝親愛的家人們，讓我在累的時候有個可以休息、再出發的地方。

接著，要感謝我的好朋友們。謝謝仲詠學長，這兩年來在他身上學習到很多，也謝謝他在課業與研究上的指導、熱心的幫忙日常生活中的大小事，還有一起修課同甘共苦、聊天閒話家常的珍貴回憶；謝謝小麥，這兩年來的照顧，從一開始認識校園、認識新朋友、課業上都幫助我很多；謝謝詠淳學長，在論文上給予的建議與幫助；謝謝孟潔、小美、卡布、upu，有這些可愛的學弟妹讓我們實驗室的氣氛非常的歡樂，還有超棒的WEAL SIGIT系列研討會；謝謝江其其，很高興能在這裡遇見這麼有趣的好朋友，並且一起渡過課表完全一樣、又歡樂又崩潰的碩一下，還有不嫌棄我總是剛好在她要考西文的時候到管五去找她聊天；謝謝蘇灌，常常對我的關心、打氣，還有常常提出無敵中肯的建議來點醒我，果然是中肯哥。

最後，要感謝口試委員—陳孟彰老師、戴碧如老師、李正帆老師，在口試中給予我的寶貴意見，讓我從其他角度更了解自己的論文。

莊璧華 謹識  
于台大資管研究所  
民國101年7月

# 論文摘要

論文題目：主題文件內人際互動關係擷取之研究

作者：莊璧華

101 年 7 月

指導教授：陳建錦 博士

由於Web2.0的發展，網際網路使用者處於資訊爆炸的時代。在面對大量文章的時候，先找出主題文件中人與人之間的互動關係將有助於閱讀者建立主題文件的背景架構以及對內容有初步的理解。為了找出人與人之間的互動關係，我們需要一個方法先辨別文字片段中是否有互動關係存在，接著再使用資訊擷取的演算法分析人物之間的互動關係，並且將同一主題文章中的人物建立其互動關係網路。

在這次的研究當中，我們將互動關係辨識定義成分類問題。結合句子中語法、語意和語境的資訊，設計出十九個語言的特徵來辨別文字片段當中是否有互動關係存在。實驗的結果顯示我們設計的互動關係辨識的方法是有效的，也優於其他著名的開放式資訊擷取系統。

關鍵字：資訊擷取、互動關係擷取、開放式資訊擷取

# THESIS ABSTRACT

## **FISER: An Effective Recognizer for Detecting Topic-dependent Interactive Relation**

By Pi-Hua Chuang

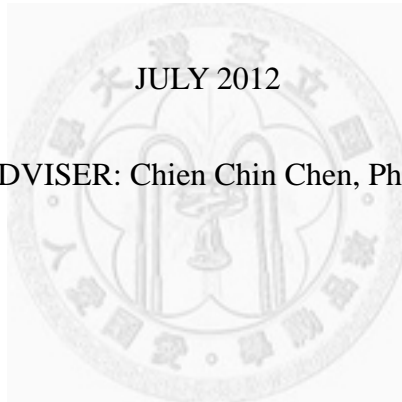
MASTER DEGREE OF BUSINESS ADMINISTRATION

DEPARTMENT OF INFORMATION MANAGEMENT

NATIONAL TAIWAN UNIVERSITY

JULY 2012

ADVISER: Chien Chin Chen, Ph.D.



Discovering the interactions between the persons mentioned in a set of topic documents can help readers construct the background of the topic and facilitate document comprehension. To discover person interactions, we need a detection method that can identify text segments containing information about the interactions. Information extraction algorithms then analyze the segments to extract interaction tuples and construct an interaction network of topic persons. In this paper, we define interaction detection as a classification problem. The proposed interaction detection method, called FISER, exploits nineteen features covering syntactic, context-dependent, and semantic information in text to detect inter-sentential and

intra-sentential interactive segments in topic documents. Empirical evaluations demonstrate that FISER outperforms many well-known open IE methods on identifying interactive segments in topic documents. In addition, the precision, recall and F1-score of the best feature combination are 72.6%, 55.6%, and 61.9% respectively.

Keywords : Information extraction, Relation detection, Open IE



# Table of Contents

致謝 .....	i
論文摘要 .....	ii
THESIS ABSTRACT .....	iii
Table of Contents .....	v
List of Tables .....	vi
List of Figures .....	vii
1. Introduction .....	1
2. Related Works .....	4
2.1 Relation Extraction .....	4
2.2 Open IE .....	7
3. Methodology .....	10
3.1 Candidate Segment Generation .....	11
3.2 Interactive Segment Recognizer .....	14
3.3 Feature Extraction .....	15
4. Performance Evaluation .....	20
4.1 Data Corpus and Evaluation Metrics .....	20
4.2 The Performance of the Features .....	24
4.3 The Best Combination of the Features .....	26
4.4 Comparison with Open IE Methods .....	29
4.5 The Effectiveness of the Features .....	34
5. Conclusion .....	38
References .....	40



# List of Tables

Table 4-1. The statistics of data corpus .....	21
Table 4-2. Ten political topics in Taiwan from 2004 to 2010.....	23
Table 4-3. Experimental result of each feature category.....	25
Table 4-4. The effect of features .....	28
Table 4-5. The features of the comparative methods .....	31
Table 4-6. The interaction detection result of compared methods .....	32



# List of Figures

Figure 3-1. The system architecture .....	11
Figure 3-2. Candidate segment generation algorithm .....	13
Figure 3-3. Examples of candidate segments .....	15
Figure 4-1. The lift curve of syntactic features .....	35
Figure 4-2. The lift curve of context-dependent features .....	36
Figure 4-3. The lift curve of semantic features .....	37



# 1. Introduction

The Web has become an abundant source of information because of the prevalence of Web2.0, and Internet users can express their opinions about topics easily through various collaborative tools, such as weblogs. Published documents provide a comprehensive view of a topic, but readers are often overwhelmed by large number of topic documents. To help readers comprehend numerous topic documents, several topic mining methods have been proposed. For instance, (Feng and Allan, 2007) grouped topic documents into clusters, each of which represents a topic incident. The clusters are then connected chronologically to form a timeline of the topic. Recently, (Chen and Chen, 2012) summarized the incidents of a topic timeline to help readers understand the story of a topic quickly.

Basically, a topic is associated with specific times, places, and persons (Nallapati *et al.*, 2004). Discovering the interactions between the persons can help readers construct the background of the topic and facilitate document comprehension. According to (Vernon, 1965), interaction is a kind of human behavior that makes people take each other into account or have a reciprocal influence on each other. Examples of person interactions include compliment, criticism, collaboration, and competition. The discovery of topic person

interactions involves two key tasks, namely *interaction detection* and *interaction extraction*.

Interaction detection first partitions topic documents into segments and identifies the segments that convey possible interactions between persons. Then, interaction extraction applies an information extraction algorithm to extract interaction tuples from the interactive segments. In this paper, we investigate interaction detection. In contrast to open IE research (Banko *et al.*, 2007), which focuses on discovering static and permanent relations (e.g., *capital of*) between entities, the interactive relations we investigate are dynamic and topic-dependent. For instance, in a topic about next generation PCs, Bill Gates and Steve Jobs were in a criticism relationship because Bill Gates criticized Steve Jobs about the usability of iPad. However, in the topic about Steve Jobs's death, they had a complimentary relationship since Gates eulogized Jobs for his remarkable contributions to human civilization. To identify dynamic interactions between persons, we define interaction detection as a classification problem. We also propose an effective interaction detection method, called FISER (Feature-based Interactive SEgment Recognizer), which employs nineteen features covering syntactic, context-dependent, and semantic information in text to detect interactive segments in topic documents. Our experiment results show that FISER can identify interactive segments accurately and the proposed features outperform those of well-known open IE systems dramatically.

The remainder of this article is organized as follows. In Section 2, we discuss traditional relation extraction, open IE and explain how them differ from our research. We describe the proposed FISER method in Section 3. In Section 4, we evaluate FISER's performance and compare with open IE methods. Then, we present our conclusions in Section 5.



