國立交通大學

資訊科學與工程研究所

碩 士 論 文

從搜尋結果進行人際關係的辨識

Identifying Human Relationship From Retrieved Snippets

研 究 生：聶家祺

指導教授：梁婷　教授

中 華 民 國　一百 年 九 月

Identifying Human Relationship From Retrieved Snippets

研 究 生：聶家祺　　Student：Chia-Chi Nieh

指導教授：梁婷　　Advisor：Tyne Liang

國 立 交 通 大 學
資 訊 科 學 與 工 程 研 究 所
碩 士 論 文

A Thesis
Submitted to Institute of Computer Science and Engineering
College of Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Computer Science

June 2008

Hsinchu, Taiwan, Republic of China

中華民國一百年九月

從搜尋結果進行人際關係的辨識

研究生：聶家祺　　指導教授：梁婷博士

國立交通大學資訊科學與工程研究所

摘要

實體間的關係辨識一直是篇章處理中的重要工作。目前所辨識的關係，有人物與組織間的工作關係、疾病和藥的關係、作者與作品的關係、蛋白質間的交互關係或是名詞間的等價關係。所使用的方法多以學習模組或樣本分析進行辨識；少部分則是利用剖析樹從句法結構中來辨識目標關係。基本上這些方法所使用的語料可分為固定的語料及動態更新的語料（如網路搜尋結果）。雖然從固定語料辨識關係可獲得較高的正確率，然而透過搜尋引擎的搜尋結果可以得到較新的資訊。在本篇論文中，我們考量人際關係常有更新，因此在搜尋引擎結果中辨識人際關係。因此我們利用 Wikipedia 建置開發語料，整理出親屬關係及工作關係的關係樣板。此外，為辨識每個人物實體所對應的領域及領域詞彙，我們利用 bootstrapping 方式從開發語料中抽取出線索詞，用以擴充查詢詞，以擷取出相關的搜尋結果。為了加速篇章處理，我們採用簡單的人名及詞性標記，並進行人稱代詞的消解。我們提出兩階段的辨識程序，第一階段透過比對樣板，第二階段從支援向量機(support vector machine, SVM)透過抽取 7 種特徵進行辨識。特徵包括線索詞的數量與位置、人物的 mutual information、及實體間的相似度。最後所提的方法在 396 個親屬關係案例的實驗的 F-score 可達到 0.86；在 175 個工作關係案例中的 F-score 則有 0.75。

Identify Human Relationship From Retrieved Snippets

Student:Jia-Chia Nieh Professor:Tyne Liang
Institute of Computer Science and Engineering
National Chiao Tung University

ABSTRACT

Identifying relation among entities is an important task in document processing. The relations identified in previous researches include co-working relations between persons and organizations, relations among diseases and medicines, relations between authors and artifacts, the interactions between proteins, and the equivalence relations among nominals etc... Most identification methods are based on machine learning algorithms or pattern matching and few are based on parsing result. Besides, the corpora used for relation identification can be static and dynamic (like search engine results). Although identifying relations from static corpus generally outperforms the methods using dynamic corpora, yet dynamic corpora contain more updated information. In this thesis, we employ retrieved snippets to identify human relationships and Wikipedia to construct developing corpus. We extract domain words from developing corpus by the bootstrapping algorithm and expand queries for accurate search results. To speed up document processing, simple methods are implemented for part-of-speech tagging, person name tagging and pronominal anaphor resolution. The proposed kinship identification is implemented by pattern matching and support vector machine (SVM). The Features to be used at identification includes the amount and position of clue words and cosine similarity of entities related to persons. The kinship identifier yields 0.86 f-score in the experiment containing 396 kinship instances and the co-working identifier

yields 0.75 f-score on 175 co-working instances.

致 謝

能夠完成這篇論文，首先要感謝我的指導教授梁婷教授，感謝老師對於碩士生的要求，並且不因為我的英文程度太差就放棄我，讓我這幾年沒有白過，有一些確實的收穫。另外也感謝我的口試委員們，葉慶龍教授、楊武教授，認真的給我論文上的建議。同時也感謝實驗室的學長們，感謝陳冠熙學長、王笙權學弟、邱泓達學弟、和李奕賢學弟不時給給我鼓勵，也陪我度過難熬的日子。最後也感謝我的家人，不因為我的進度緩慢舊則罰我，仍然讓我讀完碩士。感謝我的女朋友，在技術和智慧上都鼓勵我幫助我，也感謝神，若沒有祂的帶領我一定沒辦法撐下來。因此，僅將這篇論文獻給我所珍重的人。

Content

List of Tables

List of Figures

Chapter 1 Introduction

Relation identification is a kernel task in message understanding. The relations to be identified can be equivalence relations between nominals, relation between proteins [Rosario and Hearst, 2005], the semantic relations among nominals such as ‘is-a’ and ‘part-of-whole’ [O’Hara and Wiebe, ‘07], and the relations between authors and artifacts [Wang et al. ‘07]. In this thesis our target relation are human relationships, namely kinship and co-working relationship. Kinship relation is based on consanguinity and marriage. The co-working relationship is defined that two persons work in the same organizations.

As World Wide Web becomes popular, the information on the web has become a huge corpus for natural language processing researchers. The performance of identifying human relationship depends on the quality of training corpus. Using Google search engine can collect a large number of news, blog articles, forum articles, and home page information. The web corpus is updated rapidly so we can use its benefit to identify new human relationships.

The proposed identification contains background task and the foreground task. The background task aims to build a developing corpus for extracting kinship patterns and domain words. The developing corpus is constructed by the manual acquisition with the employment of Wikipedia. We extract syntactic patterns from developing corpus by computing frequency of clue words defined in [Tian and Qian ‘10] and part-of-speech based sequence. We also use a bootstrapping algorithm to compute the domain words by using given seeds of domains. In this thesis the domains are politic, commercial, sport, entertainment and education. In foreground processing, we use the acquired clue words and domain words to expand queries. From retrieved snippets return by Google, we

identify domain of each person occurring in snippets by domain words and domains of co-occurrence persons. The kinship identification is based on pattern matching and SVM model. There are seven features used for SVM-model, they are the average length between persons co-occurring in sentences, the amount and position of clue words. Similarly, the features used for co-working identification are the average length between persons and similarity of between related entities of persons. In experiment, kinship identifier yield 0.86 f-score and co-working identifier yield 0.75.

While using search engine as research corpus, we encounter the problems. For example, different persons have the same names, persons have several nicknames If persons have the same name; we need to distinguish their domains. We collect a domain word sets from developing corpus, and we can classify person's domain by computing word count, organization, and person frequency in the sentence.

We are also confronted with the problem of recognizing nicknames of a person, for example, "陳水扁" called "陳總統" and "陳先生". In this thesis we use a stack to store ten last names processed by the proposed system. Only complete names are stored in the stack; if we want to process a person name, we try to identify whether the name is in the stack.

The remaining parts of the thesis are described as follows. Chapter 2 is introduces the related work addressing research corpus, target relations to be identified and the identification. Chapter 3 describes the proposed method containing developing corpus construction, sentences preprocessing, the pattern extraction and the SVM-based identification. Chapter 4 describes the experiment settings and results. The conclusions and future works are in chapter 5.

Chapter 2 Related Work

In this chapter, the related work is discussed on three issues: corpus resource, target relations and identification methods.

2.1 Corpus Resource

Wikipedia has become the largest online encyclopedia in resent year. There are 375808 entries available in traditional Chinese Wikipedia in the year 2011. Gang Wang et al.[‘07] proposed a method to extract the relations from Wikipedia. They retrieve entry title, categories of entry, hyperlinks, and infobox as resource for extracting relations between authors and artifacts. The features are selected from context, categories and infobox. 10,000 XML pages are selected as experiment database, and 130,000 entities pairs are extracted.

Dat P.T. et al. [‘07] provided a relation extraction method based on Wikipedia. This paper points out that treating web document structure and syntactic subtree mining as SVM feature. There are 13 relations types extracted from Wikipedia context. They used 5975 articles from Wikipedia as experiment data and out of them 45 articles of experiment data are used as testing data. Total 39467 entities are collected from these articles, and they pick up 3300 entities randomly and tag manually as development data. Finally, accuracy of system is about 80%

Researchers also extract relations between authors, papers, and their research domain. Matsuo et al.[‘07] provided a system describing social network between researchers. The corpora Matsuo used are web pages, conference data, papers and

abstracts. They Use several mining algorithms to analyze data. Then they can output a social network graph based on their analysis.

