國 立 交 通 大 學

電 機 資 訊 學 院

資 訊 科 學 系

博 士 論 文

漢語問句偵測之量化研究

A Quantitative Study on Mandarin Question Detection

研 究 生 ：葉秉哲

指導教授 ：袁賢銘 博士

中 華 民 國 九 十 三 年 七 月

漢語問句偵測之量化研究

A Quantitative Study on
Mandarin Question Detection

研 究 生：葉秉哲　　Student: Ping-Jer Yeh

指導教授：袁賢銘 博士　　Advisor: Dr. Shyan-Ming Yuan

國 立 交 通 大 學

電 機 資 訊 學 院

資 訊 科 學 系

博 士 論 文

A Dissertation
Submitted to Department of Computer and Information Science
College of Electrical Engineering and Computer Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

in
Computer and Information Science
July 2004
Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 三 年 七 月

國立交通大學
研究所博士班
論文口試委員會審定書

本校 資訊科學 系 葉秉哲 君

所提論文 漢語問句偵測之量化研究

合於博士資格水準、業經本委員會評審認可。

口試委員：

陳國棟

朱延平

陳信希

指導教授：

系主任： 教授

中 華 民 國 93 年 7 月 14 日

Department of Computer and Information Science
College of Electrical Engineering and Computer Science
National Chiao Tung University
Hsinchu, Taiwan, R.O.C.

As members of the Final Examination Committee, we certify that we have read the dissertation prepared by Ping-Jer Yeh entitled A Quantitative Study on Mandarin Question Detection

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Dissertation Advisor:

Chairman:

Date: July 14, 2004.

漢 語 問 句 偵 測 之 量 化 研 究

研究生：葉秉哲　　　指導教授：袁賢銘 博士

國立交通大學資訊科學研究所

摘　要

「問句」是日常生活中最為人使用的語言行為之一，在電腦科學裡，舉凡人機對談、機器間對談、標點處理等次領域中，也都扮演著重要角色。少了「問句偵測與處理」此一環節，自然語言處理系統就不算完整。

由於語言本質的差異，再加上傳統上研究重心的不同，漢語的問句偵測要比英語更加困難。有鑑於此，本篇論文鎖定在這個相形之下較為基礎的議題上，並採取量化研究的角度。由於電子化語料資源的限制，本研究暫時只探討詞彙句法層次。

為了解決此一全新議題，本研究的策略是先追求召回率，再追求精確率。在召回率方面，我們先以數種統計推論及樣式比對技術進行單變數分析，成功發掘出較傳統語言學文獻所列更豐富、精確的詞彙特徵。接著我們以白箱式的雙變數分析排除部份誤判情況，以提升精確率。最後我們以數種黑箱式的語言模型技術進行複變數分析，成功分辨出更多情況。

在此研究中，我們達到不錯的召回率及精確率，並在漢語問句偵測議題上開拓一條新的量化研究途徑。

A Quantitative Study on Mandarin Question Detection

Student: Ping-Jer Yeh Advisor: Dr. Shyan-Ming Yuan
Department of Computer and Information Science
National Chiao Tung University

ABSTRACT

Question is one of the most fundamental and frequent speech acts in everyday life. It also plays an important role in sub-areas of computer science such as human-computer and computer-computer communication, and punctuation processing. An NLP application is not complete without proper detection and processing of question.

Detection of Mandarin question is more difficult than that of English due to the nature of the language itself and the research focus in the Mandarin linguistics and NLP field. It is therefore the focus of this research to undertake a quantitative study on the more fundamental problem of detecting Mandarin question. Due to limited electronic resource, the study is confined to lexico-syntactic level.

To tackle this new topic, our strategy is first trying to maximize recall and then to increase precision. To achieve higher recall, we first undertake univariate analysis on the datasets with a variety of statistical inference and pattern matching techniques. At this stage we successfully discover more comprehensive and precise features at word level than what linguistic literature has mentioned before. Next, to increase precision, we undertake white-box bivariate analysis to filter out some false positives from the previous stage. Finally we undertake black-box multivariate analysis by using several language modeling techniques. In this way we successfully discriminate more cases.

We achieve good recall and precision in the preliminary study, and pioneer the quantitative study of Mandarin question.

Keywords: Mandarin question detection, natural language processing (NLP), statistical inference, language models

ACKNOWLEDGEMENTS

千言萬語, 眞不知從何說起。

能完成這篇畢業論文, 首先最要感謝的是我的指導教授袁賢銘博士。沒有您的自由開明作風, 我不可能有機會在研究生涯多方嘗試探險; 沒有您的支持, 我不知還剩多少勇氣在一條人跡罕至的路上踽踽而行; 您的指引, 更是我在猶疑、困惑時的一盞明燈。謝謝您。其次要感謝梁婷教授和曾憲雄教授, 從論文計劃書提案開始, 一直到校外口試的各個階段, 舉凡論文結構、研究取向、技術深度等層面, 都給予我非常多的批評指教, 讓我有機會深入反思某些議題, 使論述更周延。也要感謝校內口試委員: 交大資科施仁忠教授, 校外口試委員: 中興資科朱延平教授、台大資工陳信希教授、中央資工陳國棟教授、東海資科羅文聰教授, 百忙之中給予我的寶貴建議。

在蘊釀論文方向的過程中, 我要格外感謝淸大資工張俊盛教授。旁聽您的「自然語言處理」課程, 開啓了我資訊領域的另一扇窗, 對數學模型背後論據的嚴謹要求, 也讓我對這項利器多了一分掌握。在交大語文所 Charles Lee 教授的「計算語言學」、「語料庫語言學」課堂上, 我看到了從工程領域轉行至語言學領域的榜樣。更要感謝交大語文所劉美君教授, 旁聽您的「句法學」、「詞彙語意學」、「認知語意學」等課程, 趣味橫生, 讓我感受到與理工科系截然不同的氛圍; 在不同背景的激盪下, 我最嚮往的跨領域研究得以逐漸成形。您的引介及鼓勵, 使我有機會在語文所開了兩次語料分析課程, 教學相長, 更深化我對相關議題的思維觀點。

「人獨學而無友, 則孤陋而寡聞」。研究生涯中, 若少了實驗室伙伴們, 眞難想像會是何等光景。張玉山學長、何敏煌學長不時的關照及指導, 加速我的脚步。與梁凱智、許瑞愷、蕭存喩、吳瑞祥、高子漢等人的討論及談心, 與邱繼弘、鄭明俊、彭念劬、陸振恩等人嘻笑玩鬧及組隊競賽, 讓我在實驗室的回憶滿滿。族繁不及備載, 感謝分散式系統實驗室歷年的有緣人。

在我論文口試的周邊事務上, 資科系辦的楊秀琴小姐、俞美珠小姐、陳小翠小姐, 實驗室的邱繼弘學弟、高子漢學弟、林慧雯學妹等人, 多次給予協助, 讓在巨大壓力下顯得有點迷糊的我, 得以順利渡過, 謝謝你們。

焦信達, 我親愛的多年室友 (以及「媒人」, 呵呵), 去年你博士論文誌謝詞裡提到:「在寢

室每晚睡前跟你的閒扯, 總是能帶給我許多知性上的滿足 ......謝謝你開拓了我的視野。」其實, 這句話應該是由我來說的呀。如此知交, 人生能有幾回? 眞懷念過去把盞言歡的宿舍時光: 我醉君復樂, 陶然共忘機 ......

孟君, 感謝你這些日子以來在情感上支持我, 在生活上砥勵我; 讓我學會施與受的眞諦, 讓我補完人生的缺口。這陣子疏於照顧你, 我用更長的歲月來回報。

親愛的父親、母親, 沒有你們, 我不可能沒有後顧之憂地走到現在。養育之恩是報答不完的, 僅以這篇小小的論文, 獻給你們, 以及在天上的祖母。

......

畢業論文完成, 不禁有著和古人同樣的感觸:「衆裡尋他千百度。驀然回首, 那人卻在, 燈火闌珊處。」

最後, 改編電影 The Mummy(神鬼傳奇) 裡面 Imhotep 的一句著名台詞做爲結束: “Doctoral degree is only the beginning.”

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

Symbols

$A, B, C, \ldots$	ordered arrays or associative arrays
$A[i]$	element of ordered array indexed by scalar number i
$A[w]$	element of associative array indexed (mapped) by textual word w
$a, b, c, \ldots$	scalar numbers
F	ANOVA F statistic
$F_{a,b}$	F statistic with df. a in the denominator and b in the numerator
f, g, h	functions
P	probability function
p	probability value
$r, s, t, \ldots$	character strings
$\mathcal{S}, \mathcal{W}, \ldots$	unordered sets (i.e., bags) of possibly duplicated elements
$\lvert\mathcal{S}\rvert$	scalar cardinality of $\mathcal{S}$
w	word
z	alphabet or symbol

Abbreviations

ANOVA	analysis of variance
BOW	Bilingual Ontological Wordnet project
CKIP	Chinese Knowledge Information Processing Group
df. or d.f.	degree of freedom
LLR	log-likelihood ratio
LM	language model
MOE	Ministry of Education, Taiwan
MRD	machine-readable dictionary
NLP	natural language processing
POS	part of speech
QRW	question-related word (invented by the author)
regex	regular expression
SRS	simple random sample/sampling

CHAPTER I

INTRODUCTION

This thesis presents a new topic in the field of natural language processing (NLP): Mandarin question detection. Since no such prior research exists to our knowledge, one may ask two "why" questions:

1. Why is it important in general?

2. Why is it special for the Mandarin Chinese language, in particular?

To address these questions, this chapter first presents the importance of this topic in a broader linguistic and computer science context, not limited to the Mandarin Chinese language. Afterwards we narrow our discussion down to the sole Mandarin field to see why it is still challenging. Finally we outline the overall research plans and results.

1.1 Motivation

The whole research originates from a very simple question. In a classical NLP textbook *Speech and Language Processing* [28, p. 194], there is a paragraph saying that

> There are tasks such as grammar-checking, spelling error detection, or author-identification, for which the location of the punctuation is important … In NLP applications, question-marks are an important cue that someone has asked a question. Punctuation is a useful cue for part-of-speech tagging.

However, it treats the punctuation as a *given* cue. One may then ask: What if the cue is *absent* at all? Perhaps the tasks mentioned above will be confronted with problems.

In what cases can punctuation be absent? To name but a few. In newsgroup writing, punctuation is often misused or missing, especially among the groups crowded with young people. In speech-to-text software, punctuation is usually absent from the generated text.

Among all kinds of punctuation, question marks attract the author's attention. Therefore, the goal of this study can be paraphrased as a question-detection problem.

1.2 The Question-Detection Problem

The question-detection problem is, in short, to enable computers to detect the question parts, if any, within a stream of text or utterance. Its importance is twofold: linguistic and computer science perspectives. This section will first discuss the issue from the linguistic point of view, and then, from the computer science point of view, enumerate applications that can benefit from the study of question-detection problem:

- Human-computer communication.
- Computer-computer communication.
- Punctuation processing.

1.2.1 Question: A Linguistic View

From the linguistic science perspective, the study of speech acts has been a hot topic in discourse analysis, and "question" is one of the major illocutionary acts occurred in everyday life. The deeper our understanding of the nature of a variety of question expressions in particular, the better we may form a computational linguistics model for speech acts in general, which in turn improves application of linguistics.

What do we mean by the term *question*, anyway? In *Glossary of Linguistic Terms* [35], question has two senses:

1. An *illocutionary act* that has a directive illocutionary point of attempting to get the addressee to supply information.

2. A sentence type that has a *form* (labeled interrogative) typically used to express an illocutionary act. It may be actually so used (as a direct illocution), or used rhetorically.

Obviously they reflect two main competitive schools of thought in linguistics: the first addresses the *functional* facet, while the second addresses the *formal* facet. From the functional perspective, the following two cases are both questions in spite of totally different surface forms:

(1) a. Tell me your age.
    b. How old are you?

As for the formal perspective, there are roughly three types of questions: interrogative, dubitative, and rhetorical questions. For example,

(2) a. What is this? *interrogative*
    b. Can such a diligent student fail the school entrance exams? *dubitative*
    c. Don't you understand me? *rhetorical*

1.2.2 Human-computer Communication

As for human-computer communication, a non-toy human-computer dialogue or question answering system needs to distinguish between background information and foreground queries in order to behave more like humans. In such systems, therefore, earlier stages should include at least the question detection module; subsequent processing is fragile without considering it. Now let's examine the two applications in detail.

Question answering (QA) is a fast-growing sub-task of text retrieval. Given a query, it tries to pinpoint the specific answers (noun phrases, sentences, or short passages) rather than just give a pile of relevant documents for you to browse. The QA track of Text REtrieval Conference (TREC) is one of the most famous example. Since the first QA track initiated in 1999 (TREC-8), the has been a lot of progress in this field (see [50, 51, 52, 53, 54]). Participants in this track are required to give an exact answer in response to a *factoid* question, a list of exact answers to a *list* question, and a short passage to a *definition* question. Look at the following excerpts from TREC QA tracks:

(3) a. What is the longest river in the United States? *factoid*

b. Name the highest mountain. *factoid*

c. What are 5 books written by Mary Higgens Clark? *list*

d. List the names of chewing gums. *list*

e. Name 22 cities that have a subway system. *list*

f. Who is Colin Powell? *definition*

g. What are polymers? *definition*

As reported, most QA systems first classify an incoming question into various types of query focus (e.g., quantity, name, time, and place) as suggested by its question word (e.g., what and who) or imperative verb (e.g., list and name); the expected answer types can also be predicted accordingly. Next, some systems attempt a full understanding of the text and then use logic proofs or so to verify candidate answers (e.g., [44]); still others just attempt a shallow, data-driven pattern matching against candidate answers (e.g., [33, 48]).

There is at least one limitation of these QA systems, however. They assume that a QA system receives and recognizes only canonical query forms beginning with a question word or imperative verb. But in reality, not all questions fall into this category. Take the following real-world query for example.[1] Imagine that you are asking a QA system for troubleshooting:

(4) I have installed and configured Wine, but Wine cannot
find MS Windows on my drive. Where did I go wrong?

[1]This paragraph is excerpted from *The Wine FAQ*. URL: `http://www.winehq.com/site/docs/wine-faq/index`.

It is hard to imagine that you are allowed to tell the program only the latter half "Where did I go wrong?" without the former "I have ... on my drive." Even if the unrealistic assumption was made, no program is smart enough to be able to answer the sole question "Where did I go wrong?"— the query focus is correctly identified as "where" but it is of little use here without preceding sentences. What is worse, the query focus "where" may mislead the program to an irrelevant direction of *physical places*! As a result, if the QA program fails to distinguish between foreground query and surrounding context, how can it work out a search plan to answer your "where" question?[2]

Things become even more complicated in dialogue system, in which conversation continues rather than just happens in one round, turn-taking is frequent, and a mixture of various speech acts such as illocutionary and perlocutionary may also be used freely [14]. Since natural conversation switches between both foreground and background expression frequently, it is unrealistic to assume naively that the dialogue system recognizes and accepts only query forms. Take the following excerpt from the novel *Harry Potter and the Sorcerer's Stone* for example. One day Harry Potter said to Hagrid:

(5) Everyone thinks I'm special, ... but I don't know anything about magic at all. How can they expect great things? I'm famous and I can't even remember what I'm famous for. ...

Assume for now that Hagrid is a computer. If Hagrid fails to distinguish between the two, it can never understand what Harry means by "great things" and then work out a search plan accordingly to try to comfort Harry by saying "Don' you worry, Harry. You'll learn fast enough."

[2]One may think that the QA system has a chance to function well if we force users to rephrase their query as "Where did I go wrong when I've installed and configured the Wine but it cannot find MS Windows on my drive?". It may work, but is neither practical nor user-friendly.

1.2.3 Computer-computer Communication

As for computer-computer communication, intelligent agents or software robots may need to travel around the Internet and along the way gather information on behalf of their users. Since XML and semantic webs are still young and there is no universally accepted semantic markup language for unrestricted domains, unstructured documents still dominate the Web. Therefore, a better understanding of speech acts in general and questions in particular may help software analyze unstructured documents and transform them into structured ones.

Furthermore, in multi-agent systems agent communication languages are based mostly on speech act theory (e.g., KQML defines a set of performatives for agents to communicate with [2, 29]) and temporal or first-order predicate logic (e.g., KIF [24]). Many information systems for intra- or inter-business process have also been modeled from the language/action perspective (LAP; see [56] for an overview of LAP and [16, 32] for typical applications). The study of question in natural language settings may help to enhance the expressiveness of communication facilities, finer-grained mental states, and belief-desire model of these systems.

1.2.4 Punctuation Processing

As for punctuation processing, any NLP system is not complete without punctuation processing, but punctuation has been neglected in the NLP field. For example, speech-to-text recognition software maps acoustic signals to text, but it seldom places appropriate punctuation marks in the output text. Word processors have built-in or plug-in spelling and grammar checkers, but they seldom try to check punctuation.

Some literature did recognize the importance of punctuation more or less, as we have seen in Section 1.1. However, it treats the punctuation as a given cue, and does not discuss what if the cue is absent at all.

The reason why punctuation has been neglected is that, it is such a complex coding device that challenges computers. It is, as defined in *The American Heritage Dictionary* [47], "the use of standard marks and signs in writing and printing to separate words into sentences, clauses, and phrases in order to clarify meaning." Therefore, to assign punctuation correctly involves not only syntactic but also semantic and pragmatic levels of processing. Take the following English sentences for example,

(6) a. Is this yours?
 b. What is it?
 c. I beg your pardon?
 d. This is yours? I don't think so.

To punctuate them correctly with question marks, one has to judge whether they are questions. Sentence (6a) is obviously a question because of its verb BE-initial syntactic pattern; the same for Sentence (6b) because of its WH word-initial followed by a verb BE syntactic pattern. Sentence (6c), which begins without a verb BE, an auxiliary verb, or a WH word, is regarded as a question only if the lexical meaning of the word "pardon" is taken into account. Furthermore, Sentence (6d) is regarded as a question only if the pragmatic context is taken into account. Therefore, to be perfect, it is very complicated in general.

1.3 Challenge in Mandarin

It is even more challenging for the Mandarin Chinese language because there is no syntactically decisive and reliable marker and word order in Mandarin question sentences [8], let along decisive and reliable semantic and pragmatic clues. Therefore, mainstream approaches to detecting question sentences developed on the basis of English (and possibly Indo-European languages as well) are not readily applicable here.

Now consider the English language at the syntactic level only. Questions in English have well-understood and consistent patterns, which can be easily found in books or articles on English grammar, e.g., [58]. Patterns in Table 1 are easily

Table 1: A gentle overview of English question patterns, summarized from [58]. For brevity, "AUX" means auxiliary, "SUB" means subject, and "WH" means wh words such as who, why, how, and what. Note that some oral or idiomatic expressions such as "so what?" and "say what?" are not included here

Question Type	Pattern	Example
Inverted sentence		
yes-no question	AUX-initial + SUB	Will John buy a backpack?
non-subject-extracted wh-moved ...	WH + AUX + SUB	
NP complement		What was Beth asked by Diaia?
object of P		Which girl did Beth talk to?
PP		To which girl did Beth talk?
S complement		What does Clove want?
ADJ complement		How did he feel?
do-support	DO-initial + SUB	Did John buy a backpack?
Extraction		
subject wh-question	declarative order	Who wrote the paper?
non-subject-extracted wh-moved question	*see above*	

detected by computers; full-fledged parsers are even unnecessary for this sole task. Take a commonly-available full-fledged link grammar parser from Carnegie Mellon University for example.[3] When a sentence is recognized as question, the leftmost link will be labeled with a `Wq`, `Ws`, `Wj`, or `Q` link type. Therefore, question detection in English is easy.

Things are not quite the same in Mandarin, though. What do we mean in the beginning of this section by the statement that there is no syntactically decisive and reliable marker and word order in Mandarin question sentences? As for the word order, take sentence (7) for example,

(7) a. 這 是 <u>什麼</u> ？

This is what

[3] Software, documentation, and related information of the link grammar parser can be accessed at `http://www.link.cs.cmu.edu/link/`. You can also experiment with the parser on-line.

What is this?

b. 什麼 時候 才 能 再 見面 ？

What time can again meet

What time can me meet again?

c. 你 在 吃 什麼 東西 ？

You be going eat what thing

What are you eating?

The word order of “什麼” (*shénme*; what) appears freely in sentences (7), unlike English. Things become even more complicated that “什麼” alone is not a reliable and decisive marker for question. For example,

(8) a. 你 什麼 東西 都 想 吃 。

You what thing all want eat

You want to eat everything.

b. 我 來 買 點 什麼 吧 。

I come buy CL what

Let me buy something.

To our knowledge, no prior research in Mandarin has focused on exactly the same problem. From the linguistic perspective, traditionally Mandarin linguists discuss question sentences mostly at syntactic level and identify general typology of question expressions (see Section 2.2 for details). Recently researchers have tried to model the Mandarin questions using symbolic approaches such as propositional logic and lambda calculus (see [42] for a brief review). The big picture is very likely to be correct. Not in a corpus-oriented approach, however, they fail to identify more comprehensive and precise features, and lack stronger quantitative evidence.

On the other hand, researchers from NLP and text retrieval fields also have tried to model the Mandarin question-answering problem as semantic frames and ontology [43]. But they all base on an ideal assumption that users issue no sentence other than well-formed questions. Mixed-type cases such as Sentence (4) are beyond their scope of discussion.

It is therefore the focus of this research to undertake a quantitative study on the more fundamental problem of detecting Mandarin question.

1.4 The Scope of This Study

The goal of this study is to enable computers to detect the question parts, if any, within a stream of Mandarin text or utterance. Now we would like to define the scope of this study clearly.

Textual, not prosodic. A declarative sentence may be used to express a question by rising intonation. This study considers only textual rather than prosodic issues.

Lexico-syntactic, not semantic and pragmatic. The author takes the *formal* position instead of *functional* as mentioned in Section 1.2.1 for several reasons. Quantitative studies at semantic and pragmatic levels require many machine-readable resources. Since there is no adequate Mandarin corpus with functional annotation, a quantitavie study in this direction is difficult. As for the formal perspective, quality corpora have punctuation attatched to all sentences types. Among them, interrogative, dubitative, and rhetorical questions are labeled with question marks. Such corpora are readily applicable for this quantitative study. Therefore, at this stage only lexico-syntactic issues are considered in order to narrow down the scope of discussion.

Written, not spoken. Spoken utterance has unique characteristics not equally prominent in written text. To name but a few: conversational filters, ellipsis, and interrupts. They all require additional treatement among utterance. The datasets in use for this study include some transcribed spoken utterance, but the research focus is still on written text. That said, the author thinks that parts of the overall methodology remains roughly the same even for spoken utterance.