## 2. Related Works

### 2.1 Relation Extraction

Relation extraction (RE) has been promoted by the Automatic Content Extraction (ACE) program<sup>1</sup> and is an essential research field in natural language processing or information retrieval. The task of relation extraction is to extract semantic relations between pairs of entities from text. In the program, entities are objects and there are five types of entities, namely, persons, organizations, locations, facilities, and geo-political entities. Semantic relations between entities are predefined. For instance, ORG-AFF refers affiliation relations between organizations. Previous RE methods can be classified as supervised learning approach and bootstrapping approach. The supervised approach (Chieu and Ng, 2002; Zelenko *et al.*, 2003; Culotta and Sorensen, 2004; Kambhatla, 2004) considers RE as a classification problem. Given a training corpus which consists of a set of human-tagged examples of predefined relations, a classification algorithm is employed to train a RE classifier, which assigns (i.e., classifies) a relation type to a new text segment (e.g., a sentence). Feature-based methods (Chieu and Ng, 2002; Kambhatla, 2004) and kernel-based

---

<sup>1</sup> <http://projects ldc.upenn.edu/ace/>

methods (Zelenko *et al.*, 2003; Culotta and Sorensen, 2004) are the two frequently used techniques of the supervised RE approach. The feature-based methods exploit the positive and negative relation instances in a training corpus to identify effective text features of relation extraction. (Kambhatla, 2004) integrated lexical, syntactic, and semantic features of text into a maximum entropy model to extract semantic relations between entities from the ACE official dataset. However, not all features are useful in extracting entity relations. Therefore, selecting effective features becomes an important but difficult problem. In order to address this problem, some research studies (Zelenko *et al.*, 2003; Zhou *et al.*, 2010) design kernel-based approaches to extract entity relations. A kernel represents an input text (e.g., a sentence) in a high dimensional kernel space. Each dimension represents a subsequence of the text. Given two sentences represented in a high dimensional kernel space, the kernel-based approaches calculate the similarity of sentences. The more identical subsequences between the sentences, the higher the similarity will be. The bag-of-words kernel and the tree kernel are two most frequently used kernel spaces. The difference between them is that the bag-of-words kernel focuses on the similarity (i.e., overlap) of words in the given sentences, but the tree kernel measures the common parts of the parsing trees of the given sentences. (Zelenko *et al.*; 2003) developed a tree kernel-based method which recursively matches the parsing trees of sentences in a top-down manner. (Culotta and Sorensen, 2004) extended Zelenko *et al.*'s work to compute the similarity of sentences in terms of their augmented dependency trees.

In general, compiling training dataset is laborious. The bootstrapping approaches (Agichtein and Gravano, 2000; Pantel and Pennacchiotti, 2006) eases the preparation of training dataset by training an extraction model with a small set of training seeds. Newly discovered entity relations then are included in the training dataset to refine the extraction model iteratively and to identify new entity relations. While the approach significantly reduces the effort of training dataset preparation, it prolongs the construction of extraction models and the constructed model is domain specific. Moreover, many bootstrapping methods identify entity relations based on keyword matching rules that lead to low precision and low recall (Pantel and Pennacchiotti, 2006).

The ACE program defines different kinds of relation types and subtypes, and provides official benchmarks of the relation types to promote the research of relation extraction. Many previous research studies employed syntactic information of text to detect the predefined relations and few studies considered context and semantic information in text. In this paper, we detect interactions between persons, which are diverse, topic-dependent, and cannot be defined ahead. We propose effective semantic and context-dependent text feature to detect interactive segments correctly.

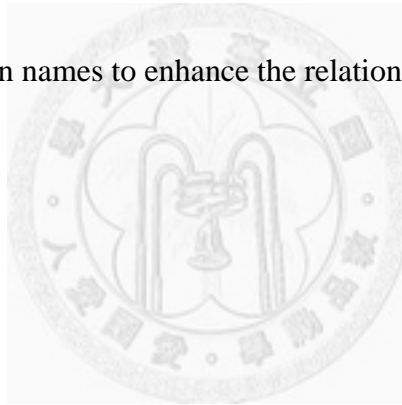


## 2.2 Open IE

Differing from the traditional RE we mentioned before, our research is closely related to open IE, which is a novel information extraction paradigm proposed by (Banko *et al.*, 2007). In open IE, the objective is to recognize the relations between entities without providing any relation-specific human input. Like our approach, open IE involves two tasks, namely, *relation detection* and *relation extraction* (Hirano *et al.*, 2007; Li *et al.*, 2008; Hirano *et al.*, 2010). The former determines whether a text segment conveys a relation between the entities, and the latter extracts relation tuples from the relation segments. (Li *et al.*, 2008) demonstrated that relation detection is critical to outputting reliable relation tuples. However, our survey of open IE literature revealed that most approaches omit relation detection, or they exploit simple heuristics to detect relation segments. For example, (Culotta and Sorensen, 2004) combined the dependency tree of a text expression with syntactic features, such as part-of-speech (POS) tags, to detect and extract relation tuples. TEXTRUNNER (Banko *et al.*, 2007) employs six syntactic features to detect relation segments in a text corpus. The drawback with the above approaches is that they do not consider text semantics, so they may not perform well in terms of relation detection (Hirano *et al.*, 2010). In (Banko and Etzioni, 2008), the authors view relation extraction as a sequence labeling problem, and employ conditional random fields (CRFs) (Lafferty *et al.*, 2001) to recognize relation expressions. Because the models are trained and tested with relation segments, in practice, a relation detection component is needed to achieve a good relation extraction performance (Li *et al.*,

2008; Hirano *et al.*, 2010). (Zhu *et al.*, 2009) proposed a statistical framework, called StatSnowball, to conduct both traditional information extraction and open IE. The framework employs discriminative Markov logic networks (MLNs) to learn the weights of relation extraction patterns, which are generally linguistic structure rules and keyword-matching rules. The framework is applied to an online open IE system called Renlifang, which was the first Chinese relation extraction service. Recently, (Etzioni *et al.*, 2011) advocated a second generation of open IE. The new research direction emphasizes the use of fine-grained linguistic analyses to improve the accuracy of open IE. For instance, (Fader *et al.*, 2011) observed that the outputs of most open IE systems are incoherent and uninformative. They proposed a second generation open IE approach called REVERB, which implements a relation phrase identifier based on syntactic and lexical constraints to improve the performance of open IE. Meanwhile, (Christensen *et al.*, 2010) observed that a large proportion of extraction errors made by open IE systems are related to incorrect or improper arguments. To address the problem, they proposed the R2A2 approach, which integrates an argument identifier with REVERB. The identifier utilizes syntactic patterns to correct the arguments of relation phrases extracted by REVERB.

Our method differs from traditional RE and existing open IE approaches in a number of respects. First, to the best of our knowledge, both traditional RE and existing open IE approaches detect static and permanent relations. By contrast, our method detects interactive segments and the interactions between persons are dynamic and topic-dependent. Second, in addition to syntactic features, we devise useful context-dependent and semantic features to detect interactive segments effectively. Finally, most approaches analyze the text between entities. For instance, (Culotta and Sorensen, 2004) constructs the dependency tree in accordance with the expression between an entity pair. Our method further considers the contexts before and after person names to enhance the relation detection performance.



### 3. Methodology

The system architecture of FISER is comprised of three key components, namely, *candidate segment generation*, *feature extraction*, and *interactive segment recognition*, as shown in Figure 3-1. At present, FISER is designed for Chinese topics. Candidate segment generation extracts important person names from a set of topic documents, and then partitions the documents into candidate segments that may contain information about the interactions between the topic persons. Next, the feature extraction component extracts representative text features from each candidate segment. The features are used by the interactive segment recognition component to classify interactive segments. We discuss each component in detail in the following sub-sections.