2.2 Target Relations and Their Related Entities

Rosario and Martin [‘05] propose a multi-way relation classification between proteins. They provide a list of interaction types; they classify each protein pair into interaction types. In this paper, they proposed a dynamic graph model as identification scheme. The method yields 50% precision better than baseline which is a trigger words approach.

Pantel and Pennacchiotti [‘06] proposed a method to extract five types of relations between nominals. It is a serial pattern generating method using very large corpora, such as Web. They perform their experiments by using TREC and CHEM datasets. The identification performance of f-score is 0.5~0.8 f-score according to different relation types and different parameters.

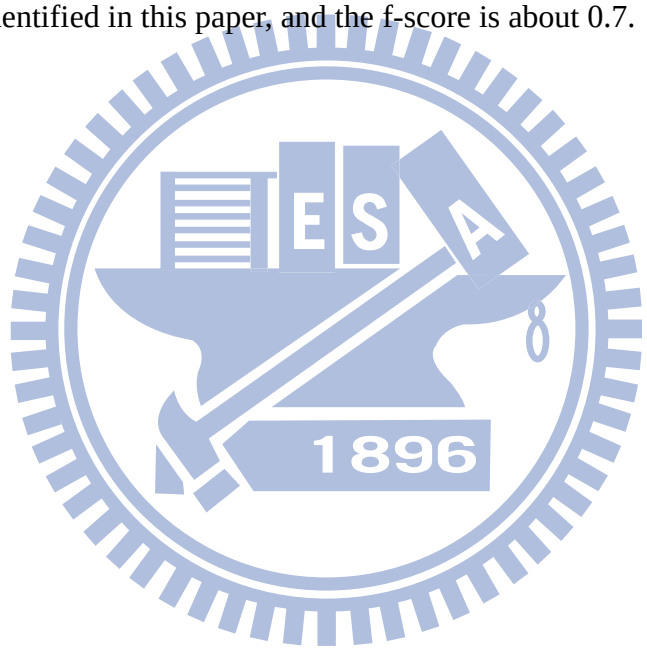
2.3 Identifications Methods

The identification used in previous work can be: pattern matching, statistical learning model, and combination model of them. It is a difficult task to find training corpora with well tagging; therefore a bootstrapping algorithm can help to build training data. [Qian et al. ‘09] proposed method to label their training corpus. First of all, they tagged a part of data, called seed data. The bootstrapping algorithm runs until no more data was tagged from dataset. The method is used to classify relation into 7 classes and 22 sub classes. The performance of this research is about 69%, and performance increases with the time of bootstrapping processing.

In [Tian and Qian ‘10], the authors extracted kinship relationship from search

engine results. They collected a set of clue words for each types of kinship relationship. The queries they used are composed of person name and clue words. The most frequent person shown in retrieved snippets has relationship to the given person. The accuracy of this method is between 0.7~0.9.

Kate and Mooney [‘10] provide a method based on card-pyramid parsing for English text. They provide a parsing diagram from the pyramid structure, and decide relation types from the parsing results. They also use pyramid parsing diagram to extract named entities. The accuracy of named entity recognition is 0.87~0.94. There are 5 types of relations to be identified in this paper, and the f-score is about 0.7.



Chapter 3 The Proposed Methods

In this chapter, we present our methods starting with how to retrieve snippet of a given person name, and preprocessing of snippet including POS tagging, named entity recognition, and 3rd-pronomial resolution. Our goal is to extract persons they have three types of relationships: kinship relation, co-working relation, and other relation to a giving person name from retrieve snippets. Kinship relationships are father, mother, couple, sister or brother etc.; and are set up by consanguinity or matrimony. Co-working relationship represents persons who work for the same project, the same organization, or they deliver a speech together.

3.1 Snippet Preprocessing

When we want to retrieve snippets from search engine, we need to solve person name disambiguation problem that different persons have the same names. To solve this problem, we expand our search word by some clue words. For a given input name, such as “吳宗憲”, we use additional information to modify his domain attribute. In fact, we know at two famous person names: “吳宗憲”, one is a professor of National Cheng Kung University Computer Science Department, and the other is a famous singer, comedian. The system requires an additional *domain word* or input domain of this person to distinguish which one is the person we want to know. We also use clue words to expand query word to increase quality of retrieve snippets. We use kinship clue word listed in

[Tian “09], after expansion of Tongyici Cilin[1], we have . Clue word database also collected a set of work clue word, such as “經理” or “主席”, in order to retrieve snippets which have more information about work related to target person. Clue words are listed in table 3.

type	Clue word set
Kinship clue word set	母親 慈母 阿媽 媽媽 娘 娘親 令堂 家母 老母 老娘 後媽 乾娘 乾媽 母女 母子 父子 父親 爸爸 老子 爹爹 爹 阿爸 生負爹地 慈父 令尊 家父 其父 嚴父 父女 丈夫 妻 愛人 夫人 老婆 太太 老公 老伴 老頭子 相公 娘子 內人 夫婦 兩口子子女 兒子 崽 愛子 犬子 令郎 少爺 女兒 閨女 姑娘 丫頭 妮愛女 小女 令愛 千金 兒女 孩子 心甘 寶貝 寵兒
title clue word set	總統 委員 助理 院長 部長 經理 總裁 股東 董事長 董座 執行長 襄理 行政祕書 業務 工程師 部長 課長 教授 職員 工作員 員工 院長 所長 校長 處長 秘書 公關 發言人 資政 顧問 任教 市長 立委 議員 委員

Table 1: List of clue words

Table 3 is an example of input person name, how we select his domain words and clue words to expand its query.

Query format	Expanded query
Person name, domain word	郝龍斌 AND(市長)
Person, kinship clue word set, domain word	郝龍斌+(父親 OR 母親…) AND(市長)
Person name, domain word set, domain word	郝龍斌+(事長 OR 議員…) AND(市長)

[1]同義詞詞林, http://ir.hit.edu.cn/phpwebsite/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=162

Table 2: System input and output sample

After query expansion, we retrieve snippets of Google search engine, and send snippets to CKIP[2] segmentation system. We handle named entity recognition by rule based on part-of-speech (POS) tags. In case of person name recognition, we collected a name database from college entrance examination result, and a surname database. For words' POS tags are 'Nb', we check if it is a person name or a person first name by matching to names in name database. In case of organization names, system identifies a noun phrase is an organization name if a NP's last word is tagged as "NC" and tagged as "+organization" by CKIP lexicon[3], then threat that NP is an organization. For example, there are three noun phrase in "交通(Na)大學(Nc)資工系(Nc)系主任(Na)", we list below.

e.g.1 交通(Na)大學(Nc)

e.g.2 交通(Na)大學(Nc)資工系(Nc)

e.g.3 交通(Na)大學(Nc)資工系(Nc)系主任(Na)

The last words of e.g.1 and e.g.2 are tagged as 'Nc' and "+organization", in fact, they are organizations, we select the longest match pattern, "交通大學資工系" as organization output. In additional, we collect organizations in wiki[4]: Taiwan organizations list and companies listed in TWC OTC market[5] and universities in Wikipedia: universities in Taipei[6] to form a organization database. We have total 3753 organizations in this

[2] 中研院詞庫小組斷詞系統, http://ckipsvr.iis.sinica.edu.tw/

[3] 中研院八萬詞詞庫, http://www.aclclp.org.tw/use_ced_c.php

[4] Wikipedia, 台灣公司列表, http://zh.wikipedia.org/zh-hant/台灣公司列表

[5] 上市上櫃公司簡稱及全名列表, 擷取自 http://www.sfi.org.tw/newsfi/intdb/menu/firstpage.asp

[6] Wikipedia, 台灣大專院校列表 http://zh.wikipedia.org/wiki/台灣大專院校列表

database.

We randomly test 400 named entities, and following table 3 is result. Most errors due to segmentation error, such as “修科大圖書館” should be “正修科大圖書館”.

	True-positive	True-negative	precision
overall	580	216	72.9
Person name	330	70	82.5
Organization name	250	146	63.1

Table 3: precision of named entities recognition

We use pronoun resolution method in [Pan and Liang, ‘09]. In this case, we only recognize singular 3rd-person pronoun. After we find a 3rd person pronoun and the antecedent, we replace pronoun with the antecedent.

In order to distinguish domains of persons extracted from snippet, we compute frequencies of domain words and named entities for each domain in appearance snippets. Domain of the neighbor title word to the person is weighted, for instance, in the case of “百戰百勝主持人胡瓜”, “胡瓜” is selected person, and the neighbor title word is “主持人”, this word is a domain word of “娛樂.”