Detection, not generation. This study aims to detect Mandarin questions in contemporary use, not to detect superficial cases, nor to generate grammatical utterance. Therefore, construction of a well-formed descriptive grammar is out of the scope of this study. That said, some research results here may provide a basis for a more thorough grammar for Mandarin questions.

1.5 Organization of this Dissertation

The goal of this study is to detect Mandarin question sentences. To put it more concretely, our task, in respect of training and validation, is to label un-punctuated input text with appropriate question marks. This dissertation is organized as follows. Chapter 2 reviews linguistic literature on Mandarin questions. Chapter 3 outlines our overall strategy, rule scheme, and training procedures. Chapter 4 discusses the datasets used in this study, why, and what kinds of pre-processing should be done on them. Chapter 5 describes our feature-selection stage at univariate level and findings, and in the meantime re-examines literature from a different angle: statistical point of view. Chapter 6 describes bivariate and multivariate stages. Chapter 7 discusses our findings and concludes our main contributions.

CHAPTER II

LINGUISTIC BACKGROUND

2.1 Question Marks in Chinese Writing System

Modern Mandarin punctuation system, inspired by the western culture, was stabilized and formalized in the 20^{th} century [57]. Since then, prescriptive guidelines have been announced by authorities in major Mandarin-speaking regions, including Taiwan and mainland China [39, 45]. In general, question marks are used at the end of three kinds of question sentences: interrogative, dubitative, and rhetorical questions. For example,[1]

(9) a. 這是什麼? *interrogative*

What is this?

b. 那麼用功的學生, 會考不上學校? *dubitative*

Can such a diligent student fail the school entrance exams?

c. 難道你還不了解我? *rhetorical*

Don't you understand me?

However, these vague statements touch only superficial mood issues. For rigorous research, we need more information on their linguistic structures.

2.2 Ways to Express Questions in Mandarin

There are, in general, two ways to express questions in Mandarin: prosodic and grammatical devices. It is not necessary for the scope and purpose of this paper to enter into a detailed discussion of the former issue. Therefore we only summarize

[1]When Chinese people write or publish text, they do not separate *characters* with spaces, i.e., *words* are written down consecutively without delimiters as shown in sentences (9). But in the linguistics literature, Chinese words are usually delimited by spaces for the sake of the research community's culture.

intonation patterns from relevant literature. In an interrogative sentence, the focal words are usually stressed and the whole sentence usually ends with a rising intonation. In a dubitative sentence, the focal words are often lengthened, possibly with a high pitch. In a rhetorical question, it is usually spoken with a sustained or falling intonation. Interested readers may consult more literature on this topic, such as (Zhang [62]; Fan [19]).

As for grammatical devices, various classification schemes have been proposed in literature. Some are compiled for educational purpose (Zhang [62]; Liu et al. [34]; Chu [11]), and others are for linguistic research purpose (Li and Thompson [30]; Lyu [37]; Fan [19]; Zhang [61]; Chu and Chi [12]). Some are classified mainly on the basis of morpho-syntactic forms (Li and Thompson [30]; Fan [19]; Chu [11]), some semantic types (Lyu [37]; Liu et al. [34]; Zhang [61]; Chu and Chi [12]), and others pragmatic functions (Zhang [62]).[2] While the big picture is now widely accepted, there is still considerable disagreement about details.

2.3 Exceptions: Question Words and Referentiality

Since no syntactically reliable marker exists in Mandarin question sentences, as mentioned in Section 1.3, exceptions are inevitable. In most of the exceptional cases, the WH words are used as indefinitives or compound relatives (Chu [11]; Chu and Chi [12]). There are roughly 5 cases as pointed out by (Chu [11]; Chu and Chi [12]). The case for indefinitives, as shown in Sentence (10), can be (and possibly can only be) identified from the context since there seems no obvious syntactic pattern. On the other hand, the case for compound relatives can be identified from syntactic patterns, as shown in Sentences (11)–(12).

(10) a. 我 要 幾個 人 幫忙。

[2]The literature review is not meant to be definite. Some of them use hybrid criteria for classifying question sentences since there is no strict dividing line between morpho-syntactic, semantic, and pragmatic issues. Most of them also discuss more than one level of linguistic issues. Here we only point out the most prominent point of view.

Wǒ yào jǐ-ge rén bāngmáng

I need how many people help

I need some people to help me.

b. 我 來 買 點 什麼 吧。

Wǒ lái mǎi diǎn shénme ba

I come buy CL what

Let me buy something.

(11) a. 什麼 事 要 做?

Shénme shì yào zuò

What things need do

What things do we need to do?

b. 沒有 什麼 事 要 做。

Méiyǒu shénme shì yào zuò

Not exist what things need do

Nothing needs to be done.

c. 什麼 事 都 要 做。

Shénme shì dōu yào zuò

What things all need do

Everything needs to be done.

d. 什麼 事 也 要 做。

Shénme shì yě yào zuò

What things also/even need do

Everything needs to be done.

(12) a. 誰 先 到?

Shéi xiān dào

Who first arrive

Who arrived first?

b. 誰 先 到, 誰 先 做。

Shéi xiān dào shéi xiān zuò

Who first arrive who first do

Let those who arrive first do it first.

c. 誰 先 到, 就 (誰) 先 做。

Shéi xiān dào jìu (shéi) xiān zuò

Who first arrive then who first do

Let those who arrive first do it first.

2.4 Exceptions: The Influence of Higher Verbs

In his articles [7, 8] Cheng investigated an interesting issue: higher verbs in a complex sentence may influence the decision whether the proceeding question form is interrogative or not. For example,

(13) a. 這 是 什麼 東西 ?

Zhè shì shénme dōngxī

This is what thing

What is this?

b. 我 來 <u>調查</u> 這 是 什麼 東西 。

Wǒ lái diàochá zhè shì shénme dōngxī

I come investigate this is what thing

Let me investigate what it is.

The verb “調查” (*diàochá*; investigate) in sentence (13b) will turn the question form in sentence (13a) into a non-question.

He concluded in [7] that *inquisitive* and *cognitive* verbs will turn a embedded question form into non-interrogative because the focus is shifted from the question form to the higher verbs; while other types of verbs may or may not have the same effect. However, in his subsequent article [8] *cognitive* verbs were classified into the “may or may not” case without further explanation.

There are still open issues regarding the influence of higher verbs. To name but a few: Is the classification scheme of verb types exhaustive, complete, and accurate? How to explain the exceptions to these higher-verb rules? Is there another theory to explain the phenomena better? We will re-examine parts of this topic in Section 6.1.

CHAPTER III

THE BIG PICTURE: RULE SCHEME AND PROCESS

In this chapter we focus on devising an overall scheme of rules and models to detect Mandarin questions. Based on this scheme, subsequent chapters will then focus on mining features relevant to questions with a variety of technologies.

3.1 Overall Strategy

To approach this task, the overall strategy adopted in this study is first trying to maximize recall and then to increase precision. In many applications recall and precision are two competitive goals. One target at one time makes the whole analysis process more focused, streamlined, and easier for performance tuning.

Another advantage of this recall-first-precision-next route is that, as we progress, we may gain more insight into some facets of question, which may not be discussed in linguistic literature from the same angle or for the same coverage. If we perform a black-box machine learning procedure from the very beginning, we may miss this opportunity. Black-box procedures may also fail to integrate knowledge from a variety of heterogeneous datasets into a seamless model.

With these ideas in mind, we outline the big picture of overall training and detection structure in Figure 1. Next we will describe the overall analysis and detection process.

Levels of Analysis. Three levels of factors are considered in this study. Univariate analysis deals with single word feature, e.g., "什麼" (*shénme*; what) and "如何" (*rúhé*; how). Bivariate analysis deals with the patterns involving two words, e.g., the compound relative "什麼" + "都" case. Multivariate analysis deals with the syntactic patterns involving three or more words.

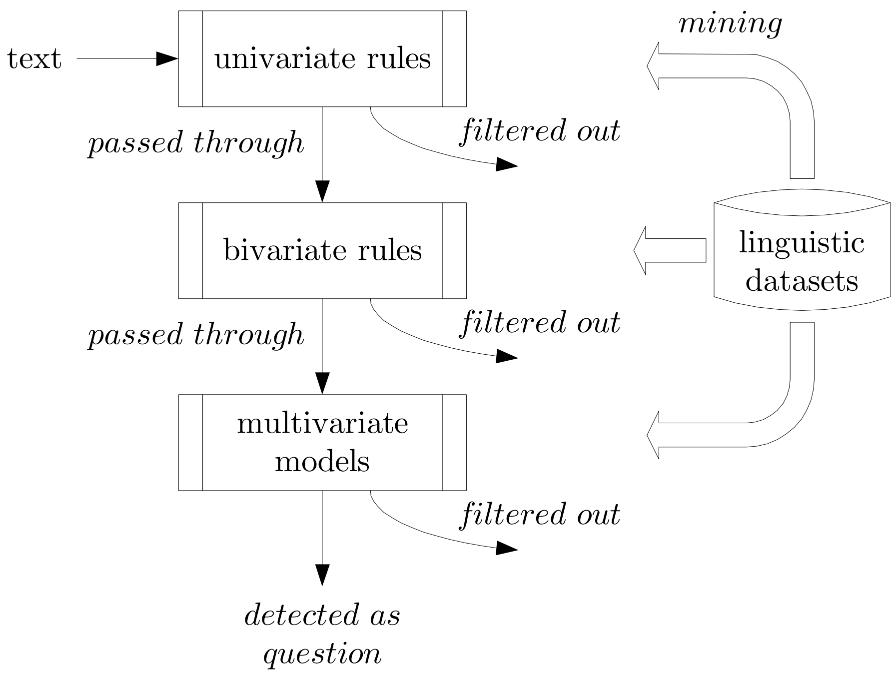


Figure 1: The big picture of overall training and detection structure

Analysis Process. To achieve higher recall, we not only review linguistic literature but also re-examine relevant issues from a new quantitative and corpus point of view, in the hope that more comprehensive and precise features than before will be discovered. Therefore, we prefer the univariate analysis to be a white box rather than a black box.

The next goal is to increase precision without hurting recall too much. As for bivariate analysis, there is still room for white box analysis. As for multivariate analysis, however, white box analysis is difficult since there are still many open issues in linguistics, let alone in NLP field. Therefore, multivariate analysis is done in a black-box approach using probability models.

Detection Process. A sentence input is first analyzed by the univariate module. Since the goal of univariate module is to maximize recall, there may be many false positives. Therefore, both true and false positives will be sent to and re-analyzed by bivariate and then multivariate modules, during which more and more false positives will be filtered out so as to increase precision.

⟨*RuleSet*⟩	::=	⟨*Rule*⟩$^{+}$	▷ disjunction of rules
⟨*Rule*⟩	::=	⟨*PositiveAtom*⟩$^{+}$ ⟨*NegativeAtom*⟩*	▷ conjunction of atoms
⟨*PositiveAtom*⟩	::=	`P` ⟨*PositivePosition*⟩ ⟨*Regex*⟩	
⟨*ExclusiveAtom*⟩	::=	`N` ⟨*ExclusivePosition*⟩ ⟨*Regex*⟩	
⟨*PositivePosition*⟩	::=	`[`	▷ head
	|	`]`	▷ tail
	|	`x`	▷ don't care
	|	`%`	▷ middle
⟨*ExclusivePosition*⟩	::=	`<`	▷ before
	|	`>`	▷ after
	|	`x`	▷ don't care
⟨*Regex*⟩	::=	⟨*any legal Perl 5.8 regular expressions*⟩	

Figure 2: Grammar of question-detection rules at univariate and bivariate level. The quantifier symbol '$*$' attached to nonterminals means "zero or more," and the symbol '$+$' means "one or more"

3.2 Syntax and Semantics of Rules

At lexico-syntactic level, Mandarin questions have a number of characteristics, according to what we have discussed in Section 1.3:

- No reliable and decisive marker.
- No reliable and decisive word order.

In addition, previous studies on Mandarin questions are seldom in a corpus-oriented approach. Therefore, they fail to identify more comprehensive and precise features.

To perform univariate and bivariate analysis, we first define a specification for rule set as a basis for analysis. The syntax of detection rules is listed in Figure 2.

The whole rule set ⟨*RuleSet*⟩ is a disjunction of a series of single rules. Since there are many exceptions in determining questions, each ⟨*Rule*⟩ is composed of a set of positive patterns ⟨*PositiveAtom*⟩ and exclusive patterns ⟨*ExclusiveAtom*⟩ if necessary. The test for a sentence by the rule is passed only when it matches every ⟨*PositiveAtom*⟩ and mismatches every ⟨*ExclusiveAtom*⟩. A positive pattern may appear only in a specific place of the clause, while a exclusive pattern may

```
1 Construct plain rules using QRWs found in Chapter 5
2 while the result does not converge do
3     Train the rules using the training set
4     if the number of false negatives is not acceptable then
5         Investigate if there is any missing feature
6     if the number of false positives is not acceptable then
7         Investigate if the rules are too general
8     Merge similar rule patterns
```

Figure 3: Overall training process of question detection at univariate and bivariate stages

only precede or procede it. Therefore, ⟨*PositiveAtom*⟩ and ⟨*ExclusiveAtom*⟩ have a ⟨*xxxPosition*⟩ field to specify this characteristic.

To handle irregular morphological patterns, we devise the patterns around regular expressions. The advantage of regular expressions is that they make rules more concise and flexible. The disadvantages are that they may over-generalize the patterns and then decrease recall or precision.

As for the syntax of regular expressions (or *regex* for short), we adopt the Perl 5.8 flavor [21] for its expressiveness and popularity. It is also considered the de facto standard in the industry that industrial-strength regex APIs or packages for other programming languages usually claim to be "Perl compatible" to one extent or another instead of compatible with POSIX's flavor.

3.3 The Training Process

In the first two stages (univariate and bivariate analysis), overall training process is iterative, as shown in Figure 3. Steps 5 and 7 are not entirely automatic since for now there remains many sophisticated facets to analyze further. For example, we discover in step 5 many subtle patterns not stated explicitly in linguistic literature before, such as the flexibility in the WH words and the lexeme "何" (*hé*; what). Due to the lack of quality machine-readable dictionaries, these patterns are hardly recognized correctly by machines.

As for the multivariate analysis, we use probability model techniques to try to discriminate questions. We will discuss the details in Section 6.2.

CHAPTER IV

DATASETS: CHOICES AND PREPROCESSING

Since it is the first time to examine this topic in a quantitative approach, at this stage we intend to acquire as accurate knowledge about Mandarin question as possible. Therefore care must be taken to insure the quality of datasets. In this chapter we discuss the reason why these datasets are chosen as our starting point, the mismatch between these datasets and our research needs, and what have to be done in order to bridge the gap.

4.1 The Corpus

The corpus used in this study is Academia Sinica balanced corpus of modern Chinese (or the Sinica corpus for short) developed by the Chinese Knowledge Information Processing Group (CKIP).[1] It comprises about 5 million words, tagged with part-of-speech (POS) information and segmented according to the draft standard in Taiwan. Further details of the corpus can be found in [10].

Clauses are the basic analysis unit used in this study. The corpus divides every complex sentence into clauses that end with commas, periods, colons, semicolons, ellipses, exclamation, or question marks. There are 20,228 question clauses (2.70%) out of total 749,984 clauses. The register distribution of question clauses is listed in Table 2. If we look at the 4^{th} column (q_i/a_i), we may find that questions are more frequent in the oral forms than written, which is quite consistent with our intuition. If we look at the 5^{th} column ($q_i/\sum_{i=1}^{n} q_i$), however, the corpus

[1] The latest public version of the Sinica corpus is 3.0, released on October 1997. Since then, there has been minor fixes on inconsistent formats, tagging, and data cleaning (e.g., about half of the file "`t820902`" was duplicated in the first release of version 3.0; this mistake was corrected in later revisions). The revision used in this study is dated April 19^{th}, 2001.

Table 2: Register distribution of question clauses in the Sinica corpus 3.0

	Clause		q_i/a_i	$q_i/\sum_{i=1}^{n} q_i$
Register	Question: q_i	All: a_i	(%)	(%)
Written	13,821	645,767	2.14	68.33
Written to be read	257	10,315	2.49	1.27
Written to be spoken	1,168	12,736	9.17	5.77
Spoken	4,915	76,470	6.43	24.30
Spoken to be written	55	2,944	1.87	0.27
Unknown	12	1,752	0.68	0.06
Total:	20,228	749,984	2.70	100.00

is biased severely toward the written forms. Therefore, our results may have the same bias, too.

The choice of Sinica corpus, however, restricts us from fuller investigation. The most serious problem is that it is not a treebank. Since there is no hierarchical information available, we cannot handle properly question clauses embedded in complex or compound sentences.

For convenience, we use the format "(*file name* : *serial number of the clause*)" to indicate where the quotation comes from. For example, "(ev7 : 121)" indicates that we quote the clause numbered 121 from the file "`ev7`" in the corpus.

4.2 The Treebank

To investigate some issues in more detail (e.g., "person," see Section 5.2.8), we refer to the CKIP Chinese treebank (or the Sinica treebank for short) as a source of syntactic and semantic information.[2] The treebank, based on a subset of the Sinica corpus as raw material, is bracketed with syntactic hierarchies and annotated with semantic roles according to the information-based case grammar (ICG) developed by the same CKIP team. It currently comprises about 54,902 trees (as claimed on their Web site) and 290,144 words. Further details of the treebank can be found in [4, 5].

[2]The latest public version of the Sinica treebank is 2.1. It can be accessed on-line at `http://treebank.sinica.edu.tw/`.

It is a pity that the Sinica treebank removes punctuation marks altogether, including intra-clause punctuation (e.g., quotation marks and parentheses) and inter-clause punctuation (e.g., periods and commas), eliminating important clues for our research. There is no relevant annotation for us to infer from, either. As a result, we sometimes need to trace these trees back to their origins in the Sinica corpus. Take the following tree numbered 47397 for example,

(14) S(theme:NP(predication:S ・的(head:S(agent:NP(Head:Nhaa: 你)|
Head:VC2: 看)|Head:DE: 的))|evaluation:Dbb: 也|Head:V_11:是|
range:NP(quantifier:DM: 這個|Head:Nab: 月亮)|particle:Td: 嗎)

Its origin in the Sinica corpus is as follows:

(15) 你(Nh) 看(VC) 的(DE) 也(D) 是(SHI)
這(Nep) 個(Nf) 月亮(Na) 嗎(T)
?(QUESTIONCATEGORY) (ev7 : 121)

It can be easily seen that two differences exist. The first is that they segment words differently: the treebank treats "這個" as one word while the corpus two words. The second is that they assign parts of speech differently: The treebank uses a full form (e.g., "Nhaa" for "你" and "Dbb" for "也") while the corpus a simplified form (e.g., "Nh" and "D"); the treebank tags the word "是" as "V_11" while the corpus tags it as a special "SHI" symbol. In case there may be still other differences, the backtracking procedure is performed solely at a level of Chinese *characters* rather than *words* as shown in Figure 4. Note that the regular expression pattern in line 9 is crafted this way in order to handle more complicated formats like the following tree numbered 914:

(16) S(evaluation:Dbb: 究竟|
`<font color="#FF0000">`agent:NP(Head:Nhaa: 你)`</font>`|
epistemics:Dbaa: 是|reason:Dj: 怎麼|
Head:VD1: 分配|particle:Ta: 的)

Let's take a closer at this regular expression pattern. The last symbol "`$`" indicates that the whole pattern is to be matched at the end of the string u. The quantifier meta-character "`*`" means *zero or more* occurrences, "`+`" means *one or more*, and "`?`" means *zero or one*. A pair of parentheses "(" and ")" groups a series of characters and is also ready for field extraction. A pattern of the form "`[` $z_1 z_2 \ldots z_n$ `]`" matches any single character in $z_1, z_2, \ldots, z_n$. On the contrary, a pattern of the form "`[^` $z_1 z_2 \ldots z_n$ `]`" matches any single character *except* for $z_1, z_2, \ldots, z_n$. The backslash "\" is an escape character. Therefore, the first parenthesis group "`([^:\)<]+)`" says that it tries to match (and also extract the content of the underlined part, if successful) a non-empty string composed of any character except for the three symbols `:` `)` `<`. The next quantified parenthesis group "`(<[^>]+>)?`" says that it tries to match an HTML tag, if any. With this carefully-crafted pattern, complicated trees such as Sentences (14) and (16) can be handled gracefully and neatly.

To our surprise, we find in the backtracking process that the textual data of the treebank are not entirely a subset of the Sinica corpus; i.e., some sentences in the treebank are not extracted from the Sinica corpus but elsewhere. Take the tree numbered 39880 for example:

(17) VP(Head:VK1: 希望|goal:S(agent:NP(Head:Nhaa: 妳)|
deontics:Dbab: 能|manner:Dh: 抽空|deixis:Dbab: 去|
Head:VC2: 看看))

The sentence cannot be found in the Sinica corpus. In consequence, some trees cannot be backtracked successfully to check if they are question clauses. These trees are excluded from this study for the sake of objectivity.

Another treebank, also based on the Sinica corpus as raw material and furthermore annotated with HowNet semantic information, does contain punctuation and provide richer semantic information [22, 23]. It currently comprises 3,178 trees and about 36,000 words. However, the sample size is too small to be useful for this study: only 8 trees are relevant to questions! Therefore we do not use this

Algorithm: Finding the punctuation of a tree from the Sinica treebank by tracing its origin back to the Sinica corpus

Input: a tree t in the Sinica treebank format

Output: associated punctuation

Begin:

$\mathcal{S}_{\text{corpus}} \leftarrow$ all clauses in the Sinica corpus,

with part-of-speech tags and delimiters removed

$\triangleright$ Split the tree t into an array of fragments U

$\triangleright$ using the vertical bar "|" as spliting points

$U \leftarrow$ split(t, "|")

$\triangleright$ Extract Chinese characters from U to string r

for each $u \in U$ **do**

  $\triangleright$ From the last ":" to the end (with optional ")" symbols) in u

  Match u against the regex pattern "`:([^:\)<]+)\)*(<[^>]+>)?$`"

  $w \leftarrow$ the first field of match result (underlined part)

  Append w to the end of r

for each $s \in \mathcal{S}_{\text{corpus}}$ **do**

  if r in s **do**

    return the last Chinese character (i.e., punctuation) of s

return not found

Figure 4: Algorithm for finding the punctuation of a tree from the Sinica treebank by tracing its origin back to the Sinica corpus. As for the syntax of regular expressions (or regex for short), we adopt the Perl 5.8 flavor [21] for its expressiveness and popularity

treebank for now.

4.3 Machine-Readable Dictionaries

Lexical semantics have influence on the determination of questions, and therefore quality machine-readable dictionaries (MRDs) would be very helpful in mining such information automatically. In addition, quality dictionaries are proved by experts (often trained in a certain degree of corpus-based lexicography); research based on them may be more accurate than solely on corpora.