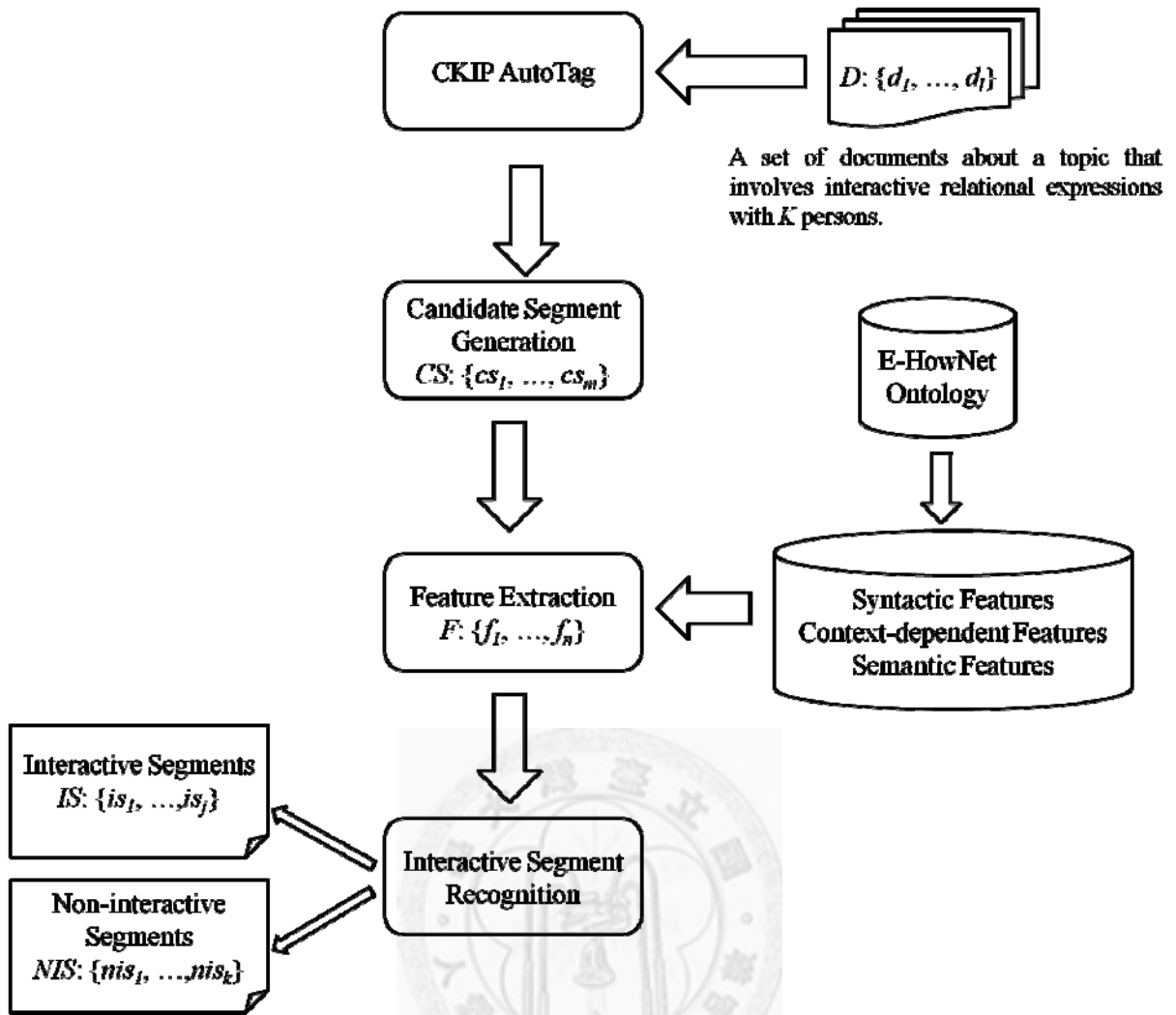


Figure 3-1. The system architecture

### 3.1 Candidate Segment Generation

Given a topic document  $d$ , we first apply the Chinese language parser CKIP AutoTag<sup>2</sup> to decompose the document into a sequence of sentences  $S = \{s_1, \dots, s_k\}$ . The parser also breaks a sentence into tokens and tags their parts-of-speech. It also labels the tokens that represent a person's name. In our experiment, we observed that many of the labeled person names rarely

<sup>2</sup> <http://ckipsvr.iis.sinica.edu.tw/>

occurred in the topic documents and the rank-frequency distribution of person names followed Zipf's law (Manning and Schütze, 1999). Low frequency names usually identify persons that are irrelevant to the topic. To discover the interactions between important topic persons, the low frequency person names are excluded. Let  $P = \{p_1, \dots, p_e\}$  denote the set of important topic person names. The sentence structure of Chinese is sophisticated and quite different from that of Western languages (Feng *et al.*, 2004). In Chinese, the main constituent of a sentence is a simple phrase (Wang *et al.*, 1998). Therefore, two or more consecutive sentences may express a coherent discourse; and an interactive segment may include a number of sentences. Figure 3-2 shows our candidate segment generation algorithm. Given a topic person name pair  $(p_i, p_j)$ , we consider two types of candidate segments: an *intra-sentential* segment in which the person names appear in the same sentence, and an *inter-sentential* segment in which the person names are distributed among consecutive sentences. The algorithm processes document sentences one by one and considers a sentence as the initial sentence of a candidate segment if it contains person name  $p_i$  ( $p_j$ ). Then, it examines the initial sentence and subsequent sentences until it reaches an end sentence that contains person name  $p_j$  ( $p_i$ ). If the initial sentence is identical to the end sentence, the algorithm generates an intra-sentential candidate segment; otherwise, it generates an inter-sentential candidate segment. However, if a period appears in the initial or end sentences, we drop the segment because a period indicates the end of a discourse in Chinese. In addition, if  $p_i$  ( $p_j$ ) appears more than once in a candidate segment, we truncate all the sentences before

the last  $p_i$  ( $p_j$ ) to make the candidate segment concise. By running all person name pairs over the topic documents, we obtain a candidate segment set  $CS = \{cs_1, \dots, cs_n\}$ .

### Candidate Segment Generation

**INPUT:**  $(p_i, p_j)$  – a topic person name pair;  $S = \{s_1, \dots, s_k\}$  – a sequence of sentences from a topic document  $d$ .

**BEGIN**

$inCandidate = false$

$cs = \{\}$

**FOR**  $l = 1$  **TO**  $l = k$

**IF**  $s_l$  contains  $p_i$  ( $p_j$ ) **&&**  $inCandidate == false$

add  $s_l$  into  $cs$

$inCandidate = true$

**ELSE IF**  $s_l$  contains  $p_i$  ( $p_j$ ) **&&**  $inCandidate == true$

$cs = \{\}$

add  $s_l$  into  $cs_n$

**ELSE IF**  $s_l$  contains  $p_j$  ( $p_i$ ) **&&**  $inCandidate == true$

add  $s_l$  into  $cs$

save  $cs$  into candidate segment set  $CS$

$inCandidate = false$

$cs_n = \{\}$

**ELSE IF**  $inCandidate == true$  **&&**  $s_l$  has a period

$cs = \{\}$

$inCandidate = false$

**END FOR**

**END**

Figure 3-2. Candidate segment generation algorithm

### 3.2 Interactive Segment Recognizer

To recognize interactive segments in  $CS$ , we treat interaction detection as a binary classification problem. In this work, we utilize the maximum entropy (ME) classification method (Berger *et al.*, 1996), which is a logistic regression-based statistical model. Let  $IS$  denote that a segment is interactive. ME classifies a candidate segment in terms of the following conditional probability:

$$p(IS|cs_l) = \frac{1}{Z(cs_l)} \exp \left( \sum_j w_j * f_j(IS, cs_l) \right), \quad (1)$$

$$Z(cs_l) = \exp \left( \sum_j w_j * f_j(IS, cs_l) + \sum_k w_k * f_k(\neg IS, cs_l) \right), \quad (2)$$

where  $f_j$  is a feature function and  $w_j$  is its weight. A feature function indicates a specific condition between  $IS$  and  $cs_l$ .  $Z(cs_l)$  is a smoothing factor that is used to normalize  $P(IS|cs_l)$  within the range  $[0,1]$ . Given a training dataset, the weights of the feature functions can be derived appropriately by the conditional maximum likelihood estimation method (Manning *et al.*, 2008). The learned weights are then used by Eq. 1 to detect interactive segments. ME has proven effective in many information extraction approaches (Chieu and Ng, 2002; Kambhatla, 2004). In addition, domain experts can design various feature functions to examine different characteristics of candidate segments. In the next section, we introduce the features used for interaction detection.



### 3.3 Feature Extraction

Generally, interactions between entities are described by verbs (Mitchell, 1997; Hatzivassiloglou and Weng, 2002), but not all verbs express interactions. For instance, the candidate segment  $cs_1$  in Figure 3-3 shown below contains more than one verb; however, the interaction between the given person names 胡錦濤(Hu Jintao) and 歐巴馬(Barack Obama) is not described by a verb. While the verb 審問(interrogated) in candidate segment  $cs_2$  in Figure. 3-3 indicates criticism between the given person names 馬英九(Ma Ying-Jeou) and 蔡英文(Tsai Ing-Wen), the segment contains another important topic person name, 宋楚瑜 (James Soong), who is irrelevant to the interaction. Because detecting interactions is so difficult, in addition to syntactic properties (e.g., verbs and their types), the semantic and context information about a candidate segment should be considered to ensure that interaction detection is successful. The following presents the proposed features including syntactic, context-dependent, and semantic information in text.