3.2 developing Corpus

We discuss training in this chapter, including collection and pattern extraction. First, we get Taipei famous person name list from Wikipedia[7]. For each name in name list, if corresponding page exists in Wikipedia, we clawed them and store as text blocks. We

[7] Wikipedia 台灣人列表, http://zh.wikipedia.org/zh-hant/台灣人列表

manually tagged person name pair with kinship relationship in these text blocks containing clue words:”父”, “母”, 子”, “女”, “姐”, “妹”, “兄”, “弟”, “姨”, “婆”, “姊”, “姑”, and “結婚”. In our observation, Wikipedia describes person’s family in the one block; therefore although the amount of clue words of clawing are only thirteen words, we can still find person name pairs of other relation types. We claw snippets of these person name pairs as our kinship relation training corpus.

Wikipedia person file	1496 persons
Text block with query words(父母子女夫妻)	1334 blocks
Person pair extracted from person file	494 pairs
Effective person pair (removing null-snippet person pairs)	346 pairs
Snippets retrieved from Google (after removing duplicate)	29205 snippets
Average #person/ word count per snippet	2.08/50.09

Table 4: Kinship relation developing corpus

We extracted syntactic pattern from kinship relationship corpus. Syntactic pattern is consisted of a part-of-speech sequence, person name information, kinship relationship clue words and wild card. We set the max length of wild card to three here, three words can tolerate with an adjective phrase with an adverb. We extract long sentences from training snippets, and pick up sentences containing kinship relationship training person pairs. To process these sentences, we remove words content information whose part-of-speech tag is not in listening part-of-speech list. Part-of-speech tags we listened are “Na”, ”VA”, ”VB”, “VC”, “VCL”, “Caa”, “DE”, “SHE,V_2”. We select these tags

here because these part-of-speech tags including most of verbs, be-verb, conjunction, prepositions and normal nouns, are important issues for human relationship extraction. Title words are normal nouns, such as father, mother, CEO and president. We collect words which are not tagged in previous step to wild card, and if length of words is more than 3 then drop this sentence from collection. We compute frequency of each pattern shown in collection, the top 9 patterns of result are list in table 5.

POS sequence	Extracting pattern amount	Coverage in relation tokens	sample
PE DE KEY PE	33	106	白崇禧之子白先勇：在花園
Na PE DE KEY PE	28	25	已故婦女部主任彭婉如的夫婿洪萬生和政大
Na PE KEY PE	25	12	２７】總統馬英九父親馬鶴凌在湖南衡山
PE _ PE DE KEY	13	74	唐美雲為「戲狀元」蔣武童之女，
PE DE KEY PE _ Na	13	102	馬英九的太太周美青，命理師稱為
Na Na PE KEY PE Na Na V_2	13	1	第一時尚名媛蔡依珊祖父蔡垂碧外遇酒女有
PE _ Na _ KEY PE	12	201	白先勇，作家。，其父白崇禧為國民黨高級
Na PE KEY PE _ Na	10	12	台中市長胡志強夫人邵曉鈴發生車禍後
Na _ PE DE KEY PE	9	87	人民法院給李敖之女李文下達了終

Table 5: Pattern system result

For co-working relationship, we looked for persons who work in the same

organization, play in the same team, or work on the same project from web page of organizations, such as NCTU, “中國國民黨” etc.. For these person name pair, we not only extract snippets which person co-occurred, but also extract snippets of single person. Here we collect snippets with only one person because we want to analysis organization and person related to only person. These entities can be an important factor to represent person’s work. In co-work relation training corpus, we collect 35 pairs per each domain, total 175 pairs. Following table 6 is the statistical data of co-working relationship training corpus.

Total person pairs	175 pairs
Total clawed sentences	27025 sentences
Average #entities / word count per snippet	2.68 persons/48.61 words

Table 6: Statistic of co-working relationship training corpus

After we built co-working developing corpus, we perform a bootstrapping algorithm to extract domain words for each domain. We select seed words of each domain as initial data to algorithm, which list in table 7. The algorithm is list in figure 1.

domain	Clue words
政治	總統,立法委員,立委,議員,議長,市長,縣長,部長,院長,資政,主席,秘書長,委員長,處長,署長,副院長,局長,中常委,中評委
經濟	總裁,執行長,董事長,董事,經理,總經理,工程師,醫生,護理長,護士,理事,協理
體育	球員,選手,棒球,網球,桌球,田徑,排球,沙排,籃球,高爾夫,手球,曲棍球,足球,得分王,競賽,比賽,奧運,好手
娛樂	星光,演唱,電視劇,模特兒,歌手,主持人,歌喉,嗓音,默劇,相聲,專輯,連續劇,偶像,藝人
教育	教授,老師,指導,研究,教鞭,審查

Table 7: Seed of bootstrapping

Input: seeds for each of domain,developing sentences(total 119529 sentences)

ListenPOS = Na,Nb,VA,VB,VC,VD

Output: domain word set for each domain

Initialization:

For all seeds, $vector_{word}$ = [1, if word's domain match; 0 otherwise]

WordSet = list of word in seeds

SentenceSet = void list

Step:

1. For sentence in developing sentences, if sentence containing word in Wordset: $vector_{word} = vector_{word} / |vector_{word}|$,and add sentence to SentenceSet
2. For word in sentence and word's POS in ListenPOS, update word data in WordSet: $vector_{word} = vector_{word} + vector_{sentence}$
3. For word in WordSet, update word: $vector_{sentence} = (\sum_{w \in sentence \& w \in wordset} vector_{word}) / count(w_{w \in sentence \& w \in wordset})$
4. If sum($vector_{word,thisstep}$-$vector_{word,prestep}$)/count(WordSet) > threshold or looptime>maxlooptime: stop and return wordset
   else: loop step 1 to 4

Figure 1: The bootstrapping algorithm

In initial step, words in wordset are the seed words. If the domain is not equal to the domain of seed word, this value of vectors is zero, otherwise is one. In step1, we use word to tag sentences. If sentence contains words which are in wordset, then the domain vector of sentence equal to the average of these vectors of words. In step 2, we extract words from tagged sentences in step 1 if word's part-of-speech in our concentration and the vector value of this word in wordset is equal to the sum of sentence vectors. After the processes of all sentences tagged in step 1 have done, we normalize domain vector value

of word in wordset to 1. Stopping conditions are the update of each loop is less than threshold or the maximum loop time occurring. Here we list the top 20 words of each domain in Table 8; complete 80 words will list in appendix A.

domain	Clue words
政治	文才,記過,薦任,內規,職等,事態,賄選,杯葛,參議員,政委,國代,補提名,警察局長,代表會,技監,密帳,頭殼,市委,縣議員,檢察長
經濟	國際商,兆豐銀,土建融,營業員,餘額,代表人,股份數,股,連接器,消費性,張發得,日盛金,事業群,純益,主機板,監察人,財務長,信案,負債表,競業
體育	公開組,木簡,陽岱鋼,系列賽,女單,混雙,總獎金,男雙,快攻,止步,外野手,女雙,擊出,內野手,敲出,日本隊,羽球,單場,未來賽,澳網
娛樂	片尾曲,超偶,烈火,文字幕,珍藏版,編曲,婚頭,虎將,翻唱,作詞,駐唱,創作人,相聲劇,搖籃曲,女聲,作曲,實力派,淚光,主題曲,歌名
教育	口試,切換,馬達,研究生,微波,超導體,指導,語言學,正教授,博士生,頁數,外文,耐心,經方,要略,指正,現址,高分子,微電子,論文

Table 8: The top 20 of domain words

3.3 Relation Identifier

Hybrid method we are talking about is combined with a pattern match system and a statistical model. For a given person pair *example,* we call this person pair as relation candidate, before we verify it relationship, we need to make sure that information of this relation candidate is enough to classify its relation. If amount of sentences corresponding to this relation candidate is less than 5, we will craw data from search engine again. The query in this re-craw will not extend by any clue words but domain words for each name are needed. After we make sure information of relation candidate is enough, we try to match all patterns from pattern database, if a match process of any sentence of relation

candidate successes, and then we can say this relation candidate has kinship relationship.

If no sentence matched pattern, we extract following feature as statistical model testing input.

Feature	
F1	Average distance between two names (this value is normalized by sentence length)
F2	Amount of total keywords
F3	Percentage of keyword shown before first name
F4	Percentage of keyword shown between two names
F5	Percentage of keyword shown after last name
F6	Average distance between keyword and first name
F7	Average distance between keyword and last name

Table 9: The features used for kinship identification

We perform an experiment to test the significance of each feature by leave-one-out strategy. We list the result in table 10.

Feature set	f-score(%)
all	86.17
all-{f1}	78.36
all-{f2}	80.67
all-{f3}	79.12
all-{f4}	77.76
all-{f5}	78.60
all-{f6}	78.43
all-{f7}	77.62

Table 10: Kinship feature impact result

In table 10, we find out the most significant feature in feature set is the average

distance between person name and clue word; and the next feature is clue word between two person names. Testing data is listed in appendix B.