The richer information an MRD has for defining and explaining words, the easier and more accurate our research will be. For instance, if an MRD tells us in plain language that the word "貴姓" (*gùixìng*; your last name) is "usually used in *asking questions*," researchers may then try to write programs accordingly to extract such clues. Furthermore, if the MRD is compiled from a modern linguistic perspective, the word may be annotated with more detailed syntactic or pragmatic information in a consistent format for ease of automated processing. For instance, the WordNet [20] (though it only focuses on the English language) annotates the word "why" with "question word," thus simplifying automated search for such interrogative expressions.

Mandarin dictionaries are seldom compiled with a modern lexicology perspective in mind, let alone Mandarin MRDs. The treatment of morphology and pragmatics is severely neglected [13, pp. 3–6]. We choose the on-line installation of the *ABC Chinese-English Dictionary* [15], under the umbrella of the Academia Sinica Bilingual Ontological Wordnet project (or the Sinica BOW for short),[3] as our primary MRD resource. The dictionary, though claimed to comprise about 60,400 words, makes only 32,691 words publicly accessible on the BOW browsing frontend (the other half may actually be on the BOW server too, but inaccessible through the dynamic pages exposed to Web browsers). The search interface at the frontend is not user-friendly—no wildcard search at all! As a result we have

[3]The service can be accessed on-line at `http://bow.sinica.edu.tw/`.

to write programs to gather page by page the list of words accessible, and then use this list to perform further search.

The dictionary translates Chinese words and idioms into equivalent English words or phrases. Although it is quite simple that no pronunciation, examples, usage notes, etc. is available, it does provide one important clue for this study: question marks. Take the words "何時" (*héshí*) and "何以" (*héyǐ*) for example, they are translated by the dictionary as "when?" and "how?; why?" respectively (notice the question marks). Given this feature, we can write programs to gather all translation entries containing the question marks as our starting point. In this way, there are totally 37 words found to be related to questions, if part-of-speech is also considered.

The disadvantage of this dictionary is that its coverage of words is too small, compared to CKIP's *Chinese Electronic Dictionary* (about 80,000 words) or even open lexicons such as libtabe (about 137,000 words; see [26]) and EZ Input big lexicon (about 100,000 words; see [18]).[4] The larger a lexicon is, the better chance we may have to extract useful information.

The Sinica BOW provides other machine-readable dictionaries as well, but they are too small in size, in under-construction or restricted-use state (e.g., CKIP's *Chinese Electronic Dictionary* and Lyu's *Eight Hundred Words of Modern Mandarin*) and/or not qualified enough for this kind of linguistic research (e.g., MOE's *Mandarin Dictionary Revised*). Therefore they are excluded from this study.

Among them, the MOE's *Mandarin Dictionary Revised* is worth a closer look. In fact, the dictionary service at BOW makes use of merely a subset of the original database. The official site for this dictionary [40] provides much larger coverage (about 166,193 words at present), richer information, and better search interface than the subset one installed at BOW. Compiled from a more traditional lexicography perspective, it provides no modern tagging or annotation system and

[4] Dr. Tsai compiles a list of lexicons available on the Internet [49]. Most of them are free or licensed as open source software. Not for linguistic purpose, though, they can still give us a rough estimate of appropriate coverage a practical Mandarin lexicon should have.

therefore is not easy to analyze automatically. That said, it does contain something useful for this study. It is therefore chosen as the auxiliary MRD in this study.[5]

In conclusion, the primary MRD in this study is the on-line installation of the *ABC Chinese-English Dictionary* at Sinica BOW site, and the auxiliary MRD is the official site for MOE's *Mandarin Dictionary Revised*.

4.4 Other Non-Electronic Resources

Sometimes it is inevitable to consult more comprehensive resources other than electronic ones about some linguistic issues. For example, we use the *Unabridged Mandarin Dictionary* [36], *Unabridged Dictionary of Chinese Characters* [59], and *Eight Hundred Words of Modern Mandarin* [38] to explore more similar cases for a certain kind of lexical semantics. Since they are not in electronic forms, exhaustive search is impossible unless plenty of labor is available.

[5]The examples in this dictionary were once considered as another source of corpus for this study. But too many quotations from ancient classics make them inappropriate here.

CHAPTER V

UNIVARIATE ANALYSIS

Our overall machine learning strategy is first trying to increase recall and then precision, as has been stated in Chapter 3. To maximize recall, we need to discover all features that may constitute a question. In Chapter 2 we have reviewed linguistic literature on Mandarin question forms, but the literature does not stand on a corpus and statistical basis. To be useful in statistical NLP methodology, however, a quantitative investigation is necessary. Another reason to conduct a quantitative survey is that the features listed in literature are neither comprehensive nor precise enough for NLP purpose. In this chapter, therefore, we re-examine several issues in quantitative point of view. It should be noted that the main purpose is to pave the way for devising programmable rules and heuristics. The fuller linguistic and qualitative study of them is beyond the scope of this research.

5.1 *Finding Question-Related Words*

As a beginning, we will examine what set of words constitutes a question sentence in a somewhat context-free manner. These "question-related words" (hereafter, QRWs) may be content words or particles. We coin the term "QRW" in order to avoid confusion with another term used frequently in linguistic literature: "question words" [30, 11], which should mean the interrogative words or WH-words (e.g., what, which, who). The set of Mandarin interrogative words is therefore a subset of QRWs.

5.1.1 Procedure

To find QRWs in a quantitative approach, the question-delection problem should be modeled first as a statistical form suitable for identifying and ranking univariate features. Since they are categorical variables, we model the problem as a statistical problem: test-of-independence of two dimensions of factors. One dimension is whether a word w_i under consideration is in a sentence s_j under consideration, and the other is whether the sentence s_j is a question. Modeled in this way, the word w_i = "什麼" (*shénme*; what) may have the following four cases:

(18) a. w_i is in a question sentence

到底(D) 什麼(Nep) 才(Da) 算是(VG) 藝術品(Na)？

b. w_i is in a non-question sentence

無論(Cbb) 發生(VJ) 了(Di) 什麼(Nep) 大(VH) 事(Na)，

c. w_i is not in a question sentence

你(Nh) 相信(VK) 嗎(T)？

d. w_i is not in a non-question sentence

再(D) 大聲(VH) 也(D) 無用(VH)。

Based on these four observations, one may undertake statistical inference procedures to test if and to what degree w_i is independent of questions.

To undertake any statistical inference, one needs to calculate, for each w_i candidate in the corpus, the number of occurrence of the four cases in sentence (18), and they can be arranged in a 2×2 contingency table (see Table 3) with 4 cells a, b, c, and d. The algorithm in Figure 5 will then generate the four variables, undertake statistical inference, and rank the results. Lines 1–2 initialize the unordered sets $\mathcal{S}_{\mathrm{Q}}$ and $\mathcal{S}_{\mathrm{NQ}}$ to store all question and non-question clauses, respectively. Line 3 initializes the unordered set $\mathcal{W}$ to hold all QRW candidates. The main loop of the algorithm in lines 4–12 iterates through each word $w_i \in \mathcal{W}$ to compute its statistic.

The framework is so general that a variety of statistical procedures can be applied. Here we apply two kinds of test procedures which have solid mathematical foundation in the field of inferential statistics. Regarding the two procedures for asymptotic χ^2 distribution, some state that the log-likelihood ratio (LLR for short) is better at sparse data [17] while others state that the Pearson's chi-square (χ^2) test is better at smaller n and more sparse tables (for literature review for this

Table 3: 2×2 contingency table for finding question-related words (QRWs)

Clauses	Is w_i in the clause? Yes	No
Ends with '?'	a	b
Ends without '?'	c	d

where $w_i \in \{\text{all words in the corpus}\}$

intermediate values:

$$
\begin{aligned}
n &= a+b+c+d \\
m_a &= (a+b)(a+c) \\
m_b &= (a+b)(b+d) \\
m_c &= (a+c)(c+d) \\
m_d &= (b+d)(c+d)
\end{aligned}
$$

and final statistics of w_i :

$$
\begin{aligned}
\text{LLR statistic} &= 2 \times \sum_{j=a}^{d} j \ln \frac{n \times j}{m_j} \\
\chi^2 \text{ statistic} &= \frac{n(ad-bc)^2}{m_a m_d} \\
\text{Frequency} &= a+c \\
\text{Precision} &= a/(a+c) \\
\text{Recall} &= a/(a+b)
\end{aligned}
$$

Algorithm: Finding question-related words from the corpus

Input: corpus

Output: associative array $C[w_1, \ldots, w_n]$ mapping from w_i to statistic of interest, where $n = |\mathcal{W}|$ and $i = 1, \ldots, n$

Begin:

$\mathcal{S}_{\mathrm{Q}} \leftarrow$ all question clauses in the corpus
$\mathcal{S}_{\mathrm{NQ}} \leftarrow$ all non-question clauses in the corpus
$\mathcal{W} \leftarrow$ all unique words in $\mathcal{S}_{\mathrm{Q}}$
for each $w_i \in \mathcal{W}$ **do**
    $a, b, c, d \leftarrow 0$
    for each $s \in \mathcal{S}_{\mathrm{Q}}$ **do**
        if w_i in s **then** $++a$
        else $++b$
    for each $t \in \mathcal{S}_{\mathrm{NQ}}$ **do**
        if w_i in t **then** $++c$
        else $++d$
    $C[w_i] \leftarrow$ compute statistic of interest for w_i via a, b, c, d
Sort C in descending order

Figure 5: Algorithm for finding question-related words

point, see [1, pp. 24, 395–397]). For completeness and comparison, both are used in this study.

It has been reported in [46, 55] that Fisher's exact test for this task is better at dealing with sparse data than some other ways to approximate theoretical χ^2 distribution such as Pearson's χ^2 test or LLR test. However, it requires a lot of computation in hypergeometric space and factorials, especially as large as factorial 749,887:

$$\begin{aligned}
\Theta &\stackrel{\text{def}}{=} \text{set of configuration from this to the most extreme case} \\
P(w_i) &= \sum_{\Theta} \frac{(a+b)!\,(a+c)!\,(b+d)!\,(c+d)!}{a!\,b!\,c!\,d!\,n!} \\
P_1(w_i) &= \sum_{\Theta} \frac{(a+b)!\,(a+c)!\,(b+d)!\,(c+d)!}{a!\,b!\,c!\,d!} \qquad \text{since } n! \text{ remains constant}
\end{aligned}$$

Even with the help of Stirling's formula:

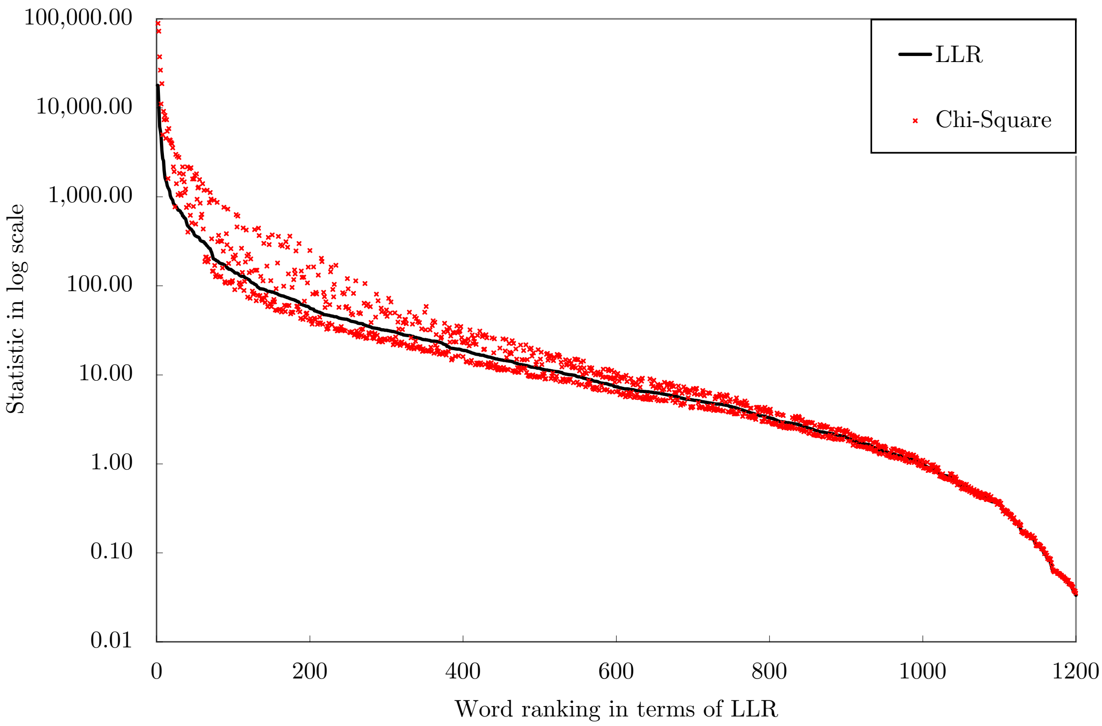
$$\begin{aligned}
x! &\sim \sqrt{2\pi}\,x^{x+0.5}\,e^{-x} \qquad \text{for } x \text{ large} \\
&= \sqrt{2\pi}\,e^{(x+0.5)\ln x - x}
\end{aligned}$$

it is still too large for a long double floating point to handle. It is therefore not very practical here.

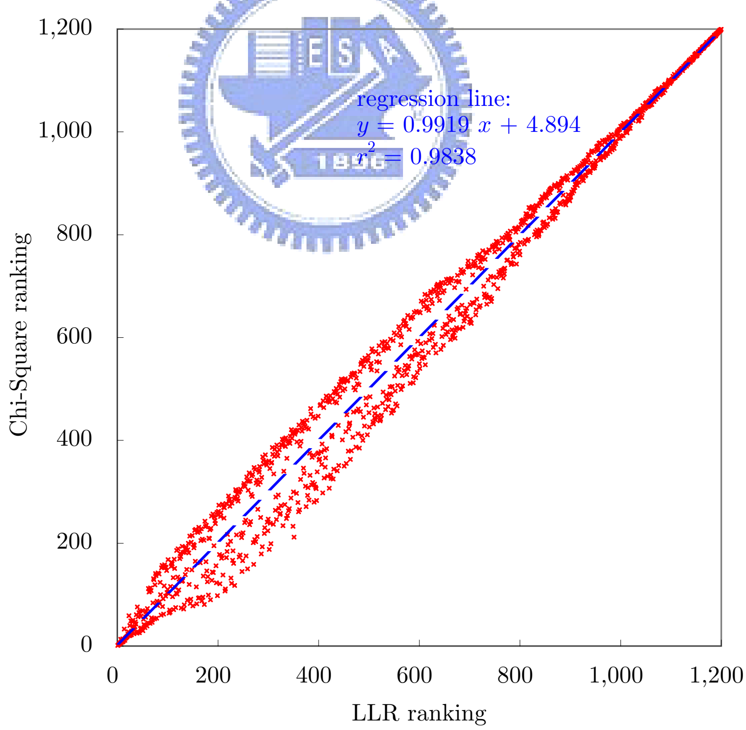
For brevity, top 40 results are listed here in Table 4, and more details can be found in Appendix A.

There are some disagreements about the rankings and statistics in the two tests. Looking at the comparison chart in Figure 6, however, the overall trend remains the same, and converges when the ranking is greater than about 1000. The correlation coefficient $r = +0.9919$ in Figure 6b further suggests a very strong association between both ranking schemes. Therefore, we will refer to the ranking in terms of LLR unless mentioned explicitly.

At first glance both the statistics for LLR and χ^2 in Table 4 seems too large. The reason is that, given a w_i, the "No" column in Table 3 may contain something that acts similar to the "Yes" column; such lurking variables have side effects that falsely magnify the statistic. Therefore, a higher χ^2 critical value (i.e., a lower Type



(a) Ranking versus statistics of LLR and Chi-Square.



(b) Ranking of LLR versus Chi-Square.

Figure 6: Comparison between LLR and Pearson's χ^2 tests on QRWs. In (a) the X-axis is arranged in terms of LLR ranking. The Y-axis shows statistics of LLR and χ^2 respectively in logarithmic scale. Here the statistics of Pearson's χ^2 test tend to be larger than that of LLR, but converge in the long run. In (b) both axes are arranged in terms of LLR and χ^2 respectively. The dashed regression line and a correlation coefficient $r = +0.9919$ suggest a very strong association between the two rankings

Table 4: Top 40 question-related words (QRWs) found by statistical inference procedures

Ranking		QRW	Statistic		Ranking		QRW	Statistic	
LLR	χ^2	w_i	LLR	χ^2	LLR	χ^2	w_i	LLR	χ^2
1	1	嗎(T)	17,956.52	88,941.79	21	23	你們(Nh)	915.39	2,173.73
2	2	呢(T)	17,798.44	72,731.19	22	28	吧(T)	851.01	1,919.64
3	3	什麼(Nep)	11,223.98	37,515.11	23	59	的(DE)	813.92	772.95
4	4	爲什麼(D)	6,163.42	26,622.68	24	20	爲何(D)	807.60	3,003.16
5	6	你(Nh)	5,464.47	11,060.08	25	40	知道(VK)	772.87	1,404.77
6	5	怎麼(D)	4,776.90	18,747.55	26	21	如何(VH)	772.15	2,857.26
7	13	不(D)	3,320.25	4,982.97	27	49	會(D)	741.07	1,057.46
8	7	誰(Nh)	2,685.79	9,161.25	28	22	會不會(D)	711.35	2,774.36
9	10	如何(D)	2,548.63	7,369.35	29	29	多少(Neqa)	709.86	1,844.61
10	8	到底(D)	1,998.58	8,221.60	30	48	又(D)	702.45	1,097.02
11	14	是否(D)	1,653.17	4,540.90	31	50	還(D)	691.14	1,036.92
12	9	怎麼辦(VH)	1,549.34	7,372.23	32	37	妳(Nh)	658.02	1,558.61
13	12	怎麼樣(VH)	1,454.17	5,464.69	33	24	還是(Caa)	652.50	2,169.88
14	33	是(SHI)	1,342.47	1,605.50	34	31	樣(Nf)	613.09	1,808.57
15	11	難道(D)	1,265.77	5,861.16	35	38	該(D)	609.59	1,471.49
16	15	哪(Nep)	1,226.55	4,376.40	36	47	對(VH)	581.03	1,107.65
17	16	何(Nes)	1,154.87	4,307.55	37	43	您(Nh)	580.09	1,223.44
18	17	哪裡(Ncd)	1,033.09	4,086.19	38	61	要(D)	564.10	761.19
19	18	有沒有(D)	973.16	3,959.01	39	25	怎麼(VH)	517.82	2,169.03
20	19	究竟(D)	944.68	3,554.16	40	76	在(P)	482.15	402.42

I error probability α) is required to claim that the result is significant. However, the raw statistic is not important at this stage, and we only refer to the statistic in terms of rankings in this section.[1]

5.1.2 Coverage Test in Terms of Recall

Before going any further, we would like to stop for a while to validate the validity of these QRWs in two ways: one is by quantitative analysis inside the corpus itself, and the other is by the MRD and a little qualitative analysis.

First, we would like to verify if these QRWs (especially those with top ranking) really cover most of the question cases in the corpus. To do this, let's examine them in terms of recall. We use the procedure outlined in Figure 7 to calculate cumulative recall of these QRWs in ascending order of their ranking; the result is shown in Figure 8.

[1]Note that in many NLP applications, as Manning and Schütze [41, p. 166] pointed out, "the level of significance itself is less useful . . . All that is used is the scores and the resulting ranking."

Algorithm: Calculating cumulative recall and precision of QRWs
Input: corpus
Output: cumulative recall and precision for $\mathcal{W}_i, i = 1, \ldots, n$
Def: $w_i \stackrel{\text{def}}{=}$ the QRW ranked i-th
$\mathcal{W}_i \equiv \{w_1, \ldots, w_i\}$
Begin:
$\mathcal{S}_{\text{Q}} \leftarrow$ all question clauses in the corpus
$\mathcal{S}_{\text{NQ}} \leftarrow$ all non-question clauses in the corpus
for $i = 1, \ldots, n$ **do**
  $a, b, c, d \leftarrow 0$
  for each $s \in \mathcal{S}_{\text{Q}}$ **do**
    if s contains any word in $\mathcal{W}_i$ **then** $++a$
    else $++b$
  for each $t \in \mathcal{S}_{\text{NQ}}$ **do**
    if t contains any word in $\mathcal{W}_i$ **then** $++c$
    else $++d$
  Compute and print recall and precision for $\mathcal{W}_i$

Figure 7: Algorithm for calculating cumulative recall and precision of QRWs

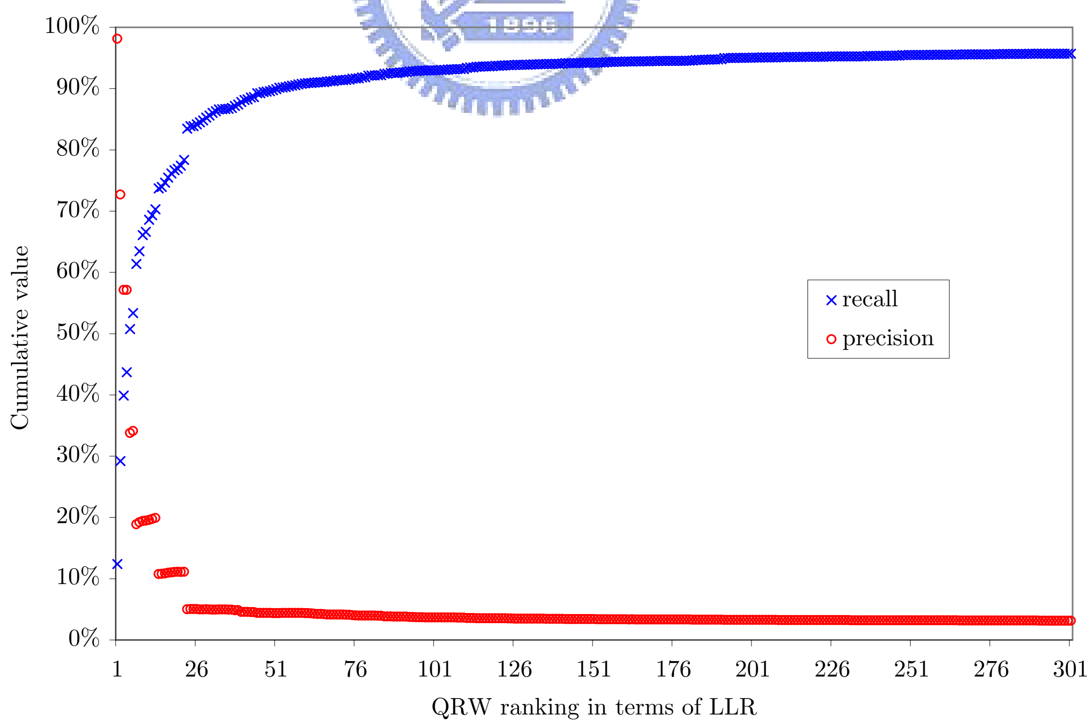


Figure 8: Cumulative recall and precision of question-related words

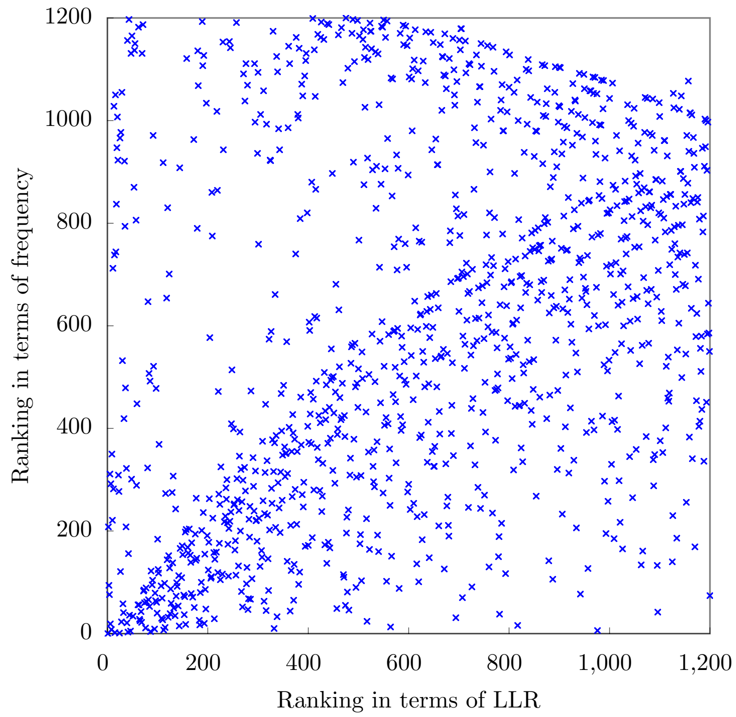


Figure 9: QRW ranking in terms of LLR vs. in terms of frequency. The correlation coefficient $r = +0.2372$ is so weak that there is little association between the two ranking

Figure 8 shows something interesting. In theory, if we accept all top n QRWs as our features, say $n = 200$, we may reach a high recall up to 95%. Though the goal at this stage is to maximize the recall, but pursuing this goal blindly may fall into the trap of overfitting. Roughly speaking, the amount of noise increases steeply when the ranking is greater than about 130. Another reason that forbids us to maximize the recall is that precision drops steeply as the recall increases.