[ $cs_1$ ] 中國(Nc, China) 國家主席(Na, paramount leader) 胡錦濤(Nb, Hu Jintao) 將(D, will) 到(VCL, bound to) 美國(Nc, United States) 訪問(VC, visit) , (COMMACATEGORY) 將會(D, will) 有與(V\_2, have) 美國(Nc, United States) 總統(Na, president) 歐巴馬(Nb, Barack Obama) 見面(VA, meet)的機會(Na, opportunity) , (COMMACATEGORY)

[ $cs_2$ ] 馬英九(Nb, Ma Ying-Jeou) 總統(Na, president) 在(P, during) 辯論會(Na, debate) , (COMMACATEGORY) 不但(Cbb, not only) 反駁(VC, retorted) 宋楚瑜的(Nb, James Soong's) 質疑(VE, questions) , (COMMACATEGORY) 且(Cbb, but also) 審問(VC, interrogated) 宇昌案爭議(Nb, Yu-Chang controversy) , (COMMACATEGORY) 指責(VC, challenge) 蔡英文的(Nb, Tsai Ing-Wen's) 道德標準(Na, moral standards) 。(PERIODCATEGORY)

Figure 3-3. Examples of candidate segments

➤ **Syntactic Features**

- **VR:** The ratio of transitive verbs to intransitive verbs between the given person names in a candidate segment.
- **NV:** The number of verbs in a candidate segment.
- **NVE:** The number of verbs between the given person names in a candidate segment.
- **SL:** The length of a candidate segment (i.e., the number of tokens).
- **VLR:** The ratio of verbs to the length of a candidate segment.
- **SP:** It is equal to 1 if the punctuation { : ; , } appears in a candidate segment; otherwise, it is 0.
- **PD:** The number of tokens between the given person names of a candidate segment; that is, the distance of the given person names.
- **MP:** It is equal to 1 if person names other than the given person names occur in a candidate segment. For instance, MP is 1 for  $cs_2$ .
- **ICS:** It is equal to 1 if a candidate segment is intra-sentential; otherwise, it is 0. For instance, ICS is 0 for  $cs_2$ .
- **FPP:** The first position of the given person name in a candidate segment. For instance, FPP is 1 for  $cs_2$ .
- **LPP:** The last position of the given person name in a candidate segment. For instance, LPP is 16 for  $cs_2$ .

➤ **Context-dependent Features**

- **NVT**: The number of verbs in the tri-windows (i.e., three consecutive tokens) including the given person names. For instance, NVT is 1 for  $cs_2$ .
- **TNV**: It is equal to 1 if a candidate segment contains a verb on an interactive verb list; otherwise, it is 0. The verb list is compiled by using the log likelihood ratio (LLR) (Manning and Schütze, 1999), which is an effective feature selection method. Given a training dataset comprised of interactive and non-interactive segments, LLR calculates the likelihood that the occurrence of a verb in the interactive segments is not random. A verb with a large LLR value is closely associated with the interactive segments. We rank the verbs in the training dataset in terms of their LLR values, and select the top 150 verbs to compile the interactive verb list.
- **TNB**: It is equal to 1 if a candidate segment contains a bigram of an interactive bigram list; otherwise, it is 0. The bigram list is compiled in a similar way to the verb list by selecting the top 150 bigrams in the training dataset based on their LLR values.

## ➤ Semantic Features

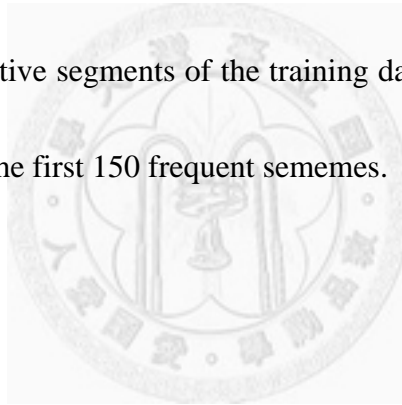
- **NPV:** This is the number of sentiment verbs in a candidate segment. Intuitively, interactions can occur with positive or negative semantics. For instance, the verb 審問 (interrogated) in *cs2* describes criticism between the given person names, and it is a sentiment verb with negative semantics. Here, we employ the NTU Chinese Sentiment Dictionary (NTUS)<sup>3</sup>, which contains 2812 positive and 8276 negative Chinese sentiment verbs compiled by linguistic experts.
- **NNA:** The number of negative adverbs (e.g., 未曾 (have never)) in a candidate segment.
- **PS:** It is equal to 1 if a sememe of a verb in a candidate segment is on an interactive sememe list; otherwise, it is 0. A sememe is a semantic primitive of a word defined by E-HowNet (Huang *et al.*, 2008), which is a Chinese lexicon compiled by Chinese linguistic experts. Basically, an interaction can be described by different synonyms. By considering the sememes of the verbs in a candidate segment, we may increase the chances of detecting interactions. For each sememe in E-HowNet, we compute its information gain (Han and Kamber, 2006) in discriminating the interactive and non-interactive segments of the training dataset. However, a sememe with a high information gain can be an indicator of non-interactive segments. Therefore, we process sememes one by one according to the order of their information gains. We compute the frequency that a sememe occurs in the interactive and non-interactive segments,

---

<sup>3</sup> <http://nlg18.csie.ntu.edu.tw:8080/opinion/pub1.html>

respectively. If the sememe tends to occur in the interactive segments, we regard it as an interactive sememe; otherwise, it is a non-interactive sememe. We compile the interactive sememe list by selecting the first 150 interactive sememes.

- **NS:** It is equal to 1 if a sememe of the verbs in a candidate segment is on a non-interactive sememe list; otherwise, it is 0. Similar to PS, the non-interactive sememe list is compiled by selecting the first 150 non-interactive sememes in the training dataset.
- **TVS:** It is equal to 1 if a sememe of the verbs in a candidate segment is on a frequent sememe list; otherwise, it is 0. We rank the sememes of verbs according to their occurrences in the interactive segments of the training dataset. The frequent sememe list is compiled by selecting the first 150 frequent sememes.



## 4. Performance Evaluation

In this section, we first present the evaluation data corpus. Then, we examine the performance of nineteen features on recognizing interactive relations. We examine the features one by one to combine the feature set that produces the best system performance. Next, we compare the proposed feature set with those of well-known open IE methods, including TEXTRUNNER, O-CRF and StatSnowball. Finally, we adopt lift curve to analyze the effectiveness of the features.

### 4.1 Data Corpus and Evaluation Metrics

In information extraction, evaluations are normally based on official corpora. Most previous information extraction studies used the Automatic Context Extraction (ACE) datasets<sup>4</sup> to evaluate system performance. However, the relations (e.g., *capital of*) defined in the datasets are static and therefore irrelevant to interactions between persons. To the best of our knowledge, there is no official corpus for interaction detection; therefore, we compiled our own data corpus for the performance evaluations.

---

<sup>4</sup> <http://www.itl.nist.gov/iad/mig/tests/ace/>

Table 4-1 shows the statistics of the data corpus which consists of ten topics related to political events in Taiwan from 2004 to 2010. The characteristics of the topics are shown in Table 4-2. Most people involved in the topics are associated with two major political parties in Taiwan, namely, the Democratic Progressive Party(DPP) and the KouMingTang(KMT). Topic *A~F* are about elections, where DPP and KMT competed for the president, mayors, legislators and city councilors. It is noteworthy that in Topic *F*, King Pu-tsung and Wang Yu-ting wrangled with each other. However, before the topic happened, they were friends and were members of KMT. Topic *G* is related to a policy that government tries to carry out a new health insurance. Topic *H* is about a protest in Taiwan that many Taiwanese protested against the visit of the China ARATS chief. Topic *I* is about scandals of corruptions and cost-overruns in the 2010 Taipei international flora exposition. Topic *J* is related to the former Taiwanese president Chen Shui-bian and his family. They involved in several corruptions and abuse of authority.

Table 4-1. The statistics of data corpus

# of topics	10
# of topic documents	500
# of tagged person names	436
# of evaluated person names	85
# of person name pairs	432
# of interactive segments (intra)	266
# of interactive segments (inter)	189
# of non-interactive segments (intra)	905
# of non-interactive segment (inter)	387

Each topic consists of 50 news documents (all longer than 250 words) downloaded from Google News. As mentioned in Sec. 3, many of the person names labeled by CKIP rarely occur in topic documents. Hence, for each topic, we selected the first frequent person names whose accumulated frequencies reached 70% of the total person name frequency count in the topic documents. In other words, the evaluated person names accounted for 70% of the person name occurrences in the examined topic and they were all representative topic persons. We extracted 1747 candidate segments from the topic documents by using the candidate segment generation algorithm. Then, we asked three experts to annotate the interactive segments. They labeled 455 segments as interactive, and the Kappa statistic of the labeling process was 0.615. The statistic indicates that our annotated data corpus is substantial. It is noteworthy that approximately 41% of the interactive segments are inter-sentential. In other words, expressions of person interactions often cross sentences. Meanwhile, 77% of the intra-sentential segments are non-interactive. That is, persons that appear in the same sentences are usually non-interactive, so considering the contexts before and after person names is necessary. Since interaction expresses are rare and sentence-crossing, detecting interactions is difficult.