Statistical model for co-working relationship is processed parallel with kinship identifier. In observation, we found that it is possible that no co-occurrence snippet for two person who have co-working relationship, but we can still find trail from organizations and people related to these persons. Extracted feature lists in table 11.

Feature	description
F1	cosine similarity of two named entities vectors
F2	cosine similarity of two named entities vectors(remove count of entities shown in co-occurring sentence)
F3	same domain
F4	the amount of the same related organization
F5	the amount of the same related person (must be in same domain)
F6	average distance between names (normalize by length of sentence)
F7	the amount co-occurrence snippet

Table 11: The features used in co-working identification

We test feature significance by leave-one-out strategy, performance listed in table 12.

Feature set	f-score(%)
all	75.59
all-{f1}	75.59
all-{f2}	75.59
all-{f3}	63.63

all-{f4}	63.63
all-{f5}	78.02
all-{f6}	77.78
all-{f7}	75.24

Table 12: Feature test for co-working

In table 12, we can find that the amount of co-occurrence organizations and domain are the most significant features, and the co-occurrence persons is the negative feature. We discuss that too many persons are occurring in the same sentence and this reason effect the result. Positive instances and negative instances are listed in appendix C.

Chapter 4 Experiment and Result

Close data testing can measure our system performance. We collect 300 relation candidates containing sentences with kinship clue words as training negative instances. In statistical model, we take person pairs and snippets from kinship relationship training data as positive instances. We select supported vector machine as model. Statistical model training and testing we use 5-folder cross-validation. Result listed in table 15 and table 16.

	Precision	Recall	f-score
Pattern based classifier	90.79	26.85	41.44
Kinship statistical classifier	90.83	80.27	85.22
Hybrid classifier	90.35	83.51	86.80

Table 13: Kinship identification performance

Pattern match /SVM/hybrid	Positive	negative	total
True	97/298/304	270/69/63	367
False	4/37/39	296/263/261	300
total	101/335/343	566/332/324	

Table 14: Instances of kinship identification

In the case of pattern system, we select top 9 patterns to build pattern based classifier, pattern are listed in table 8. If a sentence of a relation candidate matched pattern, we set up kinship relationship for this relation candidate. Hybrid classifier is to combine pattern base classifier and kinship statistical model.

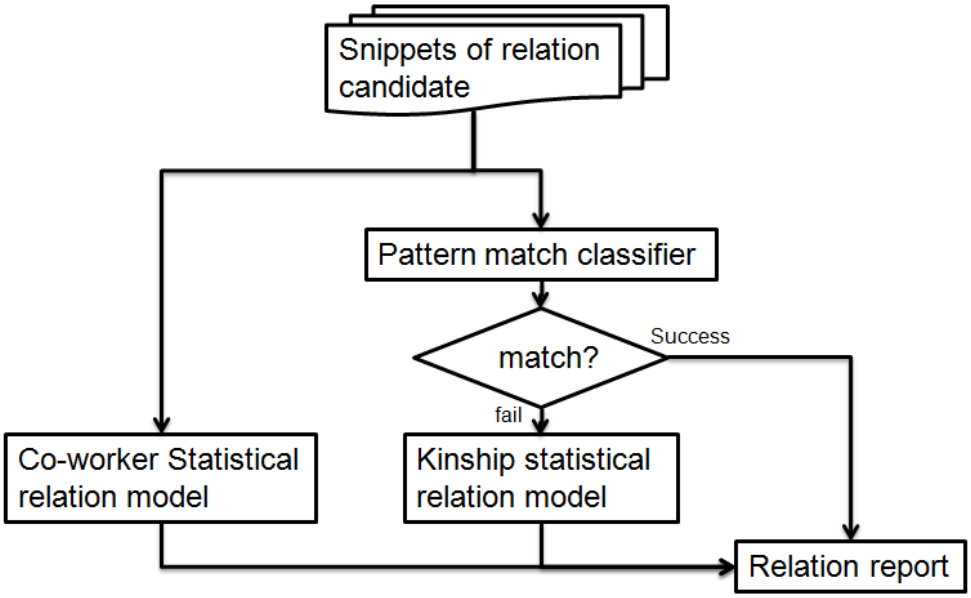


Figure 2: System process flowchart

We select positive instances of co-working relationship classifier from co-working training corpus. We randomly select 20 candidates from entities which have the same domain in pairs for each domain, and randomly select 100 relation candidates which contain more than 5 sentences as negative instances. We use five folder cross-validations to verify its performance. Performance listed in table 11.

This experiment shows our method having better performance using combination of pattern based and statistical model. For co-working relation extraction, sometimes we cannot extract feature from sentences because the co-occurrence sentences is not enough for training.

Chapter 5: Conclusion and Future Work

We extract relation from the web and classify them into kinship, co-working and other relation. We perform a real-time platform by python[8] GUI[9].

In observation, we found the result of system is limited by retrieved snippets. Data stored in search engine get closed to events people care about. This cause a data bias problem: information size of famous person is huge, but size of a common person is very small. This problem does a big effect to our system while extracting a person who is not famous, such as “白崇禧”, is a famous general of China, has 1,610,000 snippets from search engine; but his little son, “白先敬”, has only 7890 snippets as result. However we can extract relations of persons from retrieved snippets.

We can increase domain size. Lot of persons cannot classify into these domains, for example, artist, writer, and killer. If we can define domain by detail, performance of our system can be better. In another hand, relationships between human are not only 2, we can still identify more types of relationships.

In this thesis, we proposed a method identifying human relationship from retrieved snippets by a two-stage identifier. We also proposed a bootstrapping algorithm for extract domain words from corpus for a given domain.

[8] http://www.python.org/

[9] http://docs.python.org/library/tkinter.html

References

[1]. Gang Wang, Yong Yu and Haiping Zhu (2007), "PORE Positive-Only Relation Extraction from Wikipedia Text" in "Lecture Notes in Computer Science", 2007, Volume 4825/2007, Pages 580-594

[2]. Dat P.t. Nguyen, Yutaka Matsuo and Misuru Ishizuka(2007), "Relation Extraction from Wikipedia Using Subtree mining", Proceeding AAAI'07 Proceedings of the 22nd national conference on Artificial intelligence - Volume 2, Pages1414-1421

[3]. Yutaka Matsuo, Junichiro Mori and Masahiro Hamasaki (2007), "POLYPHONET: An Advanced Social Network Extraction System from the Web", in "Web Semantics: Science, Services and Agents on the World Wide Web" volume 5, issue 4, December 2007, Pages 262-278

[4]. Barbara Rosario and Marti A. Hearst (2005) "Multi-way Relation Classification: Application to Protein-Protein Interactions", in "Proceeding of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing(HLT/EMNLP)", Pages 732-739

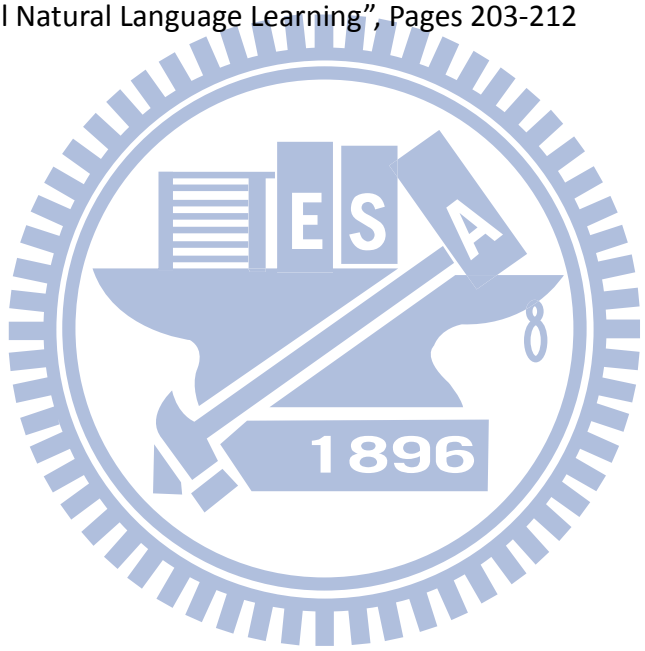
[5]. Patrick Pantel and Marco Pennacchiotti (2006) "Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations" in "Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the ACL", pages 113-120

[6]. Longhua Qian, Guodong Zhou, Fang Kong and Qiaoming Zhu (2009),

"Semi-Supervised Learning for Semantic Relation Classification using Stratified Sampling Strategy" in "Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing", Pages 1437-1445