One may also suspect that the high cumulative recall is due primarily to the effect of Zipf's law: they are merely high-frequency words. To check this, we try to calculate the strength and direction of association between these two rankings: LLR and frequency. As Figure 9 shows, the correlation coefficient $r = +0.2372$ implies the association is so weak that the high cumulative recall should not be explained in terms of frequency and the effect of Zipf's law.

5.1.3 Coverage Test by Dictionaries

Next we would validate the validity of these QRWs by the MRD we choose: the *ABC Chinese-English Dictionary*. First, we try to verify if the QRW list covers every word generated by the procedure described in Section 4.3; if some are absent, they may be added to the lexicon of our rules for completeness. Table 5 shows the result in 3 parts: match, rare occurrence, and suspect cases.

Table 5a shows 13 words which are also located in the QRW list, and most of them rank very high. Quite a good result. That said, the last 3 abnormal cases "幾," "可是," and "嗯" need further discussion.

One may wonder why the word "幾" (*jǐ*; how many) is ranked as low as 371^{st} in the previous QRW list. However, another similar but compound QRW "第幾" (ranked 233^{rd}) does not exist in the dictionary. Therefore, the low ranking of this "幾" case can be regarded as due to morphological differences.

The word "可是" (*kěshì*; but) has 3 senses in the dictionary: (1) but; yet; however, (2) Is it that . . . ? (3) be indeed. Consulting more authoritative dictionaries, it has 7 senses in the *Unabridged Mandarin Dictionary* [36], and the third sense tells us that it is "similar to '是否' (*shìfǒu*; yes or no?)." In *Eight Hundred Words of Modern Mandarin*[38], it has a sense to emphasize rhetorical questions (usually in oral situation). Nevertheless, if we look at the Sinica corpus for statistical evidence, we may find that "可是" occurs 2,482 times in the corpus and 98 times within question clauses, but none falls into such sense! It is therefore doubtful whether the sense is still common at present.

The interjection "嗯" (*en*) is defined in the dictionary as "What?; Huh?" but it is "a nasal sound" used mostly for "responding to a call" [60]. Therefore, the word is considered relevant to question only in a certain situation.

Table 5b shows 19 words which are not located in the QRW list since their occurrence in the corpus is too rare to be considered significant. Among them, two cases are worth discussion. The word "不成" (*bùchéng*) has 5 cases in the

Table 5: Using the words extracted from the *ABC Chinese-English Dictionary* to validate QRWs. The word with a symbol † means that it has multiple senses and the only one sense relevant to question is rarely used today based on the author's introspection; ‡ means that generally speaking it is irrelevant to question based on the author's introspection; * means it needs further discussion in the paper. The meanings of frequency counters a and c are the same as in Table 3

(a) Match cases

Word	QRW ranking	Word	QRW ranking	Word	QRW ranking
哪	16	何以	116	幾*	371
何	17	豈	161	可是†	476
還	31	何時	188	嗯*	985
怎樣	48	能否	191		
何處	104	抑或	204		

(b) Rare occurrence cases

Word	Freq. $a+c$	English translation
幾時	12	what time?; when?
何許	7	what kind of; what?
何如	6	(1) how about? (2) wouldn't it be better?
何故	5	why?; for what reason?
豈有此理	4	What kind of reasoning is that?; Nonsense!
借問	3	may I ask?
不成(C)*	3	(2) can it be that?
幾多	2	how many/much?
幾何	2	(1) geometry (2) how much/many?
豈敢	2	how dare?
什	1	what?
幾兒*	0	which day of the month?
何人	0	who?
何敢	0	how dare?; dare not
若何	0	How then?; What then?
哪門子	0	why?; who?; what?
干啥	0	How come?; Why?
犯得上	0	Is it worthwhile? (implying not)
犯得著	0	Is it worthwhile? (implying not)

(c) Suspect cases

Word	Frequency a	Frequency c	English translation
若干‡	1	213	(1) a certain number/amount (2) how many?
好多(P)‡	1	169	(2) how many?; how much?
何其†	1	27	how?; what?
奈何(P)†	0	11	(2) why?; for what reason?
幾許†	0	16	how much/many?

Unabridged Mandarin Dictionary, and the 5^{th} sense is used as a particle for rhetorical questions. Consequently, the word is considered relevant to question only in a certain situation, which requires word sense disambiguation to successfully distinguish. Another word “幾兒” (*jǐ'ér*) is not used in contemporary Mandarin (based on popular medium-sized dictionaries and the author's introspection) but only in the past or in a certain dialects. For example, the *Unabridged Mandarin Dictionary* traces its origin back to *The Dream of the Red Chamber*, a classic novel in the mid-18^{th} century during the Qing Dynasty.

Table 5c shows 5 words which are counter to the evidence provided by the corpus. Not only the quantitative evidence, a little qualitative study also disagrees with the *ABC Chinese-English Dictionary*. Among them, one may wonder why the word “好多” (*hǎoduō*; a lot of) ever has such a sense as “how many?; how much?”. The *Unabridged Mandarin Dictionary* again tells us that this suspect sense comes from a dialect (Beijing dialect, I guess). Therefore, it is safe to exclude this word from the QRW list in ordinary situations.

To sum up, the QRW list found so far has covered prominent information.

5.2 QRW Classification and Exploration

One drawback of statistical methodology is the risk of sampling errors. Since it is impossible to obtain a census of language utterances, some features may not be sampled enough in the datasets. Another drawback of statistical inference is the risk of Type I and II errors. Since uncertainty occurs almost everywhere, some features may be absent and misfeatures may be present just by chance. Therefore the results found so far cannot be accepted as is. Instead, we may use them as a seed to explore more cases.

On closer inspection, the QRW list found so far reveals something not addressed explicitly in linguistic literature, and also reveals some errors in contemporary NLP datasets and programs. Therefore we shall now look more carefully into a number of issues, and at the same time classify the QRWs into manageable groups.

5.2.1 Particles and Interjections

A declarative sentence can usually be turned into a question simply by appending a particle to the end. Literature on linguistics (e.g., Li and Thompson [30], Liu et al. [34], Zhang [61], Chu [11], Chu and Chi [12]) has identified several particles for this: "嗎" (*ma*, ranked 1st), "呢" (*ne*, ranked 2nd), "吧" (*ba*, ranked 22nd), "啊" (*a*, ranked 41st), though linguistic details disagree among literature.

Here we would like to undertake a quantitative survey to find if there are still other question sentence-final particles. To do this, we first use a two-word window to break all clauses (including punctuation at the end) in the corpus into 5,806,392 bigrams, and then perform an LLR test on these bigrams. Next, we sort and rank them in terms of LLR statistic, and then extract all bigrams of the form "word + ?" to see which particles co-occur most frequently with question marks. Since interjections have similar characteristics, they are also included in this survey. The result is shown in Table 6.

Based on these finding, we choose the following sentence-final particles and interjections as parts of our final QRW list.

- Normal cases: 嗎(T), 呢(T), 麼(T), 乎(T), 哪(T), 嘸(T), 否(T).
- Perfective aspects: 沒有(T), 沒(T).
- Ancient literary cases: 何(T), 邪(T).
- Ambiguous cases: 啊(T), 啊(I), 吧(T), 喔(T), 哦(T), 咦(I).

Here the greatest difficulty in deciding which is truely related to question is pragmatic issues. The same particles and interjections can also perform euphemism, irony, exclamation, or any other illocutionary act. This is obviously beyond the extent of lexico-syntactic level. Further analysis will be in Section 6.2.1.

Table 6: Sentence-final particles and interjections co-occurring with questions in the Sinica corpus, using bigram analysis with a window size = 2. The column "$w_1 w_2$" lists the ranking of the bigram "w + ?" among every possible "$w_1 + w_2$" bigram combination. The column "w_1 ?" lists the ranking of the bigram "w + ?" among every possible "w_1 + ?" bigram combination.

Word	LLR Ranking		LLR		Word	LLR Ranking		LLR	
w	w_1 ?	$w_1 w_2$	statistic	Count	w	w_1 ?	$w_1 w_2$	statistic	Count
呢(T)	1	14	26620.8898	2910	哦(T)	160	122484	18.6565	8
嗎(T)	2	17	22918.6034	2143	邪(T)	172	127308	17.7048	2
吧(T)	4	541	1752.7357	291	也(T)	207	140251	15.3693	6
呀(T)	6	973	1138.0415	208	來(T)	265	161195	12.3196	12
喔(T)	10	1333	883.7391	124	吶(T)	604	194699	8.4555	2
啊(T)	11	1488	808.7852	196	來著(T)	659	200807	7.8119	1
麼(T)	13	1691	722.7873	68	得(T)	842	212156	6.6389	1
的(T)	15	1839	679.4023	278	耳(T)	858	212156	6.6389	1
了(T)	18	2355	560.9355	288	喂(I)	951	217339	6.1114	2
沒有(T)	21	4185	350.3261	41	啊呀(I)	1056	222056	5.6327	1
啊(I)	23	4672	319.2967	66	天哪(I)	1403	233330	4.4921	1
啦(T)	33	12810	143.2852	46	咧(T)	1492	235955	4.2277	1
囉(T)	41	21671	93.2888	15	欸(I)	1801	244325	3.3836	4
喔(I)	47	25211	82.5035	22	哦(I)	2028	249101	2.9025	3
沒(T)	48	27134	77.7024	8	而已(T)	2030	249560	2.8561	5
哪(T)	49	27393	77.1720	19	嘍(T)	2093	250699	2.7413	1
去(T)	58	35309	62.4553	23	焉(T)	2180	253097	2.5007	1
乎(T)	60	35889	61.6109	9	哇(I)	2211	253792	2.4306	2
不(T)	61	36481	60.7300	6	哩(T)	2864	268327	0.9719	1
哇(T)	71	43432	52.3184	11	與否(T)	3134	272665	0.5374	1
咦(I)	72	45017	50.6755	8	耶(T)	3505	276784	0.1255	1
否(T)	75	47198	48.6054	6					
罷(T)	76	47476	48.3612	7					
哉(T)	77	49095	46.9225	7					
Huh(I)	79	49622	46.4764	4					
嗯(I)	97	63569	36.9225	21					
何(T)	101	67632	34.8572	3					
嘛(T)	114	81540	29.2220	16					
與(T)	117	84805	28.1391	3					
嘸(T)	128	102090	23.2380	2					

Note also that in practice, the POS tags may not be used as is in the univariate detection module due to differences or errors of taggers. Take two publicly-available Mandarin taggers for experiment, the Autotag 1.0 from CKIP[2] and ICTCLAS 2.0 from Institute of Computing Technology, Chinese Academic of Sciences, China:[3]

(19) a. 到了沒有?

Sinica corpus: 到(VCL) 了(Di) 沒有(T) ? (k811211 : 891)

Autotag: 到(VE) 了(Di) 沒有(VJ) ?

ICTCLAS: 到/v 了/u 沒有/v ?

The two taggers incorrectly treat this "沒有" as a verb. Therefore, a beneficial side effect of this particle-final study is to improve the quality of taggers.

5.2.2 Inconsistent Segmentation of A-not-A Questions

The Sinica corpus does not segment words in a purely consistent manner. Take the Mandarin alternative or disjunctive question form "A-not-A" for example. In some places the ranked 28^{th} entry "會不會" (*hùibúhùi*; capable or not) is treated as one word, while in other places it is segmented into 3 individual words (characters). Consider the following sentences:

(20) a. 開刀(VB) 會(D) 不(D) 會(D) 痛(VH) ? (f80013a : 1821)

b. 你(Nh) 會不會(D) 跳進去(VCL) ? (bbai : 5484)

According to the draft segmentation standard in Taiwan [27] and annotation guideline for the Sinica corpus [10, p. 19], sentence (20a) is segmented incorrectly. Similar inconsistent cases in the corpus, to name but a few, include "好不好" (*hǎobùhǎo*; good/agree or not, ranked 51^{st}), "要不要" (*yàobúyào*; want or not, ranked 70^{th}), and "可不可以" (*kěbùkěyǐ*; can or cannot, ranked 74^{th}).

[2] CKIP Autotag: executive files are available at `http://rocling.iis.sinica.edu.tw/CKIP/ws/`.

[3] ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System): documentation, technical reports, and source code are available at `http://mtgroup.ict.ac.cn/~zhp/ICTCLAS/`.

The inconsistent segmentation in the Sinica corpus raises some problems in subsequent automatic analysis. What is worse, different segmentation tools and taggers may treat them differently. For example,

(21) a. 可不可行

Autotag: 可不可行(VH)

ICTCLAS: 可/v 不/d 可行/a

b. 可不可做

Autotag: 可(D) 不可(D) 做(VC)

ICTCLAS: 可/v 不可/v 做/v

c. 可不可以做

Autotag: 可不可以(D) 做(VC)

ICTCLAS: 可不/l 可以/v 做/v

d. 可以不可以做

Sinica corpus: 可以(VH) 不(D) 可以(VH) 做(VC) (ifrien : 1002)

Autotag: 可以(D) 不可以(D) 做(VC)

ICTCLAS: 可以/v 不/d 可以/v 做/v

e. 負不負責任

Sinica corpus: 負不負責任(VH) (txi172 : 5163)

Autotag: 負(VJ) 不(D) 負責任(VH)

ICTCLAS: 負/v 不/d 負/v 責任/n

To get around inconsistent segmentation among the corpus itself and various taggers, we combine them into one word in the early text processing stage prior to subsequent detection modules. Being treated as a univariate feature also streamlines the whole analysis, though it may not adhere to segmentation standards.

The risk is that, these inconsistently segmented forms cannot be blindly merged into one word since some are not alternative questions, as shown in the following example:

(22) 雨(Na) 該(Nes) 下(VC) 不(D) 下(VC) (a472a : 2160)

Yǔ *gāi* *xià* *bú* *xià*

rain should to rain not to rain

It should rain, but it doesn't.

There is no enough evidence in the corpus for such cases. The author can only think of a few similar cases such as “要下不下” and “說下不下”.

5.2.3 A-not-A Questions and Simplified Forms

After merging, the A-not-A questions can be analyzed systematically. The patterns can be summarized as follows, where $z_1, z_2, \ldots, z_n$ denote Chinese characters, and “*” denotes zero or more occurrences:

- “z_1 不 z_1 (z_2* z_3* ...)” cases: 會不會(D), 可不可以(D), 負不負責任(VH), etc.
- “z_1 z_2 不 z_1 z_2 (z_3* z_4* ...)” cases: 可以不可以, 吃飯不吃飯, 可以不可以做, etc.
- “z_1 z_2 不 z_1” cases: 吃飯不吃, etc.
- “有沒有”.
- Simplified or grammaticalization cases: 是否, 可否, 能否, etc.

As for the simplified or grammaticalization cases, there are 3 words found in the corpus: “是否” (ranked 11th), “可否” (ranked 163rd), and “能否” (ranked 191st). One may wonder if there is still other such “z_1 否” cases. A search in the MOE's *Mandarin Dictionary Revised* shows one more case “然否”, though it is an ancient literary word. The dictionary also shows that “可否” cannot be used blindly since it is also used to compose non-question idioms such as “不置可否” and “未置可否”.

5.2.4 WH Questions

In the QRW list, the WH-family is the largest group and also has productive and tricky morphological patterns. The patterns can be summarized as follows, where z denotes a Chinese character, “|” denotes “or”, and “*” denotes zero or more occurrences:

- “(什|甚) 麼”.
- “爲 (什|甚) 麼”.
- “幹 (嘛|麼)”.
- “幹 (什|甚) 麼”.
- “怎 z_1* z_2*” cases: 怎麼(D), 怎麼辦(VH), 怎能(D), etc.
- “哪 z_1* z_2*” cases: 哪(Nep), 哪裡(Ncd), 哪些(Neqa), etc.
- “誰”.
- “啥”.
- “幾 z*” cases, except for the geometry sense of “幾何”.
- “如何”.
- “爲何”.
- “多 (少|久)”.

Having observed this list, one may want to generalize some cases in order to include more morphological variants. For example, one may want to generalize the fixed “(什|甚)麼’ form into a more flexible “z_1 麼” or even “z_1 z_2 麼”. A search in the MOE's *Mandarin Dictionary Revised* lists 32 words of such form, but “這麼” alone is not used for question. Therefore, whenever we try to generalize some cases into a regular expression pattern, we have to examine in MRDs what the pattern matches.

5.2.5 Lexical Semantics of *hé*

Words prefixed with “何” (*hé*) are yet another set of words with productive and tricky morphological patterns. A quick glance at top 120 QRWs for example, there are two groups of such cases:

- Similar to WH words: 何(Nes), 何在(VH), 何謂(VG).
- Used for emphasis or rhetorical questions: 何必(D), 何不(D).

The second group will be discussed later in Section 5.2.7. Now let's focus on the first group.

In the first group, the lexeme "何" acts as a query focus of the whole sentence, and can usually be translated into an English WH-word (we have seen such examples in Figure 5). Again one may want to generalize this observation into a more flexible "何 z" pattern. A search in the MOE's *Mandarin Dictionary Revised* lists 91 words of such form. Among them, some words are not question but people's full names since "何" is also a common Chinese surname. Therefore, proper noun detection is required if we want to filter out such cases.

In practice, some taggers segment the word "何 z" (where z is a noun lexeme) into two words "何" and "z" except for common cases such as "何時", while some taggers treat it as one word by applying morphological rules. For example,

(23) a. 何處

Sinica corpus: 何處(Nc)

Autotag: 何處(Nc)

ICTCLAS: 何處/r

b. 何方

Sinica corpus: 何方(Ncd)

Autotag: 何方(Ncd)

ICTCLAS: 何/nr 方/nr

c. 何時

Sinica corpus: 何時(Nd) *usually*

Sinica corpus: 何(Nes) 時(Na) *sometimes...*

Autotag: 何時(Nd)

ICTCLAS: 何時/r

d. 何人

Sinica corpus: 何(Nes) 人(Na)

Autotag: 何(Nes) 人(Na)

ICTCLAS: 何人/r

e. 何種

Sinica corpus: 何(Nes) 種(Nf)

Autotag: 何(Nes) 種(Nf)

ICTCLAS: 何種/r

Therefore, the pattern should be specified as “何 | 何z” to accommodate these difference.

5.2.6 Lexical Semantics of Honorifics

We find that lexical semantics have a certain influence on determining or predicting a question clause. Take the ranked 173rd entry “貴姓” (*gùixìng*; your last name) for example. Intuitively speaking, it is typically used in interrogative sentences for asking other person’s last name in a very polite manner. Statistically speaking, although it is a low frequency word (occurrence = 12, recall = 0.05%), its high precision (91.67%) suggests high validity in predicting a question sentence. The only one false positive found in the corpus is a chapter title of a textbook on conversation:

(24)	請問(VE)	您(Nh)	貴姓(VH)	。(PERIODCATEGORY)	(ebach1:80)
	Qǐngwèn	*nín*	*gùixìng*		
	ask	you	last name	period	
	May I ask your last name, please.				

In fact, either a period or a question mark is acceptable here. Different people have different opinions.

On closer inspection, the lexeme “貴” (*gùi*) has multiple senses, and the sense used for composing the word “貴姓” is labeled as an honorific. What makes things more complicated is that not all such honorific words prefixed with “貴” are used for asking questions. Consulting authoritative dictionaries such as *Unabridged Dictionary of Chinese Characters* [59] and *Unabridged Mandarin Dictionary* [36], we may find that “貴庚” (*gùigēng*; your age), “貴幹” (*gùigàn*; your intention), “貴事” (*gùishì*; your intention), and “貴處” (*gùichù*; your native place) are labeled explicitly as interrogatives, while “貴戚,” “貴賓,” “貴子,” “貴手,” “貴恙,” and “貴地” are not.

Moreover, the lexeme “貴” in this sense is defined in the *Unabridged Dictionary of Chinese Characters* as “an honorific; similar to another lexeme ‘尊’ (*zūn*).” For example, “尊姓” (*zūnxìng*; your last name) is similar to “貴姓” in that it is also used to ask other person’s last name in a polite manner. Interestingly enough, not all such honorific words prefixed with “尊” are used for asking questions; e.g., “尊翁” (*zūnwēng*; your father) and “尊容” (*zūnróng*; your face).