Table 4-2. Ten political topics in Taiwan from 2004 to 2010

Topic ID	Topic Title	Time period
<i>A</i>	The legislative by-elections of Kaohsiung and Tainan city	2011/01~2011/03
<i>B</i>	2012 presidential election	2010/12~2011/03
<i>C</i>	2010 Mayoral, city councilor and borough chief elections in the five special municipalities	2010/01~2010/12
<i>D</i>	2008 presidential election	2008/01 ~2008/03
<i>E</i>	2004 presidential election and the county magistrate by-election of Hualien	2003/12 ~2004/03
<i>F</i>	The disputation of King Pu-tsung and Wang Yu-ting	2011/01~2011/03
<i>G</i>	Yaung Chih-Liang with health insurance problems and vaccination	2009/08~2010/03
<i>H</i>	ARATS chief came to Taiwan	2008/10~2008/11
<i>I</i>	Taipei international flora exposition and scandal of Maokong Gondola	2007/07~2008/12
<i>J</i>	The corruption case of the former Taiwanese president Chen Shui-bian	2006/11~2011/08

We use 10-fold cross validation (Kohavi, 1995) to examine the effects of the proposed features. For each evaluation run, a topic and the corresponding candidate segments are selected as test data, and the remaining topics are used to train FISER. The results of the 10 evaluation runs are averaged to obtain the global performance. The evaluation metrics are the precision rate, recall rate, and F1-score (Manning *et al.*, 2008). We use F1 to determine the superiority of the features because it balances the precision and recall scores (Manning *et al.*, 2008).

## 4.2 The Performance of the Features

Table 4-3 shows the performance of the syntactic, context-dependent, semantic, and all nineteen features, denoted as  $FISER_{\text{syntactic}}$ ,  $FISER_{\text{context-dependent}}$ ,  $FISER_{\text{semantic}}$ , and  $FISER_{\text{all}}$ , respectively. As shown in the table, the syntactic features cannot detect interactive segments correctly. This is because they are incapable of discriminating between interactive segments. Since both interactive and non-interactive segments are comprised of grammatical sentences, they have similar syntactic feature values. For instance, the averages of VR for the interactive and non-interactive segments are 2.11 and 2.33 respectively; and there is no significant difference in terms of t-test with a 95% confidence level. Besides, as mentioned in Sec. 4.1, both intra-sentential and inter-sentential segments can be non-interactive. Therefore, the syntactic ICS and SP features, which are used to judge inter-sentential segments, are indiscriminative. As a result, the syntactic features are ineffective in detecting interactions. By contrast, the context-dependent features detect interactive segments successfully. We observe that the compiled interactive verb list and interactive bigram list are closely associated with person interactions, so the TNV and TNB features discriminate interactive segments effectively. Meanwhile, the verbs used to describe person interactions tend to occur immediately before or after the given person names. Thus, the context-dependent NVT feature is useful for filtering out non-interactive segments. It is noteworthy that the semantic features produce a high precision rate, but a low recall rate. Our analysis of the experimental data showed that segments containing positive or negative verbs generally reveal person

interactions; hence, the semantic feature NPV yields high detection precision. However, a significant proportion of the interactive segments do not have sentimental semantics, so the feature cannot increase the detection recall rate. While the semantic features PS, NS, and TVS try to increase the detection of interactive segments by considering the sememes of verbs, the expert-compiled E-HowNet is not comprehensive enough to identify various person interactions. Notably, FISER achieves its best performance when all the features are applied together (denoted as FISER<sub>all</sub>). In other words, the context-dependent features and semantic features do not conflict with each other. As the features detect interactive segments from different perspectives, applying them together improves the system's performance.

Table 4-3. Experimental result of each feature category

<b>Feature Category</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>
FISER <sub>syntactic</sub>	33.8%	2.3%	38.4%
FISER <sub>context-dependent</sub>	66.3%	41.5%	49.3%
FISER <sub>semantic</sub>	64.2%	16.9%	25.8%
FISER <sub>all</sub>	70.2%	54.8%	60.7%

### 4.3 The Best Combination of the Features

Table 4-4 shows the effects of the features iteration by iteration. In the first iteration, we examine the performance of each feature and list the performance of the feature that has the best F1-score in the first row of Table4-4. In the  $i$ th iteration ( $2 \leq i \leq 19$ ), the set of features selected in the  $(i-1)$ th iteration serves as the basis. Next, we examine each remaining feature combined with the basis and show the performance that has the best F1-Score in the  $i$ th row. For example, the fifth row in Table 4-4 shows the result of the top-5 effective features. Moreover, a one-tail paired t-test is applied to determine whether combining each feature with the basis improves the performance significantly. The symbol ‘\*’ means the combination improves the system performance significantly; the symbol ‘#’ means the combination decreases the performance; and, the symbol ‘-’ in the table means the performance does not change. As shown in the table, the best combination of the features is {TNB, PS, NS, TNV,SL, NV, NPV, FPP, PD, ICS} and its precision, recall, and F1-Score are 72.6%, 55.6%, and 61.9%, respectively. In the following section, we use this result to compare with other open IE systems.

It is noteworthy that TNB has the best performance of all nineteen features and it is discriminative for relation detection. The ratios of TNB in interactive segments and in non-interactive segments are 44.2% and 7.4%, respectively, and have statistically significant difference at the 99% confidence level based on z-test. PS and NS use the interactive sememe list and the non-interactive sememe list respectively to determine whether a candidate

segment is interactive or not. Similarly, the ratios of PS and NS in interactive segments and in non-interactive segments have significant difference at the 99% confidence level based on z-test. As the two features discriminate interactive segment effectively, using them increases the precision and F1-Score dramatically. SL and NV decrease the precision but improve the recall significantly and they improve the overall performance. The best combination of the features includes the syntactic, context-dependent, and semantic features. This result is highly consistent with the results shown in Table 4-3. Previous relation detection studies mainly focus on syntactic features. However, using syntactic information of text is not good enough to detect interactive segments successfully. In addition to syntactic information, we consider semantic and context information of text to improve the system's performance.

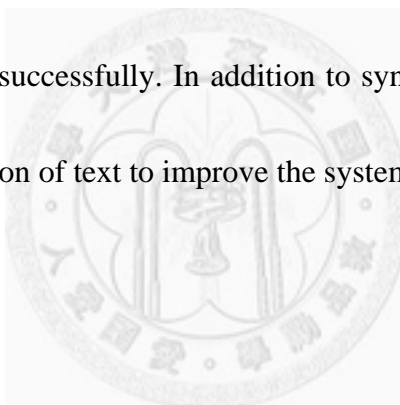


Table 4-4. The effect of features

Features	Precision	Recall	F1-Score
TNB	66.3%	41.5%	49.3%
+PS	68.4% (**, **)	45.5% (***, ***)	53.0% (***, ***)
+NS	77.4% (***, ***)	45.5% (-, ***)	55.3% (***, ***)
+TNV	77.4% (-, ***)	45.5% (-, ***)	55.3% (-, ***)
+SL	73.3% (#, ***)	50.6% (**, **)	58.9% (*, ***)
+NV	70.1% (##, *)	54.2% (**, ***)	60.0% (-, ***)
+NPV	75.9% (-, ***)	52.9% (-, ***)	60.6% (-, ***)
+FPP	68.3% (##, *)	56.2% (-, ***)	61.1% (-, ***)
+PD	72.1% (*, **)	54.5% (-, ***)	61.4% (-, ***)
+ICS	72.6% (-, **)	55.6% (-, ***)	61.9% (-, ***)
+NVE	72.6% (-, **)	55.6% (-, ***)	61.9% (-, ***)
+VLR	72.6% (-, **)	55.6% (-, ***)	61.9% (-, ***)
+MP	72.6% (-, **)	55.6% (-, ***)	61.9% (-, ***)
+LLP	72.6% (-, **)	55.6% (-, ***)	61.9% (-, ***)
+SP	72.7% (-, **)	54.6% (-, ***)	61.4% (-, ***)
+NNA	72.0% (-, **)	54.8% (-, ***)	61.3% (-, ***)
+NVT	72.2% (-, **)	54.9% (-, ***)	61.4% (-, ***)
+VR	72.4% (-, **)	54.2% (-, ***)	61.0% (-, ***)
+TVS	70.2%8 (-, *)	54.8% (-, ***)	60.7% (-, ***)

1. \*, \*\*, and \*\*\* represent right-tail paired t-tests with  $\alpha = 0.1, 0.05, \text{ and } 0.01$ , respectively
2. #, ##, and ### represent left-tail paired t-tests with  $\alpha = 0.1, 0.05, \text{ and } 0.01$ , respectively
3. (comparison with  $(i-1)$ th iteration, comparison with the first iteration)

## 4.4 Comparison with Open IE Methods

We compare the proposed feature set with those of three well-known open IE methods, namely, TEXTRUNNER (Banko *et al.*, 2007), O-CRF (Banko and Etzioni, 2008), and StatSnowball (Zhu *et al.*, 2009). We conduct two types of evaluation on the compared methods. The first type compares the original methods to examine whether the compared methods can detect interactive segments successfully. The other type compares the feature sets of the compared methods. For each compared method, we use its feature sets to train a ME classifier, to evaluate whether the features can detect interactive segments precisely. To ensure the comparison is fair, an ME classifier is trained for each feature set and 10-fold cross validation is used to obtain its global performance.