[7]. Gan Tia and Mo Qian (2009), "Research on Social Relation Extraction of Web Persons", in "2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing" Pages 606-610

[8]. Rohit J. Kate and Raymond J.Mooney (2010), "Joint Entity and Relation Extraction using Card-Pyramid Parsing", in Proceedings of Fourteenth Conference on Computational Natural Language Learning", Pages 203-212



Appendix A: domain word database

政治	娛樂	教育	經濟	體育
文才	片尾曲	口試	國際商	公開組
記過	超偶	切換	兆豐銀	木簡
薦任	烈火	馬達	土建融	陽岱鋼
內規	文字幕	研究生	營業員	系列賽
職等	珍藏版	微波	餘額	女單
事態	編曲	超導體	代表人	混雙
賄選	婚頭	指導	股份數	總獎金
杯葛	虎將	語言學	股	男雙
參議員	翻唱	正教授	連接器	快攻
政委	作詞	博士生	消費性	止步
國代	駐唱	頁數	張發得	外野手
補提名	創作人	外文	日盛金	女雙
警察局長	相聲劇	耐心	事業群	擊出
代表會	搖籃曲	經方	純益	內野手
技監	女聲	要略	主機板	敲出
密帳	作曲	指正	監察人	日本隊
頭殼	實力派	現址	財務長	羽球
市委	淚光	高分子	信案	單場
縣議員	主題曲	微電子	負債表	未來賽
檢察長	歌名	論文	競業	澳網
村長	唱作俱佳	土壤	行員	二壘
雪峰	南拳	獲益	原董事長	游擊手
彈劾	狂想曲	古文	品保	準決賽
議長	兵衛	副教授	獲利	安打
鄉長	御天	教誨	協理	會外賽

庭長	獻聲	教授	兆豐金	男單
方方土	聲帶	撰	匯通	雙打
台北縣長	情歌	薄膜	投信	衛冕
黨籍	小旦	微生物	股務	轟出
分區	曲目	老師	群創	火腿
書記長	宵	師事	券商	公開賽
司法院長	浪子	副研究員	董事	洲際盃
偽證案	試聽	集刊	觸控	守備
提名人	鴨子	組員	震出	紅襪
參議	品冠	肥料	模具	失分
會勘	合輯	論文集	期貨	力克
考試院長	客串	論文獎	副總經理	完投
國民黨籍	曲風	物理	面板	三振
特赦	粵語	期刊	出貨量	單打
巨額	名作	科普	群聯	攔網
攤商	演唱	原子	聲響	預賽
辱罵	合演	學門	聯貸	強攻
在野黨	保庇	可靠度	換股	三壘手
黨團	星光	學號	股本	冠軍戰
上山	試片	授課	年報	二軍
司令	鼓手	卜卦	國泰金	首役
幹事長	創作型	數理	產品線	台灣隊
總質詢	一修	菌種	詳閱	助攻
主席團	樂隊	實作	解任	打擊率
槍殺	唱將	社會學	控股	籃板
費案	燕姿	研究獎	實收	亞青
華柱	歌詞	磁性	語調	球路
民政	貝殼	高溫	橡膠	打者
參拜	發音	研發長	銑	中繼
徵召	嗓音	量測	積體電路	先發投手
參謀總長	丫頭	關鍵字	產能	揮出
脫黨	搖滾	行政學	產險	國訓
縣市長	專輯	藝術史	光電	法網
紅派	唱腔	取向	聚合	水手
國務	單元劇	語文	技術員	軟銀隊
省議員	納豆	廢水	神通	失誤
市議員	大碟	史學	股利	變化球

李永得	媛	元件	紡織	世錦賽
定讞	牛仔	續集	持股	日職
民進黨籍	新歌	燈光	轉讓	一壘
中樞	大道	評論人	美玉	挺進
拜票	選輯	水工	天縱	投出
收賄	填詞	數學	產物	先發
健康局	主演	普渡	相機	防禦率
立法院長	同學會	餐旅	減資	排名賽
列名	歌曲	系主任	營運長	擊敗
副議長	滾石	缺陷	畫素	輸給
胡為	旋律	探勘	證券	羽毛球
內政	床戰	原理	裹理	三壘
初選	點播	姓名學	電纜	九號球
外交部長	小琥	學系	工讀生	此役
啟臣	原唱	原名	總座	網賽
民進黨團	舞曲	研讀	半導體	小將
指揮官	卡司	學科	華紙	延長賽
提名	八點檔	終生	出貨	左投

Appendix B: kinship corpus

Positive (total 367 instances)			Negative (total 300 instances)		
白崇禧,白先敬	郭柏川,朱婉華	朱天心,朱西甯	邱毅,揚揚	李心潔,張小燕	張淑芬,陳維昭
白先勇,白崇禧	郭雪湖,郭禎祥	朱西甯,朱天文	江蕙,洪一峰	婉玲,張宗慈	林平侯,林應寅
白先勇,馬佩璋	郭雪湖,郭松棻	朱西甯,朱天衣	呂政儒,達欣	洪濬哲,韋杉	呂丹虹,兆豐
白先勇,白韻琴	郭雪湖,郭香美	陳威,陳儀	楊錫安,莊瑞雄	唐志中,大炳	陳水扁,陳唐山
白吉勝,許毓娟	郭雪湖,郭松年	陳儀,陳月芳	周宗豪,莊祥華	吳育綸,楊清順	劉薰愛,郭子乾
白冰冰,白曉燕	孔德成,孔德齊	陳大豐,陳大順	謝佳賢,林瀚	辛金展,吳佳龍	顏聖冠,林岱樺
彭淮南,賴洋珠	孔德成,孫琪方	陳定南,陳仁杰	陳鴻文,道奇	陳信安,羅興樑	東亞運,介德
彭婉如,洪萬生	孔德成,孔維鄂	陳定南,陳仰德	包宗和,歐盟	王心恬,尹蓁晞	周杰倫,游鴻明
馬英九,馬鶴凌	孔德成,孔維益	陳慧坤,陳清文	長庚,黃世聰	李宗盛,胡茵夢	朱志清,蔡福財
馬英九,秦厚修	郝龍斌,郝柏村	陳履安,陳宇廷	田鵬飛,司諾克	李千娜,江惠	張建邦,莊祥華
馬英九,周美青	林月雲,侯佩岑	陳履安,陳宇銘	張泰山,蕭一傑	林青霞,邢愛林	介德,林建平
馬英九,馬唯中	洪榮宏,洪一峰	陳履安,陳宇慷	戴家瑞,李瑞斌	蔡於儒,林錦明	兆豐,林宗勇
馬英九,馬冰如	洪一峰,陳愛玲	陳履安,陳宇慧	洪濬哲,王耀煌	黃文偉,駱國明	陳鏞基,黃智培
馬英九,馬莉君	洪曉蕾,王世均	陳履安,陳宇全	宋承憲,元斌	楊淑君,宋玉麒	吳志偉,張智峰
馬英九,周美青	洪致文,洪以南	陳水扁,陳致中	裕隆,倪雅倫	林華德,賴永吉	劉德華,張清芳
馬英九,馬唯中	胡志強,胡婷婷	陳水扁,吳淑珍	亞運會,莊智淵	郝龍斌,顏錦福	洪敬堯,洪一峰
馬英九,馬元中	胡志強,邵曉鈴	陳義信,陳鴻文	呂秀蓮,敏豪	劉德華,徐熙娣	宋慧喬,宋承憲
馬景濤,馬世媛	胡志強,胡韡耀	陳明文,陳福三	謝鳳秋,張菁雅	黃建逢,溫笙銘	伍錦霖,李彥謀
馬景濤,唐韻	胡盈禎,胡自雄	常楓,張遙	陳義信,吳復連	施易男,吳鈴山	許振榮,任弘
馬景濤,吳佳尼	李進良,胡自雄	沈葆楨,林普晴	和信,辜家	詹雅雯,陳雷	郭台銘,嚴凱泰
方耀乾,方榮欽	黃海岱,黃馬	沈葆楨,林則徐	蔡英文,史明	黃志雄,林曉儀	辜家,辜啟允
蔣經國,蔣方智怡	黃海岱,程扁	沈玉琳,吳曉純	蔡萬春,關穎	王文華,明正	張凱貞,晏紫
蔣方智怡,蔣孝勇	黃海岱,黃俊卿	施明德,陳麗珠	皮瑞,甲柯	吳岱豪,辛金展	邱宗志,倪雅倫