Therefore one may wonder if there is still any other honorific lexeme or word used for asking questions. To discover more exhaustive and precise knowledge about such interrogative use of honorific or still other lexemes, we need MRDs with quality linguistic information. To do this, we utilize the auxiliary MRD (MOE’s *Mandarin Dictionary Revised*; see Section 4.3) in the following way. The first step is trying to explore as many honorifics as possible. Since this dictionary does not label information in a consistent way for ease of software processing, we can do nothing but search exhaustively for the such terms as “敬語” (*jìngyǔ*), “敬詞” (*jìngcí*), and “敬稱” (*jìngchēng*) in the definition part of the dictionary. There are totally 189 words found, and most of them are noise, of course. The next step is filtering out words that have no question marks in their example part, and there remains 42 words unfiltered. Next we have to examine them carefully to filter out noise. Finally we obtain 9 question-related honorifics, as shown in Table 7.

Having found so many interrogative honorifics, one may want to test the validity of them. First, let’s try to verify if there is strong evidence in the Sinica corpus. “貴庚” appears only once, “尊姓大名” twice, “大號” 10 times but none falls into this sense. Therefore, the corpus can tell nothing except for the word “貴姓.”

Table 7: Honorifics found to be relevant to questions as a result of mining MOE's *Mandarin Dictionary Revised*. The word with a symbol † means that it appears but is not labeled explicitly as interrogatives in the definition part of the *Unabridged Dictionary of Chinese Characters*, while the word with a symbol ‡ means that it appears but is not labeled explicitly as interrogatives in both the definition and example parts of the *Unabridged Mandarin Dictionary*

Common case	Rare case	Multiple-sense case
貴姓	貴處	大號
貴庚	貴甲子	
貴幹	貴恙†	
尊姓大名	貴降‡	

Next, let's try to verify if there is agreement among different dictionaries. As Table 7 shows, most are in agreement except for 2 words. There may be two reasons for this. The fact that some words are only used in ancient literary context reduces agreement among today's lexicographers, and these dictionaries were not compiled from a more modern linguistic perspective.

5.2.7 Evaluative Adverbs and Rhetorical Questions

Another interesting examples showing the importance of quality MRDs are evaluative adverbs and still other words for rhetorical questions. The QRW list found so far has successfully identified a few of such kinds of words with high rankings. For instance, the former are "到底" (*dàodǐ*, ranked 10$^{\text{th}}$), "難道" (*nándào*, ranked 15$^{\text{th}}$), and "究竟" (*jìujìng* in Taiwan and *jīujìng* in mainland China, ranked 20$^{\text{th}}$); the latter are "何必" (*hébì*, ranked 42$^{\text{nd}}$), "何不" (*hébù*, ranked 112$^{\text{th}}$), "何苦" (*hékǔ*, ranked 132$^{\text{nd}}$), "何嘗" (*hécháng*, ranked 133$^{\text{rd}}$), and "何況" (*hékuàng*, ranked 687$^{\text{th}}$).

What have linguists said about these two groups of words? About the former group (evaluative adverbs):

- In [38], "到底," "難道," "難道說" (*nándàoshuō*), and "究竟" all have *emphasis* function, but only "難道" and "難道說" are used exclusively for questions; the others can also be used in other situation.

- "到底" and "究竟" are classified as adverbs with a *subjective assessment* *attribute* in the Sinica BOW database.

- “難道” is classified as a mood or modal adverb in [19].
- “究竟” is classified as an evaluative adverb in [12, p. 58].

As for the latter group (other words for rhetorical questions):

- “何必,” “何不,” “何苦,” and “何嘗” are classified as mood or modal adverbs for rhetorical questions in [38].
- “何況” is classified as a mood or modal conjunction for rhetorical questions in [38].
- “何必” is classified as an modal adverb with *question* and *necessity* properties in [9, p. 90].
- Later, [3] enumerates all modal words occurred in CKIP's *Chinese Electronic Dictionary*. “何必” and “何須” (*héxū*) are classified as modal words with an *interrogative deontical necessity* property.
- “應否” (*yīngfǒu*) is classified as a modal word with an *interrogative deontical probability* property in [3].
- “可否” (*kěfǒu*), “能否” (*néngfǒu*), “豈能” (*qǐ'néng*), and “怎能” (*zěnnéng*) are classified as modal words with an *interrogative deontical possibility* property in [3].

It can be easily seen that different linguists have minor disagreements about these words. In this study we will not engage in such debate. What we care about is trying to explore as many similar cases as possible.

To do this, again we utilize the auxiliary MRD (MOE's *Mandarin Dictionary Revised*; see Section 4.3) in a similar way. The fact that this dictionary does not classify adverbs into finer subcategories such as evaluation, manner, degree, etc.

makes it nearly impossible to explore more cases for evaluative adverbs. Instead, we will focus on exploring any other similar words used for rhetorical questions.

The first step is trying to explore as many keywords for rhetorical questions as possible. Since this dictionary does not label information in a consistent way for ease of software processing, we can do nothing but search exhaustively for the such terms as “反問” (*fǎnwèn*) and “反詰” (*fǎnjié*) in the definition part of the dictionary. There are totally 90 words found, and many of them are noise, of course. The next step is filtering out words that have no question marks in their example part, and there remains 49 words unfiltered. Next we have to examine them carefully to filter out noise. Finally we obtain 46 keywords for rhetorical questions, as shown in Table 8.

5.2.8 Person

One may wonder why the word “你” (*nǐ*; you; second person singular pronoun) is ranked 5$^{\text{th}}$ since intuitively it is irrelevant to questions. Let us look at this issue from another angle: recall. Recall is the ratio of the number of relevant items correctly identified to the total number of relevant items in the population, i.e., recall $= a/(a + b)$ in Table 4. From this point of view, Table 9 shows an interesting finding. This table lists the top 10 QRWs in terms of recalls, along with their precisions for comparison. We may see that the word “你” has such a high recall (12.93%) that for every 7.7 question clauses, there is about one clause containing the word “你”. In other words, people tend to express questions with the word “你”.

One may argue that the recall of “你” is so high simply because it is a high-frequency word and further argue that it is irrelevant to questions. The argument is partially true in that “你” is the 19$^{\text{th}}$ most frequent word in the corpus (see Appendix A). Moreover, the correlation coefficient r of recall and frequency is +0.73, and the r^2 is 0.53, indicating a modest positive association. However, it should be noticed that not only would recall and frequency but still other factors as well would affect the degree of relevance to question. Some high-recall words

Table 8: Keywords for rhetorical questions as a result of mining MOE's *Mandarin Dictionary Revised*

	Single sense case	Multiple sense case
Normal case	何干	了得
	何不	不成
	何功之有	不是
	何必	何以
	何妨	怕
	何苦	便
	何消說	哪
	何敢	豈
	何須	豈不是
	何罪之有	啊
	何樂不爲	
	與否	
	難不成	
	難道	
	嗄	
Ancient literary case	不亦	乎
	向非	何用
	何消	何有
	何得	何但
	何聊賴	何爲
	何極	何當
	庸詎	其
		哉
		爲
		烏
		盍
		與
		歟
		詎

Table 9: Top 10 question-related words in terms of recalls

	Recall		Precision			Recall		Precision	
QRW	Ranking	%	Ranking	%	QRW	Ranking	%	Ranking	%
的(DE)	1	24.71	862	1.97	你(Nh)	6	12.93	95	15.79
是(SHI)	2	19.48	329	4.85	嗎(T)	7	12.39	1	98.12
呢(T)	3	16.80	25	61.04	有(V_2)	8	9.64	375	4.38
不(D)	4	15.78	172	8.42	我(Nh)	9	6.92	497	3.66
什麼(Nep)	5	13.39	59	41.26	這(Nep)	10	6.32	455	3.86

Table 10: Relation between person and degree of relevance to questions in the Sinica corpus

		LLR	Recall		Precision	
	Pronoun	Ranking	Ranking	%	Ranking	%
2nd person	你	5	6	12.93	93	15.79
	你們	21	40	1.79	80	20.77
	妳	32	57	1.30	81	20.63
	您	37	53	1.38	92	16.39
1st person	我	111	9	6.92	497	3.66
	我們	155	23	3.40	458	3.85
3rd person	他	563	21	3.46	736	2.42
	他們	456	50	1.45	532	3.38
	她	780	61	1.24	743	2.40
	她們	1090	663	0.09	570	3.11

(such as “的” and “是”) are merely high-frequency words since they have much lower precision measures, implying a smell of stop words. We may, therefore, reasonably conclude that the word “你” is relevant to questions, due to not only high frequency but also other factors.

It can also be seen that the second person pronouns (singular “你,” plural “你們,” feminine “妳,” and honorific form “您”) have higher χ^2 statistics than the first person pronouns (singular ““我” and plural “我們”), and much higher than the third person pronouns (singular “他,” plural “他們,” and feminine “她”), as shown in Table 10. The one-way ANOVA (analysis of variance) test on the precision column gives us the p-value $= 1.08 \times 10^{-5} \ll 0.01$, indicating that there is a statistically significant difference between the precision of the second person pronouns and the first/third person pronouns. This finding further implies that the existence of second person pronouns has a higher predictive validity for question sentences.

5.2.9 Roles

It seems that thematic roles and discourse functions also determine whether the person has remarkable influence. For instance, [7] suggested that if the subject of a main clause are one of the second person pronouns, the sentence tends to be a real interrogative; if first person pronouns, non-interrogative. To see if there is really such a tendency towards the second person pronouns, we try to investigate in the Sinica treebank which roles are played more frequently by which pronouns. Detailed description of the role scheme adopted in the treebank can be found in [31].

Since the Sinica treebank removes all punctuation (as mentioned in Section 4.2), our investigation is performed in three stages. The first stage is to find out all trees containing the 4 second person pronouns by the keyword-based search on the Sinica treebank Web site, and there are totally 1,227 trees found. The trees look like the following:

(25) S(evaluation:Dbb: 究竟|agent:NP(Head:Nhaa: 你)|
epistemics:Dbaa: 是|reason:Dj: 怎麼|
Head:VD1: 分配|particle:Ta: 的)

The second stage is then trying to extract semantic roles associated with each pronoun. To simply the task of tree parsing, by issuing the pattern "`Head:Nh%`" with the "process again" and then the "filtering" command, the semantic role will be highlighted in red color as the following HTML code:

(26) S(evaluation:Dbb: 究竟|
agent:NP(Head:Nhaa: 你)|
epistemics:Dbaa: 是|reason:Dj: 怎麼|
Head:VD1: 分配|particle:Ta: 的)

Even if some complicated trees are not explicitly highlighted in the same way, for example,

(27) VP(Head:VL4: 讓|
goal:NP(predication:VP ・的 (head:VP(quantity:Daa: 只|
Head:V_2: 有|range:NP(property:A: 手提|Head:Nab: 行李))|
Head:DE: 的)|Head:Nhaa: 你))

Table 11: Distribution of 2nd person pronouns and their respective semantic roles in the Sinica treebank. The "Q" columns list the numbers of question clauses for corresponding pronoun/role pairs, while the "¬Q" non-question clauses

	你			你們			妳			您		
Role	Q	¬Q	$\frac{Q}{Q+\neg Q}$	Q	¬Q	$\frac{Q}{Q+\neg Q}$	Q	¬Q	$\frac{Q}{Q+\neg Q}$	Q	¬Q	$\frac{Q}{Q+\neg Q}$
Agent	52	196	0.21	8	16	0.33	5	27	0.16	9	23	0.28
Experiencer	32	55	0.37	2	11	0.15	5	9	0.36	0	10	0.00
Goal	12	105	0.10	3	8	0.27	0	8	0.00	1	26	0.04
Others	18	121	0.13	2	16	0.11	3	17	0.15	2	16	0.11
Theme	17	119	0.13	3	18	0.14	2	11	0.15	7	20	0.26
Total	131	596	0.18	18	69	0.21	15	72	0.17	19	95	0.17

the beginning of the result page still tells us that the role is "goal."

The last stage is then trying to trace these trees back to their origins in the Sinica corpus, as has been presented in Figure 4. Totally 1,178 trees are successfully backtracked (success rate = 96%) and have their punctuation assigned accordingly. The remaining 4% of trees are dropped away here for the sake of objectivity. Finally the distribution of the 4 second person pronouns and their respective semantic roles is listed in Table 11.

Since the sample size of the pronoun "你" is quite large, it is safe to perform a χ^2 test on the Q and ¬Q columns of it. The p-value is $1.752 \times 10^{-6} \ll 0.01$, indicating that in the "你" case there is a statistically significant relationship between the semantic roles and questions. On closer inspection, the largest component of χ^2 is for the "Experiencer-Q" cell (16.996; see Figure 10), i.e., this combination contributes to the most to the overall distance χ^2. Even if we regard the "experiencer" row as exceptional (outlier or contaminated), redoing the χ^2 test on the same table except the "experiencer" row will produce the p-value $= 0.0219 < 0.05$, still indicating a significant evidence. Therefore, it is safe to conclude that different roles of "你" does have some remarkable influence on predicting questions.

How about the other 3 pronouns? The raw counts are too small to do the same χ^2 test, but we can instead calculate the $\frac{Q}{Q+\neg Q}$ ratio (see Table 11 for statistics and Figure 12 for the boxplot) and then perform the ANOVA test on them. The one-way ANOVA test on the four $\frac{Q}{Q+\neg Q}$ columns generates the p-value $= 0.31353$,

```
Chi-Square Test: Q, not Q

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

             Q   not Q  Total
    1       52     196    248
         44.69  203.31
         1.196   0.263

    2       32      55     87
         15.68   71.32
        16.996   3.736

    3       12     105    117
         21.08   95.92
         3.913   0.860

    4       18     121    139
         25.05  113.95
         1.983   0.436

    5       17     119    136
         24.51  111.49
         2.299   0.505

Total      131     596    727

Chi-Sq = 32.187, DF = 4, P-Value = 0.000
```

Figure 10: Minitab χ^2 output for comparing the 5 roles of 2^{nd} person pronoun "你" for Table 11. The five rows are for Agent, Experiencer, Goal, Others, and Theme, respectively. The p-value is given as 0.000 here because the Minitab software rounds it to 3 decimal places, implying that $p < 0.0005$

```
One-way ANOVA: Ratio versus Role

Source  DF      SS      MS     F      P
Role     4  0.0582  0.0145  1.30  0.314
Error   15  0.1675  0.0112
Total   19  0.2256

S = 0.1057   R-Sq = 25.78%   R-Sq(adj) = 5.99%


                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level        N    Mean   StDev  -+---------+---------+---------+--------
Agent        4  0.2451  0.0780               (-----------*----------)
Experiencer  4  0.2197  0.1765             (----------*----------)
Goal         4  0.1031  0.1208  (----------*-----------)
Others       4  0.1254  0.0185    (-----------*----------)
Theme        4  0.1702  0.0605         (----------*----------)
                                -+---------+---------+---------+--------
                                0.00      0.10      0.20      0.30

Pooled StDev = 0.1057
```

Figure 11: Minitab ANOVA output for comparing the 5 roles of all 2nd person pronouns for Table 11

implying no significant effect as a whole (see Figure 11). On closer inspection of the result, however, the variance or standard deviation in the "experiencer" group is larger than all the other roles. The phenomenon has also been illustrated by Figure 12 since the range (or the "spread") of experiencer is larger than all the others. The large variance will increase the overall within group sum of squares (WSS) or mean square for error (MSE), which in turn decrease the ANOVA $F_{4,15}$ statistic since $F_{4,15} = \frac{\text{mean square for group}}{\text{mean square for error}} = \frac{\text{between group SS/(df. = 4)}}{\text{within group SS/(df. = 15)}}$. Again if we regard the "experiencer" group as exceptional (outlier or contaminated), redoing the ANOVA test on the same table except the "experiencer" group will produce the p-value $= 0.105$, indicating a certain kind of significant evidence, though the effect is not as remarkable as in the sole "你" case.

In conclusion, different roles of the second person pronouns (especially "你") have some influence on predicting questions.

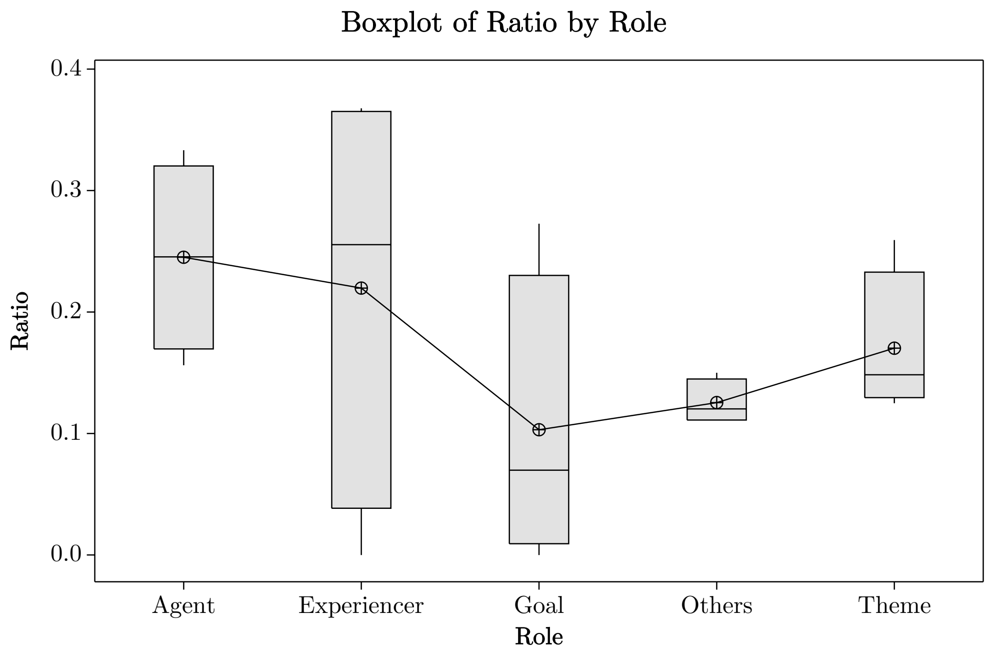


Figure 12: Boxplot of Q/(Q + ¬Q) ratio for different roles played by 4 second person pronouns. The rectangle ("box") part shows the inter-quartile range (i.e., the first quartile, the median, and the third quartile), and the whiskers draw out to the maximum and minimum values since no data is beyond 1.5 inter-quartile range to be considered outlier here. The circle part shows the mean of all data

5.3 Putting Them Together

There have been many types of QRWs found so far. Now let's put them together to see the overall recall and precision. Recall = 81.45%, and Precision = 36.79%. The next stage will focus on increasing the precision by analyzing false positives.

CHAPTER VI

BIVARIATE AND MULTIVARIATE ANALYSIS

Since the goal of previous univariate module is to maximize recall, there are many false positives. Therefore, both true and false positives are sent to and re-analyzed by bivariate and then multivariate modules, during which more and more false positives will be filtered out so as to increase precision.

6.1 Bivariate Analysis by Exception Rules

At bivariate level, exceptional cases that may be identified are mostly compound relatives and higher verbs. The issue of compound relatives can be summarized as follows:

- "$w_{\text{negation}} \ldots$ WH" cases: 不, 沒, 沒有, 別, 不管, 無論, 不論, etc.
- "WH . . . (都 |也 |就)" cases.

Since the patterns are quite consistent, they can be easily coded in the format of ⟨*ExclusiveAtom*⟩ (see Figure 2). By applying these exception rules, precision increases from 36.79% to 39.84%, and recall decreases from 81.45% to 80.12%.

In Section 2.4 we have outlined the issue of higher verbs and raised a few questions about it. Here we examine the study by Cheng [7, 8] and choose the following types of higher verbs for experiment:

- Ask-type verbs: 問, 追問, 質問, 追究, 調查, 請示.
- Test-type verbs: 探討, 討論, 研究, 實驗, 試驗, 試試, 試試看, 嘗試, 考慮, 關心.

By appending these higher verbs to the rules constructed at univariate stage, precision increases slightly from 36.79% to 36.95%, and recall decreases from

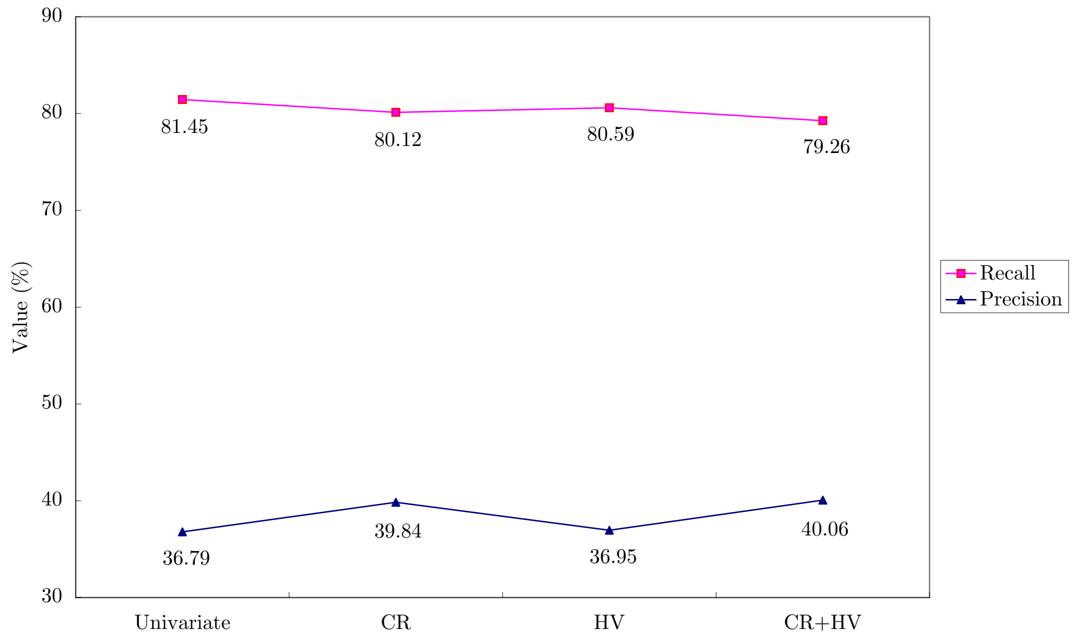


Figure 13: The results applying rules of compound relatives and higher verbs. The "CR" denotes compound relatives. The "HV" denotes higher verbs. The "CR+HV" denotes the combination of both rules

81.45% to 80.59%. Combining both higher verbs and compound relatives rules, precision is 40.06%, and recall is 79.26%. Therefore the effectiveness of higher verbs is still unclear. The results are illustrated in Figure 13.

6.2 Multivariate Analysis by Language Models

In order to reduce possible sampling errors, when there is a need to divide the dataset into a training and a test set, we select a simple random sample (SRS) of a given ratio $1/r$ as the test set; the remaining is used for training.

Figure 14 illustrates the overall flow of dataset preparation. At this stage, we collect all clauses that pass the univariate and bivariate rules. They are, of course, composed of both true and false positives. Then we divide them into a question set $\mathcal{S}'_{\mathrm{Q}}$ and a non-question set $\mathcal{S}'_{\mathrm{NQ}}$. Each one is further divided into a training set and a test set by SRS. Now our training process will focus on the two training sets: the question training set $\mathcal{S}'_{\mathrm{Q,tr}}$ and the non-question training set $\mathcal{S}'_{\mathrm{NQ,tr}}$.