TEXTRUNNER, the first open IE system, employs six syntactic features to extract the relations between entities. The features includes the part-of-speech tag sequences between an entity pair, the number of tokens and the number of stop-word between an entity pair, whether an object is a proper noun or not, and the part-of-speech tag to the left of target entity and to the right of candidate entity.

O-CRF considers syntactic features, including POS tags and regular expressions of syntax (e.g., detecting capitalization, punctuation, etc.). It also uses context words to identify relation keywords between entities. The context words consist of the word sequence between an entity pair, the conjunctions occurred in the adjacent six words left or right to the current token, and the punctuations between an entity pair.

StatSnowball also adopts syntactic features to identify relation keywords between entities. The selected features include POS tags and occurrences of nonstop words. Originally, the features of StatSnowball focus on nouns, but interactions between persons are usually described by verbs. To have a fair comparison, we adjust the features related to nouns to verbs. For example, a feature of StatSnowball is to determine whether the previous token of a target entity is a noun. We change it to examine whether the previous token of a target entity is a verb. Specifically, the features of StatSnowball examine whether the verbs between an entity pair are not stop words and their occurrences are more than a pre-defined threshold, whether the previous token of a target entity is a verb, and whether the following token of a candidate entity is a verb.

The features of the compared methods are listed in Table 4-5, and we evaluate the open IE methods on the feature sets. Notably, O-CRF and StatSnowball, which are designed for relation extraction, extract interaction keywords from a candidate segment in our experiment. Hence, a candidate segment is classified as non-interactive if no interactive keyword is extracted from it.



Table 4-5. The features of the comparative methods

Method	Description of features
TEXTRUNNER <sub>F</sub>	F <sub>1</sub> : the presence of part-of-speech tag sequences between entity pair.
	F <sub>2</sub> : the number of tokens between entity pair.
	F <sub>3</sub> : the number of stop-words between entity pair.
	F <sub>4</sub> : whether or not an object $e$ is found to be a proper noun.
	F <sub>5</sub> : the part-of-speech tag to the left of target entity $e_t$ .
	F <sub>6</sub> : the part-of-speech tag to the right of candidate entity $e_c$ .
O-CRF <sub>F</sub>	F <sub>1</sub> : the part-of-speech tags sequence between entity pair.
	F <sub>2</sub> : is there any punctuation between entity pair.
	F <sub>3</sub> : context words sequence between entity pair.
	F <sub>4</sub> : conjunctions of features occurring in adjacent positions within six words to the left and six words to the right of the current token.
StatSnowball <sub>F</sub>	F <sub>1</sub> : verbs between entity pair are all not stop word and occur more than MIN_OCCUR times.
	F <sub>2</sub> : the previous token of target entity $e_t$ is verb.
	F <sub>3</sub> : the following token of candidate entity $e_c$ is verb.

The performance results of the compared features and methods are shown in Table 4-6.

The results of TEXTRUNNER<sub>F</sub>, O-CRF<sub>F</sub>, and StatSnowball<sub>F</sub> denote the performance of the compared feature sets by training ME classifiers, and the result of TEXTRUNNER, O-CRF, and StatSnowball denote the performance of the original systems with our corpus. As shown in the table, FISER outperforms all the compared methods and feature sets. As the compared methods and feature sets simply use syntactic features, they cannot sense the semantics of person interactions in candidate segments successfully. By contrast, FISER incorporates semantic and context-dependent features, and thus achieves the best precision, recall, and F1 score. O-CRF outperforms StatSnowball and TEXTRUNNER because its feature set

considers the context information of a candidate segment. It is interesting to note that O-CRF and StatSnowball are inferior to O-CRF<sub>F</sub> and StatSnowball<sub>F</sub>, and the recall rates of O-CRF and StatSnowball are very low. Basically, O-CRF and StatSnowball employ the CRF model to learn the extraction patterns of interaction keywords. Since the non-interactive segments have no interaction keywords, only the interactive segments of the training data are useful for pattern learning. As shown in Table 4-1, most of the candidate segments are non-interactive. Thus, the learned extraction patterns cannot detect interactive segments completely, and the recall rates of the methods deteriorate. The outcome corresponds well with the observation in (Li *et al.*, 2008) that detecting relation segments is necessary to ensure that extractions of relation keywords are reliable.

Table 4-6. The interaction detection result of compared methods

<b>Features</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>
TEXTRUNNER	32.7%	2.2%	4.0%
O-CRF	42.1%	8.8%	14.6%
StatSnowball	48.1%	5.5%	9.9%
TEXTRUNNER <sub>F</sub>	48.8%	34.3%	38.9%
O-CRF <sub>F</sub>	53.2%	39.8%	43.5%
StatSnowball <sub>F</sub>	52.6%	25.2%	32.1%
FISER <sub>all</sub>	70.2%	54.8%	60.7%
FISER <sub>best</sub>	72.6%	55.6%	61.9%

Based on the experimental results, we conclude that syntactic features cannot detect interactive segments correctly. Existing open IE studies focus on discovering static and permanent relations between entities. Hence, the syntactic features of the text in entities are useful. In (Banko and Etzioni, 2008), Banko and Etzioni claim that 86% of relation expressions are in the given entities. However, according to our data corpus, only 56% of the interaction expressions are in the given person names in Chinese. In addition, Chinese sentences are complex and contain many unknown words (Ling *et al.*, 2003) that affect the correctness of the syntactic features used by the compared methods. Therefore, the compared methods are inferior in terms of detecting interactive segments. To sum up, FISER employs effective features that cover syntactic, context-dependent, and semantic information of text to detect interactive segments in topic documents successfully. Because FISER filters out non-interactive segments and discriminates between interactive segments effectively, it outperforms the compared methods.

## 4.5 The Effectiveness of the Features

In this section, we employ lift analysis (Ling and Li, 1998) to visualize the effectiveness of each feature in detecting interactive segments. Lift analysis is a popular method for measuring the effectiveness of data mining components (Jindal and Liu, 2008). To examine the effectiveness of each feature, first, we divide the corpus into training and test sets. Next, the training instances are used to build a classifier based on each feature, and the classifier computes a margin score for each test instance, after which the test instances are sorted in descending order based on the margin score and divided into 10 equal deciles. Finally, a lift curve of the feature is constructed by computing the cumulative gain ( $CG$ ) for each deciles as follow:

$$CG_j = \frac{\text{the camulated number of interactive segments at decile } j}{\text{the total number of interactive segments}} \times 100\% \quad (3)$$

where  $j$  is the index of deciles and  $0 \leq j \leq 1$ .  $CG_j$  represents the cumulative percentage of interactive segments from the first decile to decile  $j$ . A lift curve is always from the left-hand corner to the right-hand corner. If a feature is discriminative for determining interactive segments, its lift curve will climb steeply on the left-hand side and the area under the curve will be large. On the other hand, indiscriminative feature will distribute the interactive segments randomly over the deciles, so its curve will be diagonal.

The lift curves of syntactic features are around the diagonal line shown as Figure 4-1, and this result is consistent with the result shown in Table 4-3 that  $FISER_{\text{syntactic}}$  is not enough to detect interactive segments. That is because both interactive and non-interactive segments are constructed by similar grammatical rules that make the syntactic features ineffective. Previous RE studies show that syntactic features are useful in detecting static and permanent relations between entities. However, when dealing with dynamic and topic-dependent interactive relations, we have to consider other useful text information, such as context-dependent and semantic information, to detect interactive segments accurately.