傅培梅,程紹慶	黃海岱,黃俊雄	史明,林濟川	蔣介石,陳麗珠	施崇棠,黃顯雄	郭台銘,戴爾
傅培梅,程顯灝	黃俊雄,陳美霞	史哲,史英	裕隆,陳信	陳昶輔,林侑	關之琳,張學友
傅培梅,程安琪	黃俊雄,黃強華	任賢齊,陳則妤	仲村亨,溫昇豪	鄭成功,鄭芝龍	陳景峻,黃旭昇
傅培梅,程美琪	黃俊雄,黃文擇	任家萱,任容萱	謝長廷,朱立倫	李建新,郭建昇	詹江村,楊麗環
鄧麗君,鄧長安	黃俊雄,黃文耀	任家萱,任明廷	吳清基,巫秀娥	周士淵,李至明	包宗和,郝思曼
杜正勝,陳芳妹	黃志雄,洪佳君	任家萱,張承中	伊林,林志穎	王力宏,伍思凱	張菲,田麗
杜正勝,杜明夷	黃信介,黃宗河	曾國藩,曾約農	李天欣,吳鳳	武雄,林岳平	瓊瑤,陳思
田麗,陳定中	黃信介,黃至君	曾志朗,洪蘭	陳永誠,林宗勇	許慧欣,黃文華	鄭崇華,台達
唐美雲,唐冰森	黃勁連,黃三郎	曾志朗,曾允中	世大運,李欣翰	蔡康永,康熙	呂秀蓮,游錫
唐美雲,蔣武童	曾治豪,賀一航	蔡康永,蔡天鐸	吳建豪,洪百榕	蔡鴻文,施啟揚	陳泰然,顏泰崇
陶晶瑩,李李仁	紀政,張博夫	蔡依珊,蔡惠媚	陳信安,姚明	楊秋興,余陳月瑛	劉義祥,林建平
倪敏然,倪蓓蓓	蔣中正,宋美齡	連勝文,蔡依珊	林世民,潘武雄	楊淑君,賴聖蓉	游鴻明,潘偉柏
雷學明,雷倩	蔣經國,蔣中正	蔡依珊,蔡垂碧	蔡友,盧正昕	陳柏良,陳俊宇	林進春,陳秀卿
雷倩,鄭正玨	蔣中正,毛福梅	蔡依珊,蔡惠媚	葛優,王學圻	許水德,倪文亞	林道遠,趙詠華
李政道,李駿康	蔣中正,蔣緯國	蔡依珊,蔡綠峰	葉志仙,涂忠男	邱議瑩,邱信福	張立杰,田冠鈞
李政道,李仲覃	蔣中正,陳潔如	蔡依珊,陳娟娟	陳若儀,小林	蔡天鐸,偉忠	謝霆鋒,謝賢
李政道,秦惠菩	蔣中正,姚冶誠	蔡萬霖,蔡萬春	陳宏麟,蔡佳欣	林智勝,羅國輝	每日,日盛
李安,李昇	蔣經國,蔣方良	蔡萬霖,蔡萬才	裕隆,吳永仁	駱錦明,季欣麟	辜家,嘉新
李安,林惠嘉	蔣經國,蔣孝文	蔡萬霖,蔡萬春	洪秀柱,黃志雄	孫震,辜振甫	廖婉汝,吳英毅
李安,李涵	蔣經國,蔣孝武	蔡萬霖,蔡辰洲	亞運,柯秉逸	楊渡,林澄枝	莊文思,李永萍
李安,李淳	蔣經國,蔣孝勇	蔡英文,蔡潔生	羅錦興,陳裕民	甘秉宗,陳建禎	徐克,趙文卓
李登輝,李安妮	蔣經國,蔣孝章	蔡英文,張金鳳	孫永年,林仁輝	郭泰源,林華韋	郭台銘,胡釗維
賴國洲,李安妮	蔣經國,章孝嚴	宋楚瑜,陳萬水	達新,倪雅倫	黃立成,上中天	林志玲,周董
李敖,李鼎彝	蔣經國,章孝慈	宋楚瑜,宋達	隋棠,藍正龍	盧正昕,建華	傅培梅,程美琪
李敖,張桂貞	蔣緯國,戴季陶	宋楚瑜,宋鎮遠	何豪傑,范瑞君	曾文鼎,陳信	蘇嘉全,賴英照
李敖,胡茵夢	蔣緯國,石靜宜	宋楚瑜,宋鎮邁	洪曉蕾,馬國畢	李雲光,林建平	捷星,袁祥芝
李敖,李文	蔣緯國,邱愛倫	宋達民,洪百榕	郭台銘,賈伯斯	莊賦祥,陳英忠	傅哲偉,王泓翔
李敖,李戡	蔣緯國,蔣孝剛	宋逸民,宋達民	張國榮,新藝	李光陸,簡宜彬	林義傑,張洛君
李敖,李諶	蔣孝武,蔡惠媚	蘇貞昌,蘇雲英	吳宗憲,賀一航	黃志龍,林羿豪	趙豐邦,柯秉逸
李登輝,李金龍	蔣孝武,蔣友松	蘇煥智,蘇添水	簡嘉宏,子威	林子祥,吳大維	林志玲,白歆惠
李登輝,李登欽	江丙坤,江俊宏	蘇慧倫,蘇霈	郭台銘,賓士	林瀚,郭嚴文	黃炳煌,黃海寧
李登輝,李炳楠	江丙坤,江俊德	孫運璿,孫蓉昌	洪敬堯,洪榮宏	周定緯,安伯政	謝金燕,曾昱嘉
李登輝,曾文惠	江丙坤,江素華	孫運璿,孫璐西	王文華,吳東亮	杜正勝,蔣仲苓	娟娟,陳娟娟
李梅樹,李金印	江淑娜,江蕙	孫運璿,孫璐筠	廖正民,鵬舉	邱大宗,許晉哲	鄭南榕,王淑芬
李梅樹,劉清港	齊豫,齊秦	孫運璿,孫一鶴	王彩樺,官靈芝	陳義信,曹錦輝	辜振甫,辜仲諒
李鴻禧,李長庚	錢穆,錢偉長	孫運璿,孫一鴻	廖振隆,卓永財	席開,唐志中	溫嵐,許茹芸
李煥,李慶華	秦楊,黃西田	孫逸仙,孫眉	黃美珍,葉瑋庭	郁芳,何守正	陳義信,劉義傳