Next, at the training stage, we train a pair of competitive language models for both $\mathcal{S}'_{\mathrm{Q,tr}}$ and $\mathcal{S}'_{\mathrm{NQ,tr}}$. Let's call them LM_{Q} and LM_{NQ} respectively.

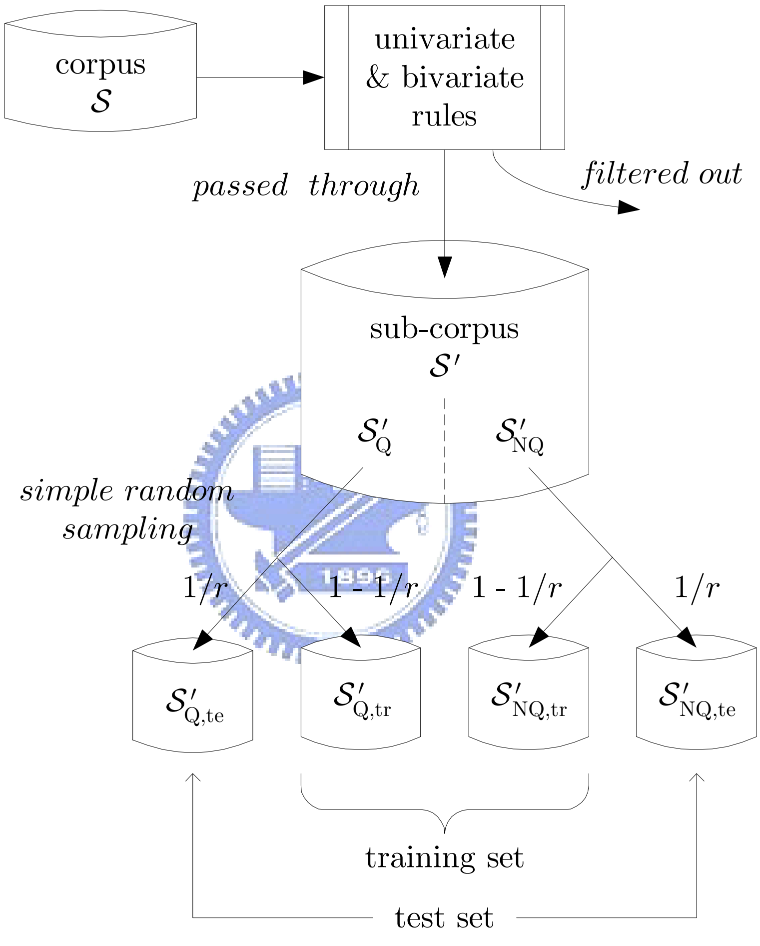


Figure 14: Prepare training and test sets by simple random sampling

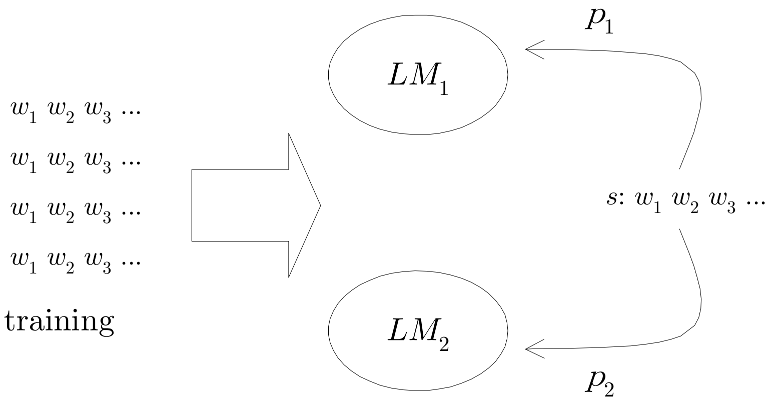
Finally, let's take a look at the detection stage.

Traditionally, perplexity (or more precisely, cross-perplexity) is used as a measure of how close a language model is to its theoretically perfect model. Let two candidate language models LM_1 and LM_2 be constructed with the same training set and then evaluated with the same test set. We say that LM_1 is, with regard to the perfect model, better at modeling the dataset than LM_2 is if perplexity values $p_1 < p_2$, and vice versa. The concept is illustrated in Figure 15a.

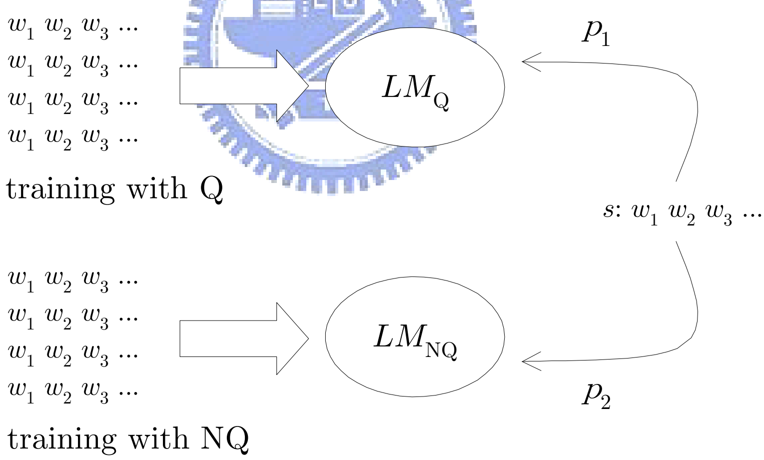
Now let's reverse the evaluation direction. Given a sentence s, it is evaluated by both LM_{Q} and LM_{NQ}, and two perplexity values p_1 and p_2 will be generated, respectively. Assume that both LM_{Q} and LM_{NQ} are good approximation to their perfect models. Since perplexity can be considered a measure of how close a language model is to s, it follows that if $p_1 < p_2$, the LM_{Q} is a better match for s than LM_{NQ}, and vice versa. Therefore we use the preplexity as a criterion to classify the s into a question (modeled by LM_{Q}) or a non-question (modeled by LM_{NQ}). The concept is illustrated in Figure 15b.

This approach works under the assumption that both LM_{Q} and LM_{NQ} are good approximation to their perfect models. It follows that the performance of this apporoach would rely on how good the language models are and how likely they will discriminate between question and non-question cases. Here we consider two types of language modeling techniques. The first is a trigram model with Good-Turing discounting and Katz backoff for smoothing (see [41, Chapter 6] and [28, Chapter 6] for more details). The second is an interpolated smoothing model since it has been reported in [6, 25] that interpolated Kneser-Ney smoothing (including higher-order n-gram models, especially 5-gram) performs better than many others in every situation they have examined. Whenever possible, we experiment with three configurations: trigram, 4-gram, and 5-gram.

There are still some variation of details that need consideration when constructing the language models. Here we consider two possible variations: tag vs. word and tag unification.



(a) Traditional use of language models.



(b) Our approach to using language models.

Figure 15: Using language models to discriminate questions

Table 12: Different configurations used in our language modeling experiments

Dataset	Good-Turing/Kats trigram	Interpolated Kneser-Ney trigram	4-gram	5-gram
word	GT-w	IKN3-w	IKN4-w	IKN5-w
tag	GT-t	IKN3-t	IKN4-t	IKN5-t
tag unification	GT-tx	IKN3-tx	IKN4-tx	IKN5-tx

Data sparseness causes problems in nearly every language model technique. Since a training set is more sparse when it is composed in terms of a series of *words* than when it is composed in terms of a series of *POS tags*, we suspect if the language model constructed in terms of POS of words is better than the one in terms of words themselves. Therefore, both approaches will be used for comparison.

In addition, at times the Sinica corpus assigns different POS tags to the same type of univariate features. Take "A-not-A" words for example, "會不會" is assigned a `D` tag while "好不好" `VH`. As a consequence, we suspect if it is inappropriate to train the language models in terms of the original tagset assignment of the corpus. To verify this, we will conduct a pair of experiments to see if there is any performance difference by unifying a variety of such tags into a single one (let's name it "XXX" tag for convenience).

Putting them together, we will experiment with several kinds of configuration, as summarized in Table 12.

Finally, the performance of language models depends on the selection of training and test sets. Therefore, the whole SRS-division/training/evaluation process is repeated n times (e.g., $n = 20$) to gain a better feeling of stability of this approach with regard to different training/test configuration.

6.2.1 Particles and Interjections

As stated in Section 5.2.1, some sentence-final particles and interjections perform not only question but also euphemism, irony, exclamation, or any other illocutionary act. Since linguists disagree with qualitative analysis and explanation of the

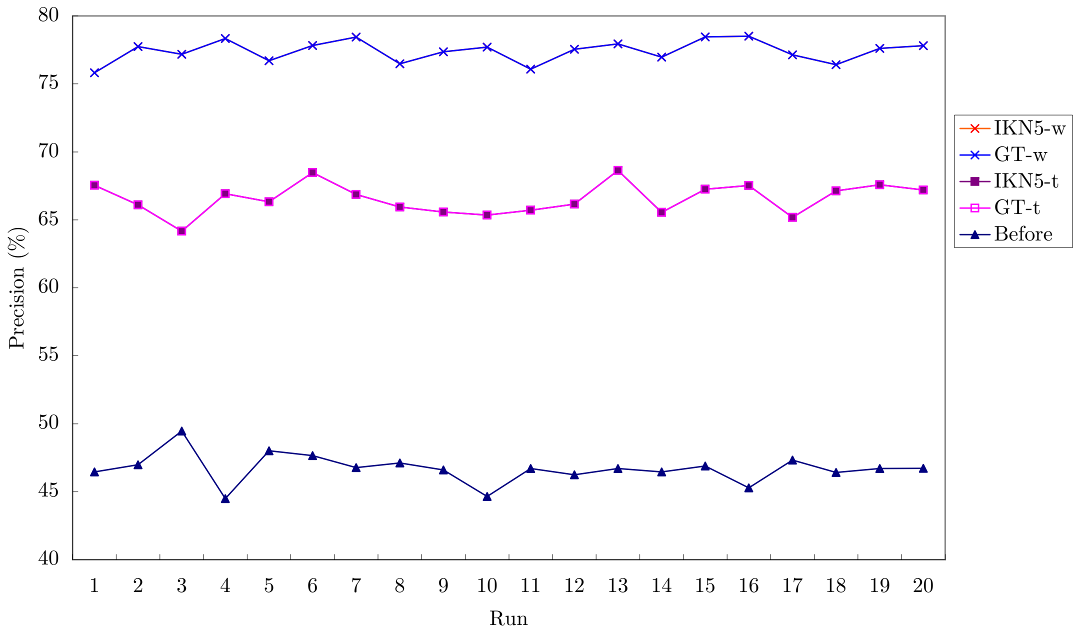


Figure 16: The result of using language models to discriminate the case of sentence-final particles. Since all IKNn-w runs produce the same outcome, only IKN5-w is shown in this figure; the same for IKN5-t.
In the "Before" cases, average precision = 46.69%, and standard deviation = 1.09. In the GT-t and IKN5-t cases, average precision = 66.56%, and standard deviation = 1.14. In the GT-w and IKN5-w cases, average precision = 77.40%, and standard deviation = 0.80.

precise way to distinguish between them, we will try another quantitative route to this.

The outcome of 20 experiments is shown in Figure 16. On average, precision increases from 46.69% to 66.56% when undertaking any language modeling technique at tag level, and to 77.40% at word level. All language modeling techniques we use at the same level have the same performance in the total 20 runs, though interpolated Kneser-Ney smoothing has lower average perplexity.

6.2.2 A-not-A Questions and Simplified Forms

The Sinica corpus assigns a variety of POS tags to different A-not-A words, e.g., 會不會(D) and 好不好(VH). Our treatment of A-not-A forms differs with that in the corpus (see Section 5.2.2). Our definition of A-not-A forms is also broader than that in the corpus (see Section 5.2.3). As a consequence, it may be inappropriate

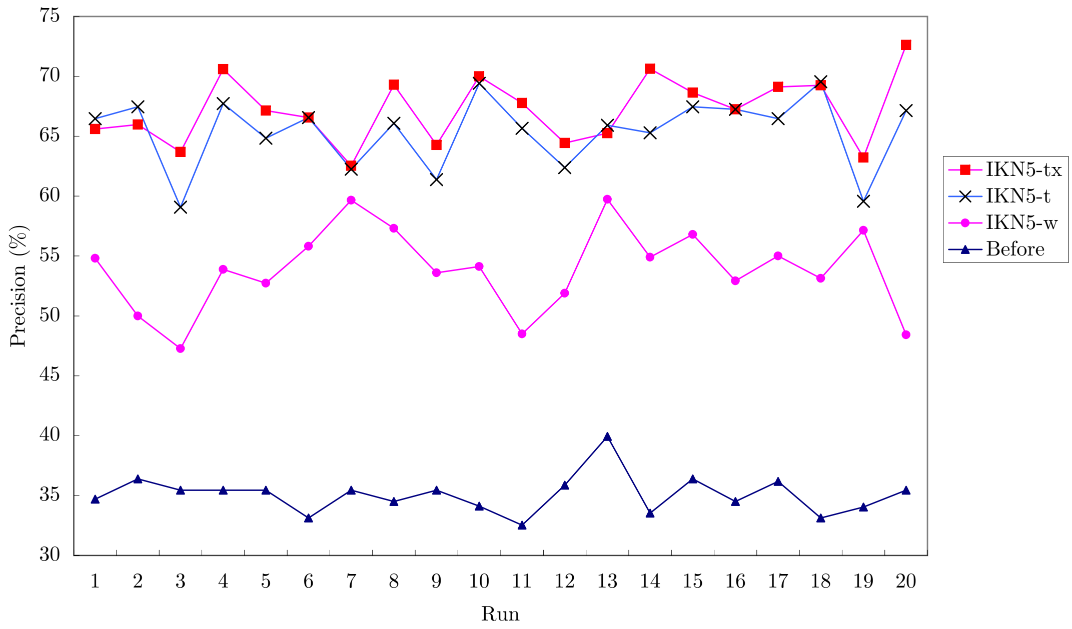


Figure 17: The result of using language models to discriminate the case of A-not-A questions. Since all IKNn-w runs produce the same outcome, only IKN5-w is shown in this figure; the same for IKN5-t and IKN5-tx.
In the "Before" cases, average precision = 35.08%, and standard deviation = 1.60.
In the IKN5-w cases, average precision = 53.88%, and standard deviation = 3.48.
In the IKN5-t cases, average precision = 65.40%, and standard deviation = 2.97.
In the IKN5-tx cases, average precision = 67.19%, and standard deviation = 2.81.
A pairwise Student's t-test on IKN5-t and IKN5-tx produces $p = 0.00051 < 0.001$, implying that there is a statistical significant improvement.

to train the language models in terms of the original tagset of the corpus. To verify this suspect, we will conduct a pair of experiments to see if there is any performance improvement by unifying a variety of A-not-A tags into a single one.

The outcome of 20 experiments is shown in Figure 17. Since all language modeling techniques used here at the same word or tag level produce the same outcome, we show only interpolated Kneser-Ney smoothing of order 5 (IKN5) for brevity. On average, precision increases from 35.08% to 53.88% when applying IKN5-w, to 65.40% when applying IKN5-t, and up to 67.19% when applying IKN5-tx. A pairwise Student's t-test on the two language models IKN5-t and IKN5-tx produces $p = 0.00051 < 0.001$, implying that there is a statistical significant improvement by unifying a variety of A-not-A tags to an fixed artificial one.

6.2.3 WH Questions

As we have seen in Table 4 and Section 5.2.4, words of this type receive a variety of POS tags in the Sinica corpus, e.g., 什麼(Nep), 爲什麼(D), and 怎麼辦(VH). As a consequence, it may be inappropriate to train the language models in terms of the tagset of the corpus. To verify this suspect, we conduct a pair of experiments to see if there is any performance improvement by unifying a variety of WH tags into a single one.

The outcome of 20 experiments is shown in Figure 18. Since all language modeling techniques used here at the same word or tag level produce the same outcome, we show only interpolated Kneser-Ney smoothing of order 5 (IKN5) for brevity. On average, precision increases from 41.23% to 69.52% when applying IKN5-w, to 72.93% when applying IKN5-t, and up to 73.97% when applying IKN5-tx. A pairwise Student's *t*-test on the two language models IKN5-t and IKN5-tx produces $p = 1.51 \times 10^{-5} < 0.001$, implying that there is a statistical significant improvement by changing the POS tags, though the improvement is only very small.

6.2.4 Evaluative Adverbs and Rhetorical Questions

As we have seen in Section 5.2.7, words of this type are mostly adverbs, if not all. However, not all adverbs that appear in a sentence belong to this type. Again we wonder if it is better to train the language models with their POS tags unified into a single one to be distinct from other type of adverbs. To verify this, we will conduct a pair of experiments to see if there is any performance improvement.

The outcome of 20 experiments is shown in Figure 19. Since all language modeling techniques used here at the same word or tag level produce the same outcome, we show only interpolated Kneser-Ney smoothing of order 5 (IKN5) for brevity. On average, precision increases from 45.59% to 61.61% when applying IKN5-w, to 64.46% when applying IKN5-t, and up to 64.64% when applying IKN5-tx. A pairwise Student's *t*-test on the two language models IKN5-t and IKN5-tx

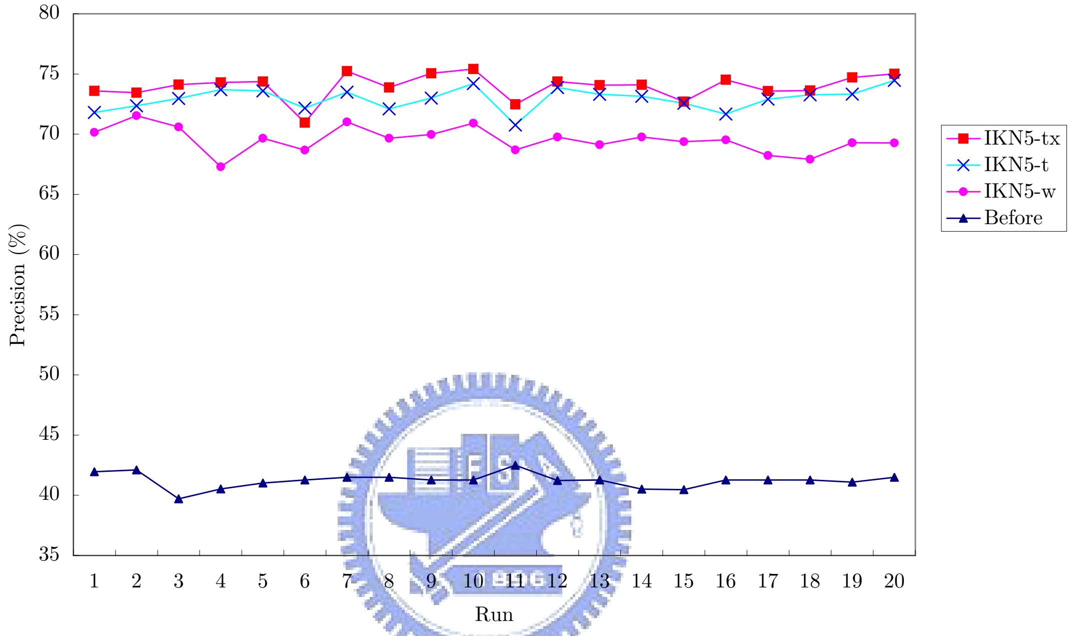


Figure 18: The result of using language models to discriminate the case of WH questions. Since all IKNn-w runs produce the same outcome, only IKN5-w is shown in this figure; the same for IKN5-t and IKN5-tx.
In the "Before" cases, average precision = 41.23%, and standard deviation = 0.61. In the IKN5-w cases, average precision = 69.52%, and standard deviation = 1.05. In the IKN5-t cases, average precision = 72.93%, and standard deviation = 0.92. In the IKN5-tx cases, average precision = 73.97%, and standard deviation = 1.05. A pairwise Student's t-test on IKN5-t and IKN-tx produces $p = 1.51 \times 10^{-5} < 0.001$, implying that there is a statistical significant improvement.

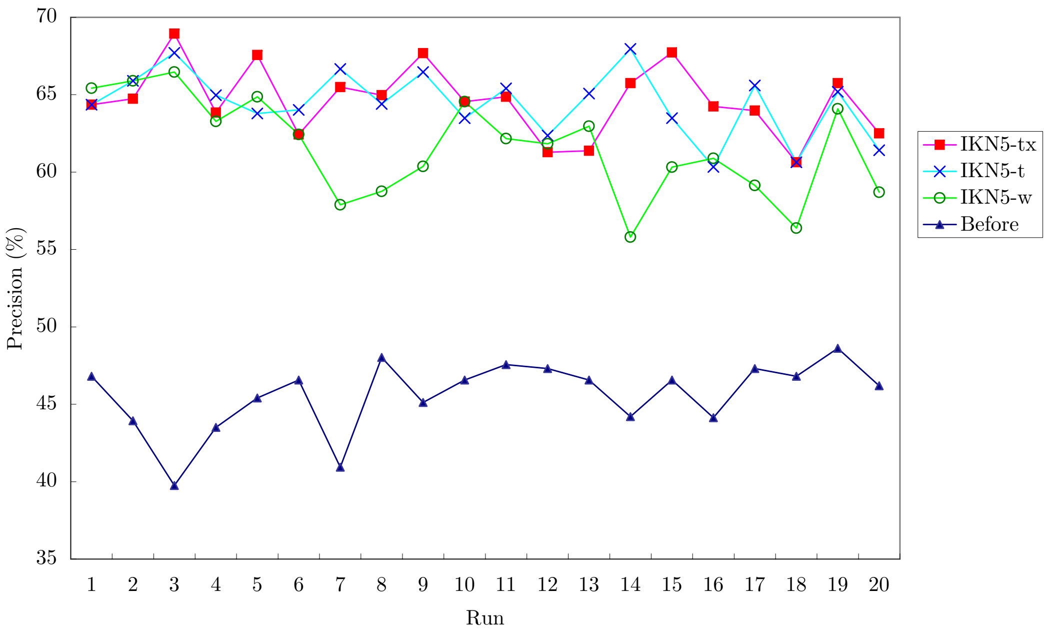


Figure 19: The result of using language models to discriminate the case of evaluative adverbs and rhetorical questions. Since all IKNn-w runs produce the same outcome, only IKN5-w is shown in this figure; the same for IKN5-t and IKN5-tx.
In the "Before" cases, average precision = 45.59%, and standard deviation = 2.29.
In the IKN5-w cases, average precision = 61.61%, and standard deviation = 3.15.
In the IKN5-t cases, average precision = 64.46%, and standard deviation = 2.11.
In the IKN5-tx cases, average precision = 64.64%, and standard deviation = 2.27.
A pairwise Student's t-test on IKN5-t and IKN5-tx produces $p = 0.357$, implying that there is no statistical significant improvement.

produces $p = 0.357$, implying that there is no statistical significant improvement by unifying the POS tags. Therefore, it may be unnecessary to unify the POS tags for such cases. In addition, standard deviations are so large in all cases that there is still room for a deeper study.