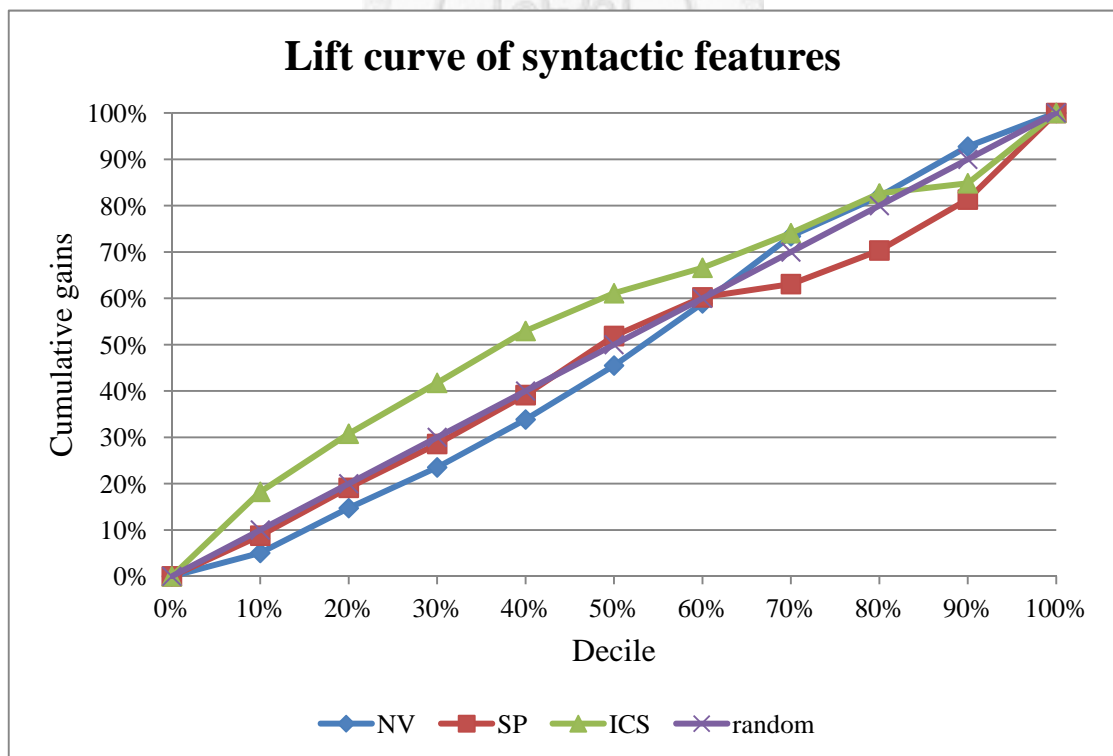


Figure 4-1. The lift curve of syntactic features

As shown in Figure 4-2, the lift curves of TNB and TNV are above the diagonal line and they discriminate interactive and non-interactive segments successfully. These two features consider the interactive verb list and bigram list consisting of verbs and bigrams which are highly associated with interaction. Hence, they filter non-interactive segments effectively. By contrast, the curve of NVT is around the diagonal line. The feature counts the number of verbs in the tri-windows before and after the given person names. Since both interactive and non-interactive segments have the similar feature values, the feature is ineffective in discriminating interactive and non-interactive segments. According to Figure 4-2 and Table 4-3, the context-dependent features are useful to identify interactive segments. Hence, using it improves the system's performance.

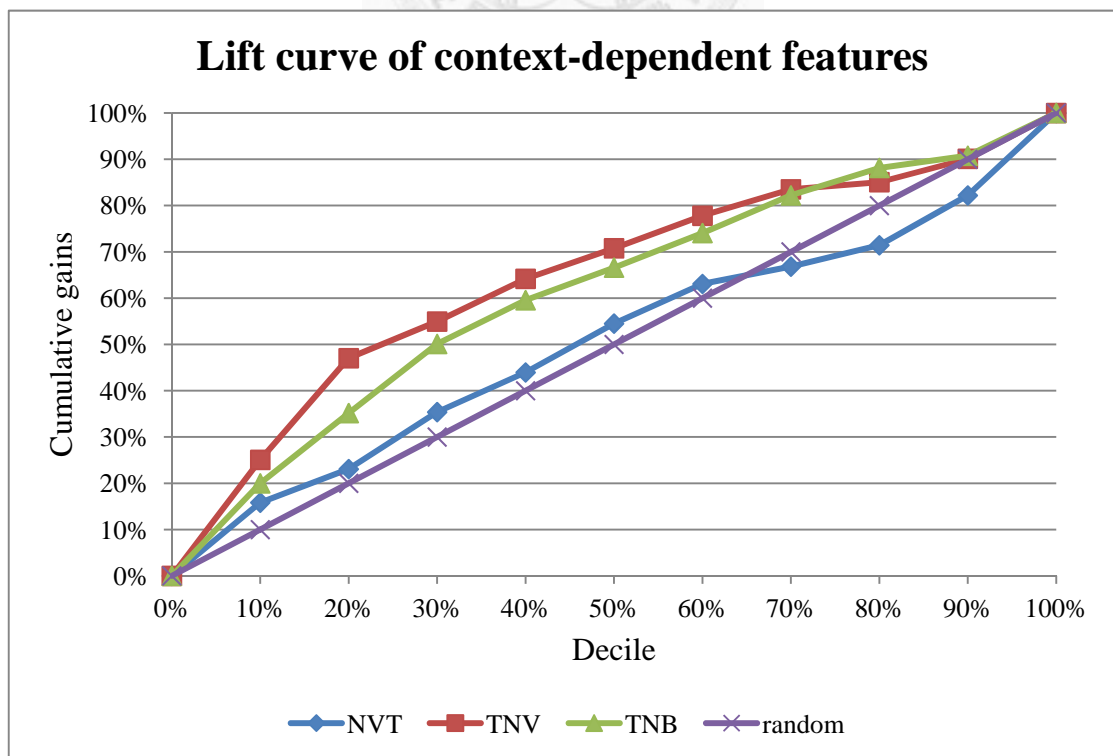


Figure 4-2. The lift curve of context-dependent features

Figure 4-3 shows that NPV is an effective semantic feature in detecting interactive segments. According to the definition of interaction we mentioned in the introduction section, interactions usually convey sentiments of people. Since NPV counts the number of sentiment verbs in a candidate segment, it is useful in detecting interactive segments. The result is consistent with the explanation of NPV in the sec. 4-2. As shown in Figure 4-3, the lift curve of NS is under the diagonal line. In other words, it is still an effective feature in identifying non-interactive segment. This is because NS examines the non-interactive sememe list. Hence, it is an effective indicator to filter non-interactive segments.

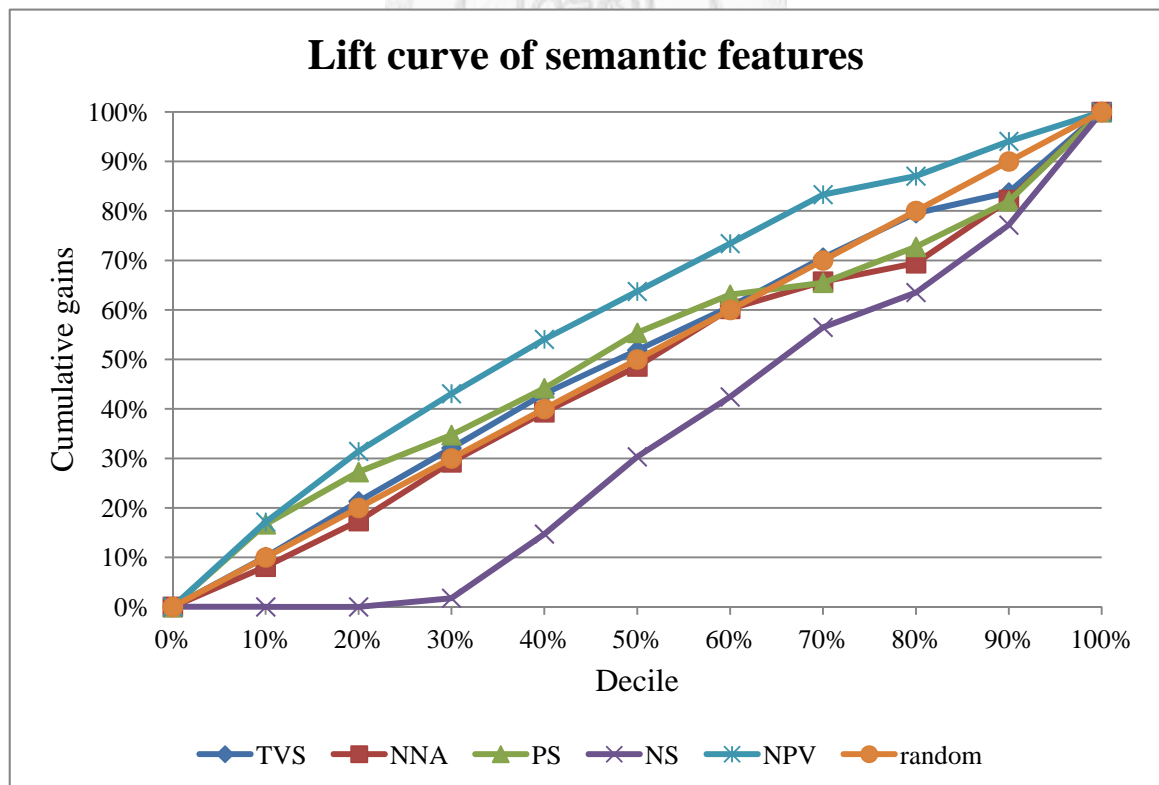


Figure 4-3. The lift curve of semantic features

## 5. Conclusion

A topic is associated with specific times, places, and persons. Discovering the interactions between the persons would help readers construct the background of the topic and facilitate document comprehension. In this paper, we have proposed an interaction detection method called FISER, which employs nineteen features covering syntactic, context-dependent, and semantic information in text to detect interactive segments in topic documents.