李慶安,李慶華	丘逢甲,丘龍章	孫逸仙,盧慕貞	徐牧基,游象富	陳國,吳東亮	蔣榮先,曾元顯
李遠哲,李澤藩	邱毅,謝京叡	孫逸仙,宋慶齡	元大,京華	邱宗志,小邱	黃志龍,羅德
李遠哲,李遠川	瓊瑤,陳錦春	隋棠,隋華	劉謙,賀一航	賴聲川,蕭艾	宋達民,李明依
李心潔,彭順	瓊瑤,陳壯飛	歐陽龍,傅娟	傅子純,宋逸民	張清芳,何守正	吳岱豪,邱啟益
李鎮源,莊絜	蕭敬騰,蕭百城	顏聖冠,顏錦福	李勝木,胡崇賢	洪茂峰,段維新	溫昇豪,張艾嘉
李宣榕,彭思維	謝森中,謝延禧	楊傳廣,楊博宗	呂政儒,黃萬隆	王彩樺,侯怡君	蘇貞昌,陳郁秀
連震東,連雅堂	謝東閔,謝敏初	楊傳廣,楊玉蘭	郭信良,蔡旺詮	蘇永欽,賴浩敏	徐若瑄,卜學亮
趙蘭坤,連震東	謝東閔,謝孟雄	楊傳廣,周美玲	李政穎,雷洪	官靈芝,周蕙	宗華,陳松勇
梁靜茹,張智成	謝東閔,謝孟雄	楊維哲,楊柏因	謝震武,范瑞君	呂景義,蔡於儒	王力宏,吳佩慈
梁靜茹,趙元	謝東閔,林澄枝	楊維哲,楊啟東	王文洋,羅雪映	李文媛,李驥	邱坤良,蘇金煌
廖琛,廖文毅	謝長廷,游芳枝	洪文棟,楊麗花	富邦,蔡明興	劉香慈,姚元浩	趙駿亞,許效舜
廖文毅,廖承丕	楊洋,熊天平	林淑芬,楊秋興	金渡訓,司諾克	建華,陳博志	黃能富,曾憲雄
廖溫仁,廖文毅	熊海靈,熊秀珍	楊秋興,陳哲男	徐若瑄,張秀卿	榮輝,蘇雍傑	許慧欣,國信
廖文奎,廖文毅	徐生明,徐政明	楊英風,楊朝木	達新,智峰	成龍,李連傑	唐美雲,唐冰森
廖文毅,李惠容	許世賢,張進通	伊能靜,楊淑婉	陳志清,林玉芬	潘瑋柏,周傑倫	新黨,黃信介
林佳龍,許文龍	張博雅,許世賢	殷琪,殷之浩	吳志揚,吳鴻麟	邱淑媞,蘇龍麒	胡志強,謝鳳秋
林平侯,林應寅	許世賢,張文英	殷琪,郭英聲	陳信安,倪雅倫	郭靜,方炯鑌	吳建豪,陳慧貞
林榮三,林鴻堯	許秋暘,許遠東	王拓,王醒之	蘇嘉全,林佳龍	張雨生,黃舒駿	陳水扁,杜正勝
林榮三,林鴻聯	許效舜,許義隆	王小棣,王昇	曾文鼎,裕隆	丁守中,鄭文玲	吳敦義,蔡甘連
林獻堂,林文欽	許效舜,許朝宗	王祖賢,王耀煌	蔡長海,明基友達	蕭萬長,金平	毛治國,邵琪偉
林毅夫,陳雲英	詹雅雯,詹雅芸	王祖賢,王仁峰	高慧君,洪榮宏	林益全,謝佳賢	張亞雯,何霞
林志玲,林繁男	張博雅,許世賢	王育德,王育霖	江天祐,陳昶輔	于蓓,林景庸	陳水扁,余政憲
林志玲,吳慈美	張博雅,張進通	王永慶,王天來	張誌洋,李柏憲	邱宗志,建仁	蔡辰威,關穎
林志玲,林志鴻	張光正,張光直	王永慶,王永在	陳建仁,李元貞	陳亭妃,林鳳營	陳文鍾,趙建銘
林青霞,林維良	張光直,張光誠	王永慶,李寶珠	金木,徐生明	高宇蓁,丞琳	梁偉銘,周俊三
林青霞,邢愛林	張惠妹,張惠春	王永慶,王瑞華	錢章杉,張智峰	吳育綸,吳協昌	國隆,郭岱琦
林青霞,邢言愛	張俊宏,張慶沛	王永慶,郭月蘭	王永慶,何壽川	王乾發,中堂	高志綱,盧銘銓
陳若儀,林志穎	張俊宏,許榮淑	王永慶,王文洋	白歆惠,棠棠	吳建豪,張芸京	宋楚瑜,章孝嚴
林岳平,陳惠珊	張艾嘉,魏景蒙	王永慶,林明珠	林鴻堯,林鴻聯	台新,何壽川	陳淑樺,劉若英
劉若英,劉若玉	張艾嘉,王靖雄	王永慶,羅雪貞	郝龍斌,沈世宏	李彥松,蘇雍傑	李登輝,文茜
劉若英,劉詠堯	張泰山,張全伯	王永慶,羅文源	周泓諭,周士淵	楊哲宜,李啟	康熙,曾志偉
劉松藩,劉雲騰	張泰山,張可洛	王永慶,羅雪映	吳奇隆,施易男	施文彬,王識賢	甘秉宗,康富斌
劉兆玄,劉兆凱	張泰山,吳靜怡	王建民,吳嘉妗	吳奇隆,何雨雯	馮小剛,張國立	蔡琴,齊豫
陳郁秀,盧修一	張雅琴,劉國平	王建民,王鵬硯	蔡旺詮,邱莉莉	楊秋興,林岱樺	李杜軒,林瀚
盧修一,盧佳慧	張雅琴,張瑞起	王馨平,林翠	張秀卿,陳雷	邱茂男,邱銘璋	陳祖培,世華
盧修一,盧佳君	張錫銘,張達雄	王馨平,曾江	張博雅,涂德錡	廖喜代,王雙玉	施啟揚,翁岳生
盧修一,盧佳德	章孝嚴,章孝慈	王貞治,王登美	吳鈴山,小林	陳建州,蘇健忠	吳婉茹,兆祥

盧彥勳,錢瓊文	章孝慈,周錦華	翁倩玉,翁俊明	賴清德,邱莉莉	程文欣,魏軼力	羅興稼,李雲光
盧彥勳,盧威儒	章孝嚴,黃美倫	翁倩玉,翁炳榮	楊錫安,陳智盛	亞都,運通	吳秉叡,老李
盧彥勳,盧慧源	陳水扁,趙建銘	翁倩玉,翁祖模	謝霆鋒,林志穎	邱議瑩,陳亭妃	呂政儒,辛金展
盧彥勳,許素芬	趙建銘,趙建勳	翁金珠,劉學軒	簡毓瑾,于洋		
盧彥勳,錢瓊文	趙建銘,趙玉柱	吳伯雄,吳鴻麟			
羅璧玲,羅青哲	趙又廷,趙樹海	吳伯雄,吳聖昌			
高慧君,高一生	鄭成功,鄭芝龍	吳伯雄,楊毛治			
辜振甫,辜顯榮	鄭成功,鄭鴻逵	吳鴻麒,吳鴻森			
辜寬敏,辜顯榮	鄭成功,鄭渡	吳毅,吳清源			
辜濂松,辜岳甫	鄭成功,鄭榮	吳國楨,吳經明			
辜濂松,林瑞慧	鄭成功,鄭經	吳國楨,黃金濤			
辜濂松,辜仲諒	鄭成功,鄭克舉	吳國楨,黃卓群			
辜濂松,辜仲瑩	鄭芝龍,黃程	吳奇隆,馬雅舒			
辜濂松,辜仲立	鄭經,唐顯悅	吳尊,吳成典			
辜濂松,辜仲玉	鄭南榕,鄭竹梅	吳玟萱,呂向榮			
辜振甫,張安平	鄭豐喜,吳繼釗	吳敦義,吳奚			
辜振甫,嚴倬雲	鍾理和,鍾浩東	吳敦義,蔡令怡			
辜振甫,辜懷群	鍾麗緹,嚴錚	余光中,余佩珊			
辜振甫,辜懷箴	周華健,周厚安	余光中,余季珊			
辜振甫,辜懷如	周杰倫,葉惠美	余政憲,余登發			
辜振甫,辜啟允	周杰倫,周耀中	余政憲,余陳月瑛			
辜振甫,辜成允	朱懷冰,朱邦復	郁芳,陳昱羲			
辜仲諒,辜仲瑩	朱銘,王愛	庾澄慶,庾家麟			
關穎,陳國和	朱銘,朱李記	庾澄慶,張正芬			
關穎,蔡淑媛	朱銘,陳富美	伊能靜,庾澄慶			
郭泰源,張瓊姿					

Appendix C: co-working corpus

Positive			Negative		
馬英九,蕭萬長	范光煌,賴耀群	潘威倫,高志綱	沈世宏,周慧敏	游錫,文彬	貝慕敦,何敏華
蔡英文,鄭文燦	黃永仁,林鐘雄	張泰山,林益全	丁守中,羅雪貞	沈君山,胡其湘	辜寬敏,陳致存
郝龍斌,莊文思	錢復,李永振	呂政儒,吳岱豪	馬晉,殷之浩	吳明道,吳鳳	王志高,孔德成
郝龍斌,任孝琦	蔡宏圖,蔡鎮宇	陳信安,吳岱豪	杜家濱,楊丞琳	賴耀群,台成	文本,方耀乾
郝龍斌,楊錫安	汪國華,熊明河	吳育綸,林書宏	明仁,盧正昕	世新,陳富美	郭富城,華原朋美
馬英九,吳敦義	汪國華,陳祖培	莊佳容,詹詠然	陳亭妃,張子慧	紀佳松,盧佳慧	張啟娜,海青
陳水扁,呂秀蓮	卓火土,王雪華	楊宗樺,謝政鵬	馬唯中,許榮淑	黃平洋,蔡於儒	方介民,黃石城
王清峰,陳履安	周永明,王雪華	方介民,李勝木	世新,郭英聲	貫一,周宗豪	韋帆,麥朝成
郝柏村,林洋港	賀陳旦,王雪華	程文欣,簡毓瑾	柯俊雄,黃培閎	謝欣霓,葉春智	張雨生,胡蓉蔚
李登輝,連戰	卓火土,任偉光	李勝木,簡毓瑾	邱茂男,張子	劉建國,王正旭	熊明河,觀音
彭明敏,謝長廷	張榮發,張國政	吳珈慶,楊清順	趙世璋,吳宗榮	呂西鈞,葉慶隆	李啟,米迪亞
王金平,曾永權	張榮發,鄭光遠	楊清順,趙豐邦	陳淑華,鄭成功	歐元,賴國弘	項潔,李戡
郭蔡文,毛治國	張榮發,陳泰仁	李惠堂,陳昌源	林憲銘,莫瑞	凱悅,程建中	陽森,王小棣
簡太郎,李明峰	蔡琴,梁文音	張子岱,林尚義	蔡明興,溫世仁	李朝枝,廖文彬	郭瑤琪,崔湧
邱文達,蕭美玲	潘璋柏,邰正宵	羅智聰,黃文偉	林靖軒,王復蓉	林芳郁,陳明德	蘇起,姚嘉文
蕭美玲,江宏哲	王心如,王彩樺	呂昆錡,陳柏良	李仁龍,唐雪舫	管碧玲,許義隆	嚴爵,孫協志
蘇永欽,賴浩敏	黃美珍,賴銘偉	陳一富,戴耀章	黃俊傑,廖溫仁	林淑如,陳珊妮	楊謹華,宋美齡
吳敦義,陳沖	戴佩妮,梁靜茹	羅仲仁,林育豪	簡明仁,冠信	辜寬敏,亞運會	郭柏成,李國泰
王金平,曾永權	吳建豪,楊宗緯	戴家瑞,詹詠引	吳忠泰,楊維哲	李進良,湖人	范植谷,鄭志龍
蘇永欽,賴浩敏	許慧欣,蘇慧倫	李開園,白世輝	四星,黃淑英	顏鴻森,尹祚芊	許水德,劉胡榮