CHAPTER VII

CONCLUDING REMARKS

7.1 Discussion

For now our system has average recall 76.26% and precision 73.43%. Let us devote a little more space to examining the result and possible ways to improve the performance.

7.1.1 False Negatives

Our experience shows that the most fatal obstacle to improve recall is pragmatic issues. Some false negatives are truly our faults, but others (especially those with sentence-end particles) can also be considered euphemism, irony, exclamation, or any other illocutionary act. For example,

(28) a. 台大醫院新大樓沒注意過?
 b. 你不跟我學武藝了?
 c. 你是活得不耐煩了?

To reduce the number of false negatives, the most challenge is that programs should be able to recognize some kinds of speech acts.

7.1.2 False Positives

Our experience shows that the most fatal obstacle to improve precision is referentiality. As stated in Section 2.3, there seems no obvious syntactic pattern to identify indefinitives. Another difficulty is again in pragmatic issues. To reduce the number of false positives, we may need to experiment with some more powerful multivariate models to identify indefinitives more accurately, and a little tree parsing may also help.

The issue of higher verbs also needs study. For example, the verb “決定” (*juédìng*; decide) is not listed in the literature mentioned in Section 2.4, but the following sentence in the Sinica corpus is a non-question:

(29) 才 決定 如何 走下去 。

The inclusion of higher verbs, as suggested in Section 6.1, improves the precision only by 0.16%. Therefore, the validity of higher verb list is doubtful.

Finally, some words have multiple senses, and not all senses function as question. For example, “究竟,” “幾何,” and “到底” cannot be treated blindly as question without some process of word sense disambiguation.

7.1.3 Clause or Sentence Boundary

The syntax of Mandarin is very flexible compared to Indo-European languages. Such flexibility, however, increases the search space for parsing. As stated in Section 1.3, there is no syntactically decisive and reliable marker and word order in Mandarin question sentences. In real setting, therefore, given a series of clauses, a question-detection program must be able to identify the beginning and the end of every sentence. Otherwise it may be confused about complex sentences or serial clauses, and has trouble dealing with alternative questions spanning over several clauses.

7.2 Summary

In this study we have pointed out the problem of detecting Mandarin question sentences and reviewed relevant linguistic literature. Then we have outlined our strategy to approach this problem: to increase recall first and then to increase precision. We have presented our statistical approaches and procedure, and discussed our findings. The lack of appropriate machine-readable dictionaries and electronic resources limits our pursuit of several subtle issues.

This is a new topic in NLP community as far as we know. Our contributions are twofold. In the linguistic field, we re-examine relevant topics from a new quantitative point of view, and discover more comprehensive and precise features. In the NLP field, we demonstrate several techniques that is useful for this problem, and achieve good recall and precision in the preliminary study.

APPENDIX A

LIST OF QUESTION-RELATED WORDS

This appendix details the top 300 results generated by the procedure discussed in Section 5.1. In the beginning, the four counters a, b, c, and d are accumulated in the following way: for every $w_i \in$ {all words in the corpus},

	Is w_i in the clause?	
Clauses	Yes	No
Ends with '?'	a	b
Ends without '?'	c	d

Next, compute a series of intermediate values:

$$\begin{aligned} n &= a+b+c+d \\ m_a &= (a+b)(a+c) \\ m_b &= (a+b)(b+d) \\ m_c &= (a+c)(c+d) \\ m_d &= (b+d)(c+d) \end{aligned}$$

Finally, calculate LLR statistic, χ^2 statistic, precision, and recall of w_i in the following way:

$$\begin{aligned} \text{LLR statistic} &= 2 \times \sum_{j=a}^{d} j \ln \frac{n \times j}{m_j} \\ \chi^2 \text{ statistic} &= \frac{n(ad-bc)^2}{m_a m_d} \\ \text{Frequency} &= a+c \\ \text{Precision} &= a/(a+c) \\ \text{Recall} &= a/(a+b) \end{aligned}$$

Table 13: List of top 300 question-related words (QRWs)

Ranking		QRW	Statistic		Count				Recall	Precision
LLR	χ^2	w_i	LLR	χ^2	a	b	c	d	(%)	(%)
1	1	嗎(T)	17,956.52	88,941.79	2,507	17,721	48	729,611	12.39	98.12
2	2	呢(T)	17,798.44	72,731.19	3,398	16,830	2,169	727,490	16.80	61.04
3	3	什麼(Nep)	11,223.98	37,515.11	2,708	17,520	3,855	725,804	13.39	41.26
4	4	爲什麼(D)	6,163.42	26,622.68	1,149	19,079	593	729,066	5.68	65.96
5	6	你(Nh)	5,464.47	11,060.08	2,615	17,613	13,946	715,713	12.93	15.79
6	5	怎麼(D)	4,776.90	18,747.55	998	19,230	835	728,824	4.93	54.45
7	13	不(D)	3,320.25	4,982.97	3,191	17,037	34,695	694,964	15.78	8.42
8	7	誰(Nh)	2,685.79	9,161.25	661	19,567	931	728,728	3.27	41.52
9	10	如何(D)	2,548.63	7,369.35	761	19,467	1,735	727,924	3.76	30.49
10	8	到底(D)	1,998.58	8,221.60	400	19,828	276	729,383	1.98	59.17
11	14	是否(D)	1,653.17	4,540.90	530	19,698	1,393	728,266	2.62	27.56
12	9	怎麼辦(VH)	1,549.34	7,372.23	257	19,971	62	729,597	1.27	80.56
13	12	怎麼樣(VH)	1,454.17	5,464.69	323	19,905	328	729,331	1.60	49.62
14	33	是(SHI)	1,342.47	1,605.50	3,940	16,288	77,339	652,320	19.48	4.85
15	11	難道(D)	1,265.77	5,861.16	219	20,009	71	729,588	1.08	75.52
16	15	哪(Nep)	1,226.55	4,376.40	289	19,939	354	729,305	1.43	44.95
17	16	何(Nes)	1,154.87	4,307.55	259	19,969	271	729,388	1.28	48.87
18	17	哪裡(Ncd)	1,033.09	4,086.19	217	20,011	180	729,479	1.07	54.66
19	18	有沒有(D)	973.16	3,959.01	198	20,030	145	729,514	0.98	57.73
20	19	究竟(D)	944.68	3,554.16	210	20,018	213	729,446	1.04	49.65
21	23	你們(Nh)	915.39	2,173.73	362	19,866	1,381	728,278	1.79	20.77
22	28	吧(T)	851.01	1,919.64	365	19,863	1,579	728,080	1.80	18.78
23	59	的(DE)	813.92	772.95	4,999	15,229	248,739	480,920	24.71	1.97
24	20	爲何(D)	807.60	3,003.16	182	20,046	193	729,466	0.90	48.53
25	40	知道(VK)	772.87	1,404.77	501	19,727	3,635	726,024	2.48	12.11
26	21	如何(VH)	772.15	2,857.26	175	20,053	189	729,470	0.87	48.08
27	49	會(D)	741.07	1,057.46	990	19,238	12,904	716,755	4.89	7.13
28	22	會不會(D)	711.35	2,774.36	152	20,076	134	729,525	0.75	53.15
29	29	多少(Neqa)	709.86	1,844.61	247	19,981	754	728,905	1.22	24.68
30	48	又(D)	702.45	1,097.02	684	19,544	7,154	722,505	3.38	8.73
31	50	還(D)	691.14	1,036.92	769	19,459	8,872	720,787	3.80	7.98
32	37	妳(Nh)	658.02	1,558.61	262	19,966	1,008	728,651	1.30	20.63
33	24	還是(Caa)	652.50	2,169.88	167	20,061	258	729,401	0.83	39.29
34	31	樣(Nf)	613.09	1,808.57	181	20,047	396	729,263	0.89	31.37
35	38	該(D)	609.59	1,471.49	236	19,992	867	728,792	1.17	21.40
36	47	對(VH)	581.03	1,107.65	341	19,887	2,214	727,445	1.69	13.35
37	43	您(Nh)	580.09	1,223.44	280	19,948	1,428	728,231	1.38	16.39
38	61	要(D)	564.10	761.19	979	19,249	14,751	714,908	4.84	6.22
39	25	怎麼(VH)	517.82	2,169.03	102	20,126	65	729,594	0.50	61.08
40	76	在(P)	482.15	402.42	769	19,459	55,146	674,513	3.80	1.38

Ranking		QRW	Statistic		Count				Recall	Precision
LLR	χ^2	w_i	LLR	χ^2	a	b	c	d	(%)	(%)
41	57	啊(T)	468.43	804.87	351	19,877	2,927	726,732	1.74	10.71
42	27	何必(D)	460.89	2,076.27	83	20,145	34	729,625	0.41	70.94
43	65	那(Nep)	447.72	620.46	695	19,533	9,830	719,829	3.44	6.60
44	26	麼(T)	435.34	2,128.01	69	20,159	11	729,648	0.34	86.25
45	68	有(V_2)	435.24	509.48	1,949	18,279	42,564	687,095	9.64	4.38
46	34	請問(VE)	426.64	1,590.91	96	20,132	101	729,558	0.47	48.73
47	58	呀(T)	407.77	779.29	239	19,989	1,546	728,113	1.18	13.39
48	35	怎樣(VH)	404.89	1,589.15	86	20,142	74	729,585	0.43	53.75
49	32	怎能(D)	380.65	1,744.45	67	20,161	24	729,635	0.33	73.63
50	69	能(D)	369.10	495.63	672	19,556	10,323	719,336	3.32	6.11
51	30	好不好(VH)	363.41	1,809.61	55	20,173	5	729,654	0.27	91.67
52	52	喔(T)	359.12	940.58	124	20,104	372	729,287	0.61	25.00
53	42	怎樣(D)	355.61	1,257.22	85	20,143	108	729,551	0.42	44.04
54	41	甚麼(Nep)	354.78	1,292.21	82	20,146	94	729,565	0.41	46.59
55	36	沒有(T)	351.10	1,567.45	64	20,164	28	729,631	0.32	69.57
56	56	那裡(Ncd)	341.46	848.60	127	20,101	435	729,224	0.63	22.60
57	67	那(Dk)	325.11	584.29	219	20,009	1,643	728,016	1.08	11.76
58	63	眞的(D)	320.43	638.18	173	20,055	1,015	728,644	0.86	14.56
59	39	何在(VH)	316.73	1,409.06	58	20,170	26	729,633	0.29	69.05
60	73	好(VH)	316.41	435.28	515	19,713	7,452	722,207	2.55	6.46
61	45	哪些(Neqa)	312.59	1,186.21	69	20,159	68	729,591	0.34	50.36
62	112	等(Cab)	311.49	187.41	18	20,210	7,949	721,710	0.09	0.23
63	105	以(P)	299.55	210.48	85	20,143	12,914	716,745	0.42	0.65
64	106	及(Caa)	297.46	210.39	90	20,138	13,200	716,459	0.44	0.68
65	44	幹什麼(VA)	286.22	1,188.63	57	20,171	38	729,621	0.28	60.00
66	113	時(Ng)	282.54	185.67	42	20,186	9,429	720,230	0.21	0.44
67	108	之(DE)	274.44	202.31	125	20,103	14,862	714,797	0.62	0.83
68	55	能不能(D)	268.65	860.27	72	20,156	125	729,534	0.36	36.55
69	46	哪兒(Ncd)	263.49	1,120.95	51	20,177	30	729,629	0.25	62.96
70	51	要不要(D)	261.41	944.40	61	20,167	72	729,587	0.30	45.86
71	84	做(VC)	245.85	336.88	411	19,817	6,022	723,637	2.03	6.39
72	131	則(D)	238.51	145.97	17	20,211	6,402	723,257	0.08	0.26
73	129	後(Ng)	221.13	146.49	36	20,192	7,631	722,028	0.18	0.47
74	53	可不可以(D)	201.66	914.99	36	20,192	14	729,645	0.18	72.00
75	119	也(D)	200.22	169.42	440	19,788	29,024	700,635	2.18	1.49
76	141	並(Cbb)	198.04	126.80	23	20,205	6,100	723,559	0.11	0.38
77	137	於(P)	196.81	132.61	38	20,190	7,244	722,415	0.19	0.52
78	54	何謂(VG)	191.01	868.56	34	20,194	13	729,646	0.17	72.34
79	87	在(VCL)	190.76	312.87	164	20,064	1,538	728,121	0.81	9.64
80	79	啊(I)	189.30	369.00	107	20,121	662	728,997	0.53	13.91

Ranking		QRW	Statistic		Count				Recall	Precision
LLR	χ^2	w_i	LLR	χ^2	a	b	c	d	(%)	(%)
81	142	已(D)	185.03	126.10	40	20,188	7,124	722,535	0.20	0.56
82	85	那麼(D)	180.25	322.75	123	20,105	933	728,726	0.61	11.65
83	140	各(Nes)	179.32	127.59	59	20,169	8,294	721,365	0.29	0.71
84	86	這麼(D)	178.58	317.47	124	20,104	957	728,702	0.61	11.47
85	124	一(Neu)	176.64	159.12	1,033	19,195	54,432	675,227	5.11	1.86
86	100	這樣(VH)	176.59	246.22	274	19,954	3,849	725,810	1.35	6.65
87	60	豈不(D)	175.33	763.41	33	20,195	17	729,642	0.16	66.00
88	152	表示(VE)	173.08	111.24	21	20,207	5,408	724,251	0.10	0.39
89	157	因此(Cbb)	170.85	107.10	16	20,212	4,937	724,722	0.08	0.32
90	77	眞(VH)	168.49	386.36	71	20,157	297	729,362	0.35	19.29
91	92	那麼(Dk)	163.81	289.96	115	20,113	897	728,762	0.57	11.36
92	62	怎會(D)	159.10	729.59	28	20,200	10	729,649	0.14	73.68
93	117	這(Nep)	159.04	178.76	1,279	18,949	31,847	697,812	6.32	3.86
94	158	所以(Cbb)	153.70	106.97	41	20,187	6,469	723,190	0.20	0.63
95	154	由(P)	153.69	109.61	52	20,176	7,207	722,452	0.26	0.72
96	97	那些(Neqa)	153.08	262.12	117	20,111	989	728,670	0.58	10.58
97	153	因爲(Cbb)	152.88	110.11	57	20,171	7,511	722,148	0.28	0.75
98	147	年(Nf)	151.01	114.55	95	20,133	9,769	719,890	0.47	0.96
99	104	應該(D)	147.23	211.50	201	20,027	2,620	727,039	0.99	7.13
100	110	事(Na)	146.80	200.69	251	19,977	3,712	725,947	1.24	6.33
101	168	並(D)	143.90	90.49	14	20,214	4,206	725,453	0.07	0.33
102	102	知(VK)	140.29	222.40	134	20,094	1,367	728,292	0.66	8.93
103	64	豈不是(D)	139.34	629.96	25	20,203	10	729,649	0.12	71.43
104	70	何處(Nc)	139.16	459.57	36	20,192	57	729,602	0.18	38.71
105	66	何(D)	138.96	608.06	26	20,202	13	729,646	0.13	66.67
106	116	去(D)	138.07	179.01	316	19,912	5,365	724,294	1.56	5.56
107	155	中(Ng)	136.60	108.49	142	20,086	11,944	717,715	0.70	1.17
108	163	每(Nes)	132.31	96.65	57	20,171	6,989	722,670	0.28	0.81
109	160	與(Caa)	130.66	105.40	163	20,065	12,853	716,806	0.81	1.25
110	96	回(Nf)	128.82	264.12	66	20,162	362	729,297	0.33	15.42
111	133	我(Nh)	128.06	141.17	1,400	18,828	36,897	692,762	6.92	3.66
112	71	何不(D)	127.86	446.84	31	20,197	41	729,618	0.15	43.06
113	169	將(P)	127.86	88.64	33	20,195	5,308	724,351	0.16	0.62
114	151	就(D)	127.59	111.82	494	19,734	28,406	701,253	2.44	1.71
115	114	去(VCL)	124.83	185.24	150	20,078	1,803	727,856	0.74	7.68
116	75	何以(D)	122.34	423.30	30	20,198	41	729,618	0.15	42.25
117	103	要(VC)	122.20	216.21	86	20,142	672	728,987	0.43	11.35
118	186	本(Nes)	120.86	74.83	10	20,218	3,378	726,281	0.05	0.30
119	101	多(Dfa)	119.19	226.77	71	20,157	466	729,193	0.35	13.22
120	134	的(T)	118.32	140.91	514	19,714	10,962	718,697	2.54	4.48

Ranking		QRW	Statistic		Count				Recall	Precision
LLR	χ^2	w_i	LLR	χ^2	a	b	c	d	(%)	(%)
121	189	該(Nes)	116.85	72.57	10	20,218	3,296	726,363	0.05	0.30
122	109	叫(VG)	113.23	201.07	79	20,149	612	729,047	0.39	11.43
123	130	過(Di)	110.67	146.29	227	20,001	3,667	725,992	1.12	5.83
124	170	但(Cbb)	109.06	87.16	123	20,105	10,055	719,604	0.61	1.21
125	180	地(DE)	108.01	80.17	55	20,173	6,229	723,430	0.27	0.88
126	174	將(D)	107.25	82.68	81	20,147	7,720	721,939	0.40	1.04
127	74	幹嘛(VA)	106.49	428.71	22	20,206	17	729,642	0.11	56.41
128	126	沒(D)	104.06	150.14	140	20,088	1,807	727,852	0.69	7.19
129	194	許多(Neqa)	103.59	68.11	16	20,212	3,516	726,143	0.08	0.45
130	127	錢(Na)	102.34	149.77	130	20,098	1,619	728,040	0.64	7.43
131	182	兩(Neu)	99.45	78.59	99	20,129	8,472	721,187	0.49	1.16
132	72	何苦(D)	96.78	444.44	17	20,211	6	729,653	0.08	73.91
133	80	何嘗(D)	96.04	367.79	21	20,207	20	729,639	0.10	51.22
134	183	著(Di)	94.75	77.58	144	20,084	10,651	719,008	0.71	1.33
135	139	在(D)	92.60	127.98	151	20,077	2,175	727,484	0.75	6.49
136	190	三(Neu)	92.56	71.42	71	20,157	6,727	722,932	0.35	1.04
137	192	如果(Cbb)	92.43	68.42	46	20,182	5,268	724,391	0.23	0.87
138	149	想(VE)	91.89	114.06	285	19,943	5,444	724,215	1.41	4.97
139	89	多久(Nd)	91.84	300.78	24	20,204	39	729,620	0.12	38.10
140	187	爲(VG)	91.65	72.80	97	20,131	8,118	721,541	0.48	1.18
141	200	全(Neqa)	90.40	65.07	34	20,194	4,467	725,192	0.17	0.76
142	181	都(D)	90.26	78.88	345	19,883	19,938	709,721	1.71	1.70
143	120	算(VG)	89.45	166.40	56	20,172	388	729,271	0.28	12.61
144	214	如(P)	87.85	58.33	15	20,213	3,093	726,566	0.07	0.48
145	83	爲甚麼(D)	86.98	349.66	18	20,210	14	729,645	0.09	56.25
146	91	那裡(D)	86.76	291.60	22	20,206	33	729,626	0.11	40.00
147	202	中(Ncd)	86.68	63.68	40	20,188	4,749	724,910	0.20	0.84
148	148	來(VA)	86.37	114.49	175	20,053	2,809	726,850	0.87	5.86
149	215	研究(Na)	85.87	58.22	18	20,210	3,272	726,387	0.09	0.55
150	78	知不知道(VK)	85.51	374.18	16	20,212	8	729,651	0.08	66.67
151	198	其(Nep)	84.82	65.96	71	20,157	6,497	723,162	0.35	1.08
152	191	上(Ncd)	84.82	70.07	145	20,083	10,343	719,316	0.72	1.38
153	196	次(Nf)	84.51	66.23	77	20,151	6,817	722,842	0.38	1.12
154	225	同時(Nd)	83.67	53.61	10	20,218	2,606	727,053	0.05	0.38
155	165	我們(Nh)	82.59	93.03	688	19,540	17,169	712,490	3.40	3.85
156	227	未(D)	82.35	52.85	10	20,218	2,578	727,081	0.05	0.39
157	115	憑(P)	81.99	180.53	37	20,191	171	729,488	0.18	17.79
158	82	幹嘛(D)	81.26	359.44	15	20,213	7	729,652	0.07	68.18
159	222	仍(D)	79.58	54.05	17	20,211	3,056	726,603	0.08	0.55
160	221	項(Nf)	78.75	54.44	20	20,208	3,250	726,409	0.10	0.61

Ranking		QRW	Statistic		Count				Recall	Precision
LLR	χ^2	w_i	LLR	χ^2	a	b	c	d	(%)	(%)
161	93	豈(D)	78.60	288.92	18	20,210	20	729,639	0.09	47.37
162	231	不過(Cbb)	78.23	50.93	11	20,217	2,571	727,088	0.05	0.43
163	94	可否(D)	77.39	280.62	18	20,210	21	729,638	0.09	46.15
164	223	內(Ncd)	77.25	53.78	21	20,207	3,284	726,375	0.10	0.64
165	162	跟(P)	77.07	98.92	190	20,038	3,322	726,337	0.94	5.41
166	205	此(Nep)	77.01	60.70	75	20,153	6,476	723,183	0.37	1.14
167	228	前(Ng)	75.53	51.26	16	20,212	2,891	726,768	0.08	0.55
168	88	哪(Ncd)	75.29	312.77	15	20,213	10	729,649	0.07	60.00
169	208	者(Na)	74.97	59.81	84	20,144	6,891	722,768	0.42	1.20
170	95	怎(D)	74.09	271.92	17	20,211	19	729,640	0.08	47.22
171	136	如此(Dfa)	73.71	136.03	47	20,181	332	729,327	0.23	12.40
172	206	上(Ng)	73.34	60.11	114	20,114	8,379	721,280	0.56	1.34
173	81	貴姓(VH)	72.66	361.90	11	20,217	1	729,658	0.05	91.67
174	243	非常(Dfa)	72.22	48.56	14	20,214	2,669	726,990	0.07	0.52
175	232	系統(Na)	71.25	50.90	25	20,203	3,406	726,253	0.12	0.73
176	164	問(VE)	70.86	95.99	128	20,100	1,943	727,716	0.63	6.18
177	235	活動(Na)	70.50	50.21	24	20,204	3,322	726,337	0.12	0.72
178	146	意思(Na)	70.32	117.23	58	20,170	524	729,135	0.29	9.97
179	125	曉得(VK)	70.03	150.39	33	20,195	162	729,497	0.16	16.92
180	135	這麼(Dfa)	69.47	136.05	39	20,189	239	729,420	0.19	14.03
181	99	啥(Nep)	68.40	246.76	16	20,212	19	729,640	0.08	45.71
182	210	和(Caa)	68.24	58.86	216	20,012	13,052	716,607	1.07	1.63
183	209	而(Cbb)	66.89	59.21	325	19,903	17,883	711,776	1.61	1.78
184	226	只(Da)	66.81	53.58	80	20,148	6,413	723,246	0.40	1.23
185	90	豈能(D)	66.39	298.54	12	20,216	5	729,654	0.06	70.59
186	229	公司(Nc)	65.63	51.18	57	20,171	5,144	724,515	0.28	1.10
187	253	而且(Cbb)	62.32	43.12	16	20,212	2,585	727,074	0.08	0.62
188	128	何時(Nd)	62.23	147.20	25	20,203	97	729,562	0.12	20.49
189	144	用(Na)	61.92	124.44	33	20,195	190	729,469	0.16	14.80
190	257	人員(Na)	61.87	41.94	13	20,215	2,361	727,298	0.06	0.55
191	143	能否(D)	60.76	124.93	31	20,197	169	729,490	0.15	15.50
192	195	了(T)	60.49	67.46	595	19,633	15,308	714,351	2.94	3.74
193	251	下(Ng)	60.19	44.07	27	20,201	3,252	726,407	0.13	0.82
194	244	卻(D)	59.78	48.48	82	20,146	6,279	723,380	0.41	1.29
195	238	所(D)	58.86	49.48	129	20,099	8,571	721,088	0.64	1.48
196	156	修(VC)	58.84	108.00	38	20,190	272	729,387	0.19	12.26
197	123	蹺家(VA)	58.82	162.21	19	20,209	50	729,609	0.09	27.54
198	260	便(D)	58.42	41.40	19	20,209	2,695	726,964	0.09	0.70
199	98	豈非(D)	57.74	249.05	11	20,217	6	729,653	0.05	64.71
200	255	一(D)	56.03	43.00	41	20,187	3,965	725,694	0.20	1.02