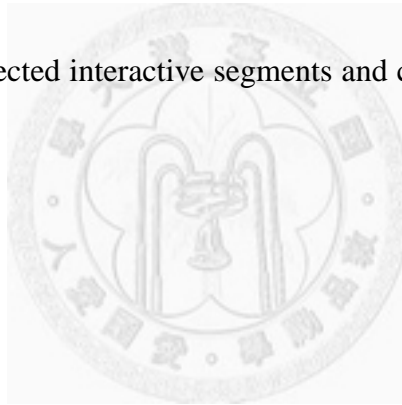
Our method differs from previous relation detection studies in three respects. First, instead of detecting static and permanent relations, our method detects interactive segments and the interactions between persons are dynamic and topic-dependent. Second, in addition to syntactic features, we devise useful context-dependent and semantic features to detect interactive segments effectively. Finally, most previous approaches analyze the text between entities, but our method further considers the contexts before and after person names to enhance the relation detection performance.

We present an effective recognizer which consider syntactic, context-dependent and semantic information for detecting topic-dependent interactive relation. The experiment results demonstrate the efficacy of FISER which outperforms well-known open IE methods



dramatically. By using all nineteen features, the precision, recall, and F1-Score are 70.2%, 54.8%, and 60.7%, respectively; Meanwhile, the best combination of the features is {TNB, PS, NS, TNV,SL, NV, NPV, FPP, PD, ICS} and its precision, recall, and F1-Score are 72.6%, 55.6%, and 61.9%, respectively.

In the future work, we will employ sophisticated syntactic features, such as the dependency tree of a sentence, to enhance FISER's syntactic features. Moreover, external knowledge bases will be incorporated into E-HowNet to increase the detection of interactive segments. We will also investigate using information extraction algorithms to extract interaction tuples from the detected interactive segments and construct an interaction network of topic persons.



## References

Agichtein, Eugene and Gravano, Luis, " Snowball: extracting relations from large plain-text collections," In *Proceedings of the 5th ACM conference on Digital libraries*, 85-94, (2000).

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

Banko, Michele and Etzioni, Oren, "The tradeoffs between open and traditional relation extraction," In *Proceedings of the 46th Annual Meeting on Association for Computational Linguistics on Human Language Technologies*, 28-36, (2008).

Berger, Adam L., Pietra, Vincent J. Della and Pietra, Stephen A. Della, "A maximum entropy approach to natural language processing," *Comput. Linguist.*, **22**, 39-71, (1996).

Chen, Chien Chin and Chen, Meng Chang, "TSCAN: A content anatomy approach to temporal topic summarization," *IEEE Transactions on Knowledge and Data Engineering*, **24**, 170-183, (2012).

Chieu, Hai and Ng, Hwee, "A maximum entropy approach to information extraction from semi-structured and free text," In *Proceedings of the 18th National Conference on Artificial intelligence*, 786-791, (2002).

Christensen, Janara, Mausam, Soderland, Stephen and Etzioni, Oren, "Semantic role labeling for open information extraction," In *Proceedings of the NAACL HLT 2010 1st International Workshop on Formalisms and Methodology for Learning by Reading*, 52-60, (2010).

Culotta, Aron and Sorensen, Jeffrey, "Dependency tree kernels for relation extraction," In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 423-429, (2004).

Etzioni, Oren, Fader, Anthony, Christensen, Janara, Soderland, Stephen and Mausam, "Open information extraction: the second generation," In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 3-10, (2011).

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

Feng, Ao and Allan, James, "Finding and linking incidents in news," In *Proceedings of the 16th ACM International Conference on Information and Knowledge Management*, 821-830, (2007).

Feng, Haodi, Chen, Kang, Deng, Xiaotie and Zheng, Weimin, "Accessor variety criteria for Chinese word extraction," *Comput. Linguist.*, **30**, 75-93, (2004).

Han, Jiawei, and Kamber, Micheline, *Data mining Concepts and Techniques*: Morgan Kaufmann Publishers, 2nd edn., 2006.

Hatzivassiloglou, Vasileios and Weng, Wubin, "Learning anchor verbs for biological interaction patterns from published text articles," *International Journal of Medical Informatics*, **67**, 19-23, (2002).

Hirano, Toru, Asano, Hisako, Matsuo, Yoshihiro and Kikui, Genichiro, "Recognizing relation expression between named entities based on inherent and context-dependent features of relational words," In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 409-417, (2010).

Hirano, Toru, Matsuo, Yoshihiro and Kikui, Genichiro, "Detecting semantic relations between named entities in text using contextual features," In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 157-160, (2007).

Huang, Shu-Ling, Chung, You-Shan and Chen, Keh-Jiann, "E-HowNet: the expansion of HowNet," In *Proceedings of the 1st National HowNet Workshop*, 10-22, (2008).

Jindal, Nitin and Liu, Bing, "Opinion spam and analysis," In *Proceedings of the international conference on Web search and web data mining*, 219-230, (2008)

Kambhatla, Nanda, "Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations," In *Proceedings of the 42nd Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 178-181, (2004).

Kohavi, Ron, "A study of cross-validation and bootstrap for accuracy estimation and model selection," In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1137-1143, (1995).

Lafferty, John D., McCallum, Andrew and Pereira, Fernando C. N., "Conditional random fields: probabilistic models for segmenting and labeling sequence data," In *Proceedings of the 18th International Conference on Machine Learning*, 282-289, (2001).

Li, Wenjie, Zhang, Peng, Wei, Furu, Hou, Yuexian and Lu, Qin, "A novel feature-based approach to Chinese entity relation extraction," In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, 89-92, (2008).

Ling, Charles and Li, Chenghui, "Data Mining for Direct Marketing: Problems and Solutions," In *Knowledge Discovery and Data Mining*, 73-79, (1998).

Ling, Goh Chooi, Asahara, Masayuki and Matsumoto, Yuji, "Chinese unknown word identification using character-based tagging and chunking," In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 197-200, (2003).

Manning, Chris and Schütze, Hinrich, *Foundations of statistical natural language processing*: MIT Press, Cambridge, Massachusetts, 1st edn., 1999.

Manning, Christopher D., Raghavan, Prabhakar and Schütze, Hinrich, *Introduction to information retrieval*: Cambridge University Press, Cambridge, U.K, 2nd edn., 2008.

Mitchell, T.M., *Machine learning*: McGraw-Hill, New York, 1st edn., 1997.

Nallapati, Ramesh, Feng, Ao, Peng, Fuchun and Allan, James, "Event threading within news topics," In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, 446-453, (2004).

Pantel, Patrick and Pennacchiotti, Marco, "Espresso: leveraging generic patterns for automatically harvesting semantic relations," In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 113-120, (2006).

Vernon, G.M., *Human interaction: an introduction to sociology*: Ronald Press Co., New York, 1st edn., 1965.

Wang, Yuan-Kai, Chen, Yi-Shiou, and Hsu, Wen-Lian, "Empirical study of Mandarin Chinese discourse analysis: an event-based approach," In *Proceedings of 10th IEEE International Conference on Tools with Artificial Intelligence*, 466-473, (1998).

Zelenko, Dmitry, Aone, Chinatsu and Richardella, Anthony, " Kernel methods for relation extraction," *The Journal of Machine Learning Research*, **3**, 1083-1106, (2003).

Zhou, Guodong, Qian, Longhua and Fan, Jianxi, "Tree kernel-based semantic relation extraction with rich syntactic and semantic information," *Information Sciences*, **180**, 1313-1325, (2010).

Zhu, Jun, Nie, Zaiqing, Liu, Xiaojiang, Zhang, Bo and Wen, Ji-Rong, "StatSnowball: a statistical approach to extracting entity relationships," In *Proceedings of the 18th International Conference on World Wide Web*, 101-110, (2009).