陳進利,陳永祥	蔡依林,蕭敬騰	吳志平,張建邦	莊瑞雄,張桂貞	尹啟銘,曹約文	莊佳容,周平
陳明文,蔡同榮	楊蒨時,詹子晴	黃建逢,江天祐	余怡德,鄭玄	王識賢,辜岳甫	陳奕迅,瑪麗
陳其邁,羅文嘉	李佳薇,黃小琥	呂姜耀強,王明俊	柯明道,謝長峻	陳國,林月雲	美和,呂景義
余政憲,段宜康	馮翊綱,宋少卿	李嗣涔,陳泰然	楊蒨,翁家明	志鵬,陳芳妹	梅若穎,謝霆鋒
許添財,陳明文	梅若穎,高煜玟	包宗和,李嗣涔	康仁俊,施珮君	安妮,周倩	吳靜宜,李耀宗
羅文嘉,段宜康	宋少卿,葛文彬	李嗣涔,張小珍	彭佳慧,白韻琴	林宜,威秀	三商美邦,李駿康
邱議瑩,蔡其昌	張曼玉,梁朝偉	吳妍華,林一平	富士康,徐重仁	蘇永欽,邵恩新	小宇,金尚東
詹春柏,蔣孝嚴	張國榮,梁朝偉	林一平,謝漢萍	廖元豪,黃偉聖	黃震智,謝金河	冠信,曾江
蔣孝嚴,林豐正	關之琳,梅艷芳	吳妍華,許千樹	古煥球,林昭安	經國,夏川里美	李冠儀,吳國楨
曾永權,黃敏惠	章子怡,宋慧喬	吳妍華,裘性天	黃望修,台成	郭添財,蔡惠媚	伊能靜,柳信美
江丙坤,林豐正	劉德華,周星馳	吳妍華,傅恆霖	每日,天主教	吳鳳,林義守	洪百榕,葉歡
馬英九,廖了以	劉德華,謝霆鋒	吳妍華,張翼	施羅德,江蕙	蕭煌奇,林濟川	洪榮宏,麥朝成
廖了以,洪秀柱	范冰冰,謝霆鋒	簡榮宏,曾煜棋	傅賢,帝大	張孝威,辜懷箴	奇摩,鄔禎琳
廖了以,林瑞德	李宇春,劉德華	黃琬姝,簡榮宏	廣順,蘇雲英	毛治國,陳麗珠	楊錫安,廖文奎
廖了以,林鴻池	周星馳,朱茵	黃琬姝,紀珮詩	陳世卿,舒淇	陳國,郭莉莉	黃秋蓮,胡茵夢
沈振來,施崇棠	倪齊民,崔浩然	紀珮詩,白師瑜	鄧麗君,何潤東	余光中,李宗南	洪慶隆,林炳坤
林宗梹,施崇棠	江淑娜,高宇蓁	張明峰,陳登吉	陶坊,李威	戴佩妮,黃忠義	陳美鳳,宋光宇
黃震智,戴正吳	傅子純,何豪傑	陳耀宗,陳玲慧	瓊瑤,蔣孝文	顏聖冠,王龍雄	蔡正元,何壽川
戴正吳,盧松青	徐熙娣,蔡康永	黃煌煇,歐善惠	于洋,羅智	楊蒨時,倪敏然	程顯灝,廖溫仁
徐牧基,遊象富	張芯瑜,趙駿亞	蘇炎坤,詹寶珠	黃國書,陳忠憲	簡嘉宏,林吟蔚	高苑,台成
黃震智,李光陸	王少偉,許孟哲	魏健宏,謝璧妃	明道,李威	張誌家,啟德	李學林,溫賢昌
李光陸,劉燈桂	張菲,黃品源	鄭建鴻,陳文村	陳信安,錢韋杉	陳文鍾,陳秀卿	彭明敏,彭顯鈞
呂芳銘,蔣浩良	安心亞,劉薰愛	余怡德,古煥球	黃國書,簡傳	王振堂,蔡瀚毅	陳子敬,陳詩欣
蔣浩良,程天縱	安心亞,陳漢典	余怡德,王立邦	洪詩涵,陶嘉明	徐旭東,蔣勳	郭榮宗,林惠嘉
郭台銘,帶正吳	王心恬,王婷儀	宋震國,彭明輝	蔣公,孫蓉昌	詹詠引,羅錦龍	洪冬桂,鄭竹梅
魏美玉,徐光曦	何宛庭,何美君	李國賓,林昭安	施啟揚,張亞雯	葉志仙,柯秉逸	張學友,舜子
魏美玉,呂丹虹	白歆惠,洪曉蕾	顏嗣鈞,陳雅淑	賴瑟珍,蔣方智怡	陳信安,李澤藩	伍錦霖,蔡佩玲
呂丹虹,馬進福	隋棠,阮經天	顏嗣鈞,廖婉君	潘威倫,謝錫	白師瑜,金仁寶	台亞,奇摩
馬進福,賴昭銑	隋棠,林志玲	吳家麟,許永真	張學友,楊培安	柏楊,李耀宗	吳清山,楊秋忠
徐光曦,張瑛鶯	賈凡,黃春雄	金仲達,賴尚宏	吳宏偉,牟宗三	麗貞,王郁倫	任偉光,仁寶
邱創興,溫機榮	張學雷,錢一飛	廖建能,許雅三	李國賓,陳恩鐘	黃秀,黃能富	老子,張勻
洪慶隆,張慶年	陳信安,曾文鼎	吳宗憲,曾永華	曹爾忠,楊貴媚	鄭韶婕,張錫銘	清華大華,楊永斌
孫芳珠,張慶年	張學雷,邱宗志	洪茂峰,吳宗憲	趙少康,陳維齡	郭智宏,楊啟東	吳思華,張智星
任子平,陳國根	蔡峻銘,田壘	詹寶珠,曾永華	許榮賢,劉清港	南華,楊英風	胡志強,旺福
陳天祿,陳晴慧	邱大宗,許智超	陳建富,鄭憲宗	劉嘉發,張榮麟	劉姿麟,林慧萍	林煜清,陳宏
張耀明,程名臻	葉志仙,郭泰源	鄭憲宗,謝文峰	釋憲,黃基雄	素顏,柯宇綸	李建億,林毅夫
蔡萬霖,劉鳳嬌	黃志龍,林羿豪	鄭憲宗,曾新穆	馮建三,傅達仁	周湯豪,李烈	謝志偉,陳正

汪國華,董成城 蔡萬霖,錢復	潘威倫,陳冠宇	孫永年,蔣榮先	謝欣霓,劉慶東 朱茵,蔣武童 羅見順,黃勁連 陳泰仁,蔡福財 朱敬一,鍾理和 陳馮富,戴季陶 陳文茜,楊永斌 古煥球,顏鴻森 馮國華,任明廷 高煜玟,蔡垂碧	華誼,李諶 王效蘭,陳子敬 國眾,王耀煌 蔡健雅,林英傑 艾莉絲,易樺 鄧麗君,李娜 吳清基,資安 許仁杰,羅智 達賢,朱浤源 段津華,高政華	汪道淵,陳玉梅 蔡清雲,張鈞甯 劉建國,劉謙 李千娜,唐美雲 許雅三,林理慧 亞運,許晉哲 李羅權,陳永昌 黃秋蓮,萬通 世衛,李瑞斌