Ranking		QRW	Statistic		Count				Recall	Precision
LLR	χ^2	w_i	LLR	χ^2	a	b	c	d	(%)	(%)
201	262	市場(Nc)	55.57	40.35	23	20,205	2,882	726,777	0.11	0.79
202	274	包括(VK)	55.25	37.24	11	20,217	2,064	727,595	0.05	0.53
203	178	怕(VK)	54.95	81.02	68	20,160	832	728,827	0.34	7.56
204	121	抑或(Caa)	54.79	163.58	16	20,212	34	729,625	0.08	32.00
205	252	目前(Nd)	54.76	43.31	56	20,172	4,752	724,907	0.28	1.16
206	259	使(VL)	52.60	41.61	54	20,174	4,576	725,083	0.27	1.17
207	171	樣子(Na)	52.32	84.71	47	20,181	456	729,203	0.23	9.34
208	176	看法(Na)	51.94	81.26	52	20,176	549	729,110	0.26	8.65
209	276	之後(Ng)	51.80	36.96	18	20,210	2,466	727,193	0.09	0.72
210	138	有無(VJ)	51.03	131.88	18	20,210	56	729,603	0.09	24.32
211	270	大學(Nc)	50.94	38.29	30	20,198	3,185	726,474	0.15	0.93
212	249	很(Dfa)	50.65	44.54	223	20,005	12,532	717,127	1.10	1.75
213	269	二(Neu)	50.51	38.34	33	20,195	3,348	726,311	0.16	0.98
214	107	幹麼(D)	49.61	204.04	10	20,218	7	729,652	0.05	58.82
215	111	何去何從(VA)	49.59	187.55	11	20,217	11	729,648	0.05	50.00
216	277	必須(D)	48.70	36.88	31	20,197	3,181	726,478	0.15	0.97
217	172	哪(T)	48.62	84.39	36	20,192	295	729,364	0.18	10.88
218	184	有(D)	47.55	75.72	45	20,183	455	729,204	0.22	9.00
219	217	現在(Nd)	47.39	57.01	193	20,035	4,025	725,634	0.95	4.58
220	197	意義(Na)	47.38	66.16	74	20,154	1,040	728,619	0.37	6.64
221	292	今年(Nd)	47.15	32.58	12	20,216	1,949	727,710	0.06	0.61
222	281	中心(Nc)	47.08	35.30	27	20,201	2,901	726,758	0.13	0.92
223	285	名(Nf)	46.97	33.22	15	20,213	2,149	727,510	0.07	0.69
224	272	小(VH)	46.81	37.89	63	20,165	4,857	724,802	0.31	1.28
225	204	找(VC)	46.69	61.57	98	20,130	1,597	728,062	0.48	5.78
226	132	何方(Ncd)	46.48	143.78	13	20,215	25	729,634	0.06	34.21
227	122	哪(D)	45.96	162.49	11	20,217	14	729,645	0.05	44.00
228	278	使用(VC)	45.88	36.60	52	20,176	4,253	725,406	0.26	1.21
229	167	發言(VA)	45.84	91.07	25	20,203	148	729,511	0.12	14.45
230	185	懂(VK)	45.64	75.17	39	20,189	363	729,296	0.19	9.70
231	161	遠景(Na)	45.23	104.02	19	20,209	79	729,580	0.09	19.39
232	286	國內(Nc)	45.09	33.12	21	20,207	2,485	727,174	0.10	0.84
233	118	第幾(Neu)	45.08	170.50	10	20,218	10	729,649	0.05	50.00
234	289	無(VJ)	45.05	32.94	20	20,208	2,422	727,237	0.10	0.82
235	304	昨天(Nd)	44.95	31.21	12	20,216	1,897	727,762	0.06	0.63
236	302	技術(Na)	44.36	31.56	15	20,213	2,085	727,574	0.07	0.71
237	216	還是(D)	44.11	57.43	100	20,128	1,686	727,973	0.49	5.60
238	296	電腦(Na)	43.32	32.09	22	20,206	2,499	727,160	0.11	0.87
239	287	進行(VC)	43.31	32.95	29	20,199	2,912	726,747	0.14	0.99
240	282	工作(Na)	43.21	34.96	58	20,170	4,476	725,183	0.29	1.28

Ranking		QRW	Statistic		Count				Recall	Precision
LLR	χ^2	w_i	LLR	χ^2	a	b	c	d	(%)	(%)
241	299	無法(D)	42.97	32.00	23	20,205	2,550	727,109	0.11	0.89
242	173	嬉皮(Na)	42.76	83.76	24	20,204	147	729,512	0.12	14.04
243	301	較(Dfa)	42.72	31.69	22	20,206	2,483	727,176	0.11	0.88
244	177	那麼多(Neqa)	42.41	81.16	25	20,203	162	729,497	0.12	13.37
245	305	資料(Na)	42.35	31.15	20	20,208	2,351	727,308	0.10	0.84
246	218	啦(T)	42.31	57.00	79	20,149	1,218	728,441	0.39	6.09
247	211	一下(Nd)	42.28	58.67	68	20,160	971	728,688	0.34	6.54
248	145	咦(I)	42.22	120.84	13	20,215	31	729,628	0.06	29.55
249	166	看待(VC)	42.03	92.47	19	20,209	88	729,571	0.09	17.76
250	220	買(VC)	41.41	55.43	80	20,148	1,254	728,405	0.40	6.00
251	312	歲(Nf)	41.35	30.17	18	20,210	2,201	727,458	0.09	0.81
252	316	其實(D)	41.00	29.77	17	20,211	2,129	727,530	0.08	0.79
253	240	孩子(Na)	40.28	48.99	149	20,079	3,015	726,644	0.74	4.71
254	310	提供(VD)	39.94	30.62	29	20,199	2,815	726,844	0.14	1.02
255	245	覺得(VK)	39.89	47.14	193	20,035	4,228	725,431	0.95	4.37
256	175	猜(VE)	39.77	82.42	20	20,208	107	729,552	0.10	15.75
257	329	服務(VC)[+nom]	39.60	27.86	12	20,216	1,769	727,890	0.06	0.67
258	322	國際(Nc)	39.41	28.72	17	20,211	2,088	727,571	0.08	0.81
259	150	乎(T)	39.38	113.61	12	20,216	28	729,631	0.06	30.00
260	303	美國(Nc)	39.07	31.26	46	20,182	3,713	725,946	0.23	1.22
261	320	資訊(Na)	38.96	29.03	21	20,207	2,321	727,338	0.10	0.90
262	337	來(Ng)	38.34	26.55	10	20,218	1,602	728,057	0.05	0.62
263	236	至於(P)	37.96	50.15	79	20,149	1,282	728,377	0.39	5.80
264	321	發現(VE)	37.88	28.88	26	20,202	2,584	727,075	0.13	1.00
265	242	東西(Na)	37.84	48.64	93	20,135	1,622	728,037	0.46	5.42
266	295	與(P)	37.82	32.21	101	20,127	6,374	723,285	0.50	1.56
267	331	地區(Nc)	37.77	27.80	18	20,210	2,107	727,552	0.09	0.85
268	291	可(D)	37.55	32.70	142	20,086	8,251	721,408	0.70	1.69
269	334	家(Nf)	37.55	26.77	13	20,215	1,785	727,874	0.06	0.72
270	201	可以(VH)	36.68	63.82	27	20,201	220	729,439	0.13	10.93
271	159	怎麼說(VH)	36.54	106.45	11	20,217	25	729,634	0.05	30.56
272	347	六(Neu)	36.20	25.70	12	20,216	1,686	727,973	0.06	0.71
273	207	好處(Na)	36.05	59.93	30	20,198	273	729,386	0.15	9.90
274	343	多(Neqa)	35.54	26.01	16	20,212	1,925	727,734	0.08	0.82
275	203	會(VL)	35.36	62.51	25	20,203	196	729,463	0.12	11.31
276	317	學生(Na)	35.31	29.73	80	20,148	5,273	724,386	0.40	1.49
277	323	向(P)	35.21	28.66	51	20,177	3,840	725,819	0.25	1.31
278	213	些(Dfb)	34.71	58.34	28	20,200	248	729,411	0.14	10.14
279	266	自己(Nh)	34.63	38.74	333	19,895	8,516	721,143	1.65	3.76
280	314	更(D)	34.55	29.87	117	20,111	6,971	722,688	0.58	1.65

Ranking		QRW	Statistic		Count				Recall	Precision
LLR	χ^2	w_i	LLR	χ^2	a	b	c	d	(%)	(%)
281	342	服務(VC)	33.80	26.04	26	20,202	2,465	727,194	0.13	1.04
282	338	經濟(Na)	33.79	26.46	31	20,197	2,742	726,917	0.15	1.12
283	349	第二(Neu)	33.56	25.48	22	20,206	2,230	727,429	0.11	0.98
284	319	被(P)	33.53	29.08	119	20,109	7,015	722,644	0.59	1.67
285	250	人生(Na)	33.36	44.48	66	20,162	1,046	728,613	0.33	5.94
286	179	翹課(VA)	33.36	80.53	13	20,215	48	729,611	0.06	21.31
287	328	天(Nf)	33.34	27.97	72	20,156	4,809	724,850	0.36	1.48
288	193	談談(VE)	33.22	68.21	17	20,211	93	729,566	0.08	15.45
289	368	下(Ncd)	32.94	23.61	12	20,216	1,605	728,054	0.06	0.74
290	361	整(Neqa)	32.92	24.28	16	20,212	1,856	727,803	0.08	0.85
291	354	開始(VL)	32.85	25.19	24	20,204	2,324	727,335	0.12	1.02
292	219	用途(Na)	32.35	56.11	24	20,204	197	729,462	0.12	10.86
293	230	打算(VF)	32.28	51.16	31	20,197	317	729,342	0.15	8.91
294	372	以(Cbb)	32.11	23.43	14	20,214	1,711	727,948	0.07	0.81
295	358	以及(Caa)	32.10	24.93	27	20,201	2,470	727,189	0.13	1.08
296	359	對於(P)	32.06	24.91	27	20,201	2,469	727,190	0.13	1.08
297	352	網路(Na)	31.93	25.32	34	20,194	2,845	726,814	0.17	1.18
298	241	管(VE)	31.74	48.70	34	20,194	377	729,282	0.17	8.27
299	224	那兒(Ncd)	31.70	53.74	25	20,203	217	729,442	0.12	10.33
300	247	活(VH)	31.68	45.43	44	20,184	578	729,081	0.22	7.07

REFERENCES

[1] Agresti, A., *Categorical Data Analysis*. Hoboken, New Jersey: John Wiley & Sons, second ed., 2002.

[2] ARPA Knowledge Sharing Initiative, *Specification of the KQML Agent-Communication Language*. ARPA Knowledge Sharing Initiative, External Interfaces Working Group, July 1993.

[3] Chang, L.-l., "Modal words in modern Mandarin," Tech. Rep. 93-06, Institute of Information Science, Academia Sinica, Taipei, June 1993.

[4] Chen, F.-Y., Tsai, P.-F., Chen, K.-J., and Huang, C.-R., "Construction of the Sinica treebank," *Computational Linguistics and Chinese Language Processing*, vol. 4, no. 2, pp. 87–104, 1999.

[5] Chen, K.-J., Luo, C.-C., Chang, M.-C., Chen, F.-Y., Chen, C.-J., Huang, C.-R., and Gao, Z.-M., "Sinica treebank: Design criteria, representational issues and implementation," in *Treebanks: Building and Using Parsed Corpora* (Abeillé, A., ed.), ch. 13, pp. 231–248, Kluwer Academic Publishers, 2003.

[6] Chen, S. F. and Goodman, J. T., "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 13, pp. 359–394, Oct. 1999.

[7] Cheng, R. L., "Chinese question forms and their meanings," *Journal of Chinese Linguistics*, vol. 12, pp. 86–147, 1984. Also in *Temporal and Spatial Relations, Questions and Negatives in Taiwanese and Mandarin*, Yuan-Liou Publishing Co., Taipei, pp. 273–312, 1997.

[8] Cheng, R. L., "Mandarin and Taiwanese question sentences," in *Temporal and Spatial Relations, Questions and Negatives in Taiwanese and Mandarin*, pp. 357–402, Taipei: Yuan-Liou Publishing Co., 1997.

[9] Chinese Knowledge Information Processing Group, "Analysis of Chinese word classes," Tech. Rep. 93-05, Institute of Information Science, Academia Sinica, Taipei, 1993.

[10] Chinese Knowledge Information Processing Group, "A guide to the Academia Sinica balanced corpus of modern Chinese," Tech. Rep. 95-02/98-04, Institute of Information Science, Academia Sinica, Taipei, 1998.

[11] Chu, C. C., *A Concise Grammar of Mandarin Chinese*. Taipei: Wu-Nan Book Inc, 1999.

[12] Chu, C. C. and Chi, T.-J., *A Cognitive-Functional Grammar of Mandarin Chinese*. Taipei: Crane Publishing Company, Inc, 1999.

[13] Chu, C.-N., *Mandarin Lexicology.* Taipei: Wu-Nan Book Inc, Oct. 1999.

[14] Churcher, G. E., Atwell, E. S., and Souter, C., "Dialogue management systems: A survey and overview," Tech. Rep. 97.06, University of Leeds, UK, Feb. 1997.

[15] DeFrancis, J., ed., *ABC Chinese-English Dictionary.* Honolulu: University of Hawaii Press, 1996.

[16] Dietz, J. L., "The constituents of business interaction–generic layered patterns," *Data & Knowledge Engineering*, vol. 47, pp. 301–325, 2003.

[17] Dunning, T., "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*, vol. 19, no. 1, pp. 61–74, 1993.

[18] EZ Information, "The affiliated Mandarin phrase center of EZ Input." `http://phrasecenter.freehosting.net/`.

[19] Fan, X. and others, eds., *Studies on Mandarin Sentence Types.* Taiyuan, Shanxi, China: Shuhai Publishing House, 1998.

[20] Fellbaum, C., ed., *WordNet: An Electronic Lexical Database.* Cambridge, Mass: The MIT Press, 1998. The service can be accessed on-line at `http://www.cogsci.princeton.edu/~wn/`.

[21] Friedl, J. E. F., *Mastering Regular Expressions.* CA: O'Reilly & Associates, second ed., 2002.

[22] Gan, K. W. and Tham, W. M., "General knowledge annotation based on How-net," *Computational Linguistics and Chinese Language Processing*, vol. 4, no. 2, pp. 39–86, 1999.

[23] Gan, K. W. and Wong, P. W., "Annotation guidelines of Chinese message structures based on HowNet," tech. rep., Department of Computer Science, Hong Kong University of Science and Technology, Hong Kong, Nov. 2000.

[24] Genesereth, M. R. and Fikes, R. E., "Knowledge Interchange Format version 3.0 reference manual," Tech. Rep. Logic-92-1, Interlingua Working Group, DARPA Knowledge Sharing Effort, June 1992.

[25] Goodman, J. T., "A bit of progress in language modeling," *Computer Speech and Language*, vol. 15, pp. 403–434, Oct. 2001.

[26] Hsiao, P.-H., Tsai, C.-H., Hsieh, T.-H., Yeh, P.-J., and Tan, K.-S., "The TaBE project." `http://xcin.linux.org.tw/libtabe/`.

[27] Huang, C.-R., Chen, K.-J., Chen, F.-Y., and Chang, L.-L., "Segmentation standard for Chinese natural language processing," *Computational Linguistics and Chinese Language Processing*, vol. 2, pp. 47–62, Aug. 1997.

[28] Jurafsky, D. and Martin, J. H., *Speech and Language Processing.* New Jersey: Prentice Hall, 2000.

[29] Labrou, Y. and Finin, T., "A proposal for a new KQML specification," Tech. Rep. CS-97-03, Computer Science and Electrical Engineering Department, University of Maryland Baltimore County, Baltimore, Maryland, Feb. 1997.

[30] Li, C. N. and Thompson, S. A., *Mandarin Chinese: A Functional Reference Grammar.* New York: University of California Press, 1981.

[31] Lin, F.-W., "Some reflections on the thematic system of information-based case grammar (ICG)," Tech. Rep. 92-01, Institute of Information Science, Academia Sinica, Taipei, Aug. 1992.

[32] Lind, M. and Goldkuhl, G., "The atoms, molecules and fibers of organizations," *Data & Knowledge Engineering*, vol. 47, pp. 327–348, 2003.

[33] Litkowski, K. C., "Use of metadata for question answering and novelty tasks," in *Proceedings of the 12th Text REtrieval Conference (TREC 2003)*, (Gaithersburg, Maryland), pp. 161–170, Nov. 2003.

[34] Liu, Y., Pan, W., and Gu, W., *Modern Chinese Grammar.* Beijing: Foreign Language Teaching and Research Press, 1983. Reprinted in traditional Chinese by Shih Ta Book Ltd, Taipei, 1996.

[35] Loos, E. E., Anderson, S., Dwight H., Jr., D., Jordan, P. C., and Wingate, J. D., "Glossary of linguistic terms," in *LinguaLinks Library, Version 5.0*, SIL International, 2003. Its content can also be accessed on-line at `http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/`.

[36] Luo, Z. and others, eds., *Unabridged Mandarin Dictionary.* Shanghai: The Publishing House of the Unabridged Chinese Dictionary, 1993. Reprinted in traditional Chinese by Taiwan Doughua Bookstore, Taipei, 1997.

[37] Lyu, S., *An Outline of Chinese Grammar.* Beijing: The Commercial Press, 1982.

[38] Lyu, S. and others, eds., *Eight Hundred Words of Modern Mandarin.* Beijing: The Commercial Press, revised ed., 1999.

[39] Mandarin Promotion Council, *A Manual of Punctuation Marks Revised.* Taipei: Ministry of Education, 1987.

[40] Mandarin Promotion Council, *Mandarin Dictionary Revised.* Taipei: Ministry of Education, fourth ed., Apr. 1998. The on-line version can be accessed at `http://www.edu.tw/mandr/clc/dict/`.

[41] Manning, C. D. and Schütze, H., *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press, Cambridge, MA, 1999.

[42] Mao, J.-j., Chen, Q.-l., and Lu, R.-z., "Formal representation and semantics of modern Chinese interrogative sentences," *Lecture Notes in Computer Science*, vol. 2588, pp. 65–74, 2003.

[43] Meng, I.-H., *Design and Study of Semantic Discovery Methods for Extracting Knowledge from Free Text Information.* Ph.D. dissertation, National Chiao Tung University, Hsinchu, Taiwan, July 2003.

[44] Moldovan, D., Harabagiu, S., Girju, R., Morarescu, P., Lacatusu, F., Novischi, A., Badulescu, A., and Bolohan, O., "LCC tools for question answering," in *Proceedings of the 11th Text REtrieval Conference (TREC 2002)*, (Gaithersburg, Maryland), pp. 388–397, Nov. 2002.

[45] National Standard of the People's Republic of China, *GB/T 15834-1995: Use of Punctuation Marks*, 1995.

[46] Pedersen, T., "Fishing for exactness," in *Proceedings of the South Central SAS User's Group Conference (SCSUG-96)*, (Austin, TX), pp. 188–200, Oct. 1996.

[47] Pickett, J. P. and others, eds., *The American Heritage Dictionary of the English Language.* Boston, MA: Houghton Mifflin Company, fourth ed., 2000.

[48] Shen, T.-Z., *A Study of Why Questions in Web-Based Question Answering System.* Master thesis, National Taiwan University, Taipei, Taiwan, 2003.

[49] Tsai, C.-H., "A review of Chinese word lists accessible on the Internet." online document, Oct. 2003. http://www.geocities.com/hao510/wordlist/.

[50] Voorhees, E. M., "Overview of the TREC-8 question answering track," in *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, (Gaithersburg, Maryland), pp. 77–82, Nov. 1999.

[51] Voorhees, E. M., "Overview of the TREC-9 question answering track," in *Proceedings of the 9th Text REtrieval Conference (TREC-9)*, (Gaithersburg, Maryland), pp. 71–79, Nov. 2000.

[52] Voorhees, E. M., "Overview of the TREC 2001 question answering track," in *Proceedings of the 10th Text REtrieval Conference (TREC 2001)*, (Gaithersburg, Maryland), pp. 42–51, Nov. 2001.

[53] Voorhees, E. M., "Overview of the TREC 2002 question answering track," in *Proceedings of the 11th Text REtrieval Conference (TREC 2002)*, (Gaithersburg, Maryland), pp. 57–67, Nov. 2002.

[54] Voorhees, E. M., "Overview of the TREC 2003 question answering track," in *Proceedings of the 12th Text REtrieval Conference (TREC 2003)*, (Gaithersburg, Maryland), pp. 54–68, Nov. 2003.

[55] Weeber, M., Baayen, R. H., and Vos, R., "Extracting the lowest-frequency words: Pitfalls and possibilities," *Computational Linguistics*, vol. 26, no. 3, pp. 301–317, 2000.

[56] Winograd, T., "A language/action perspective on the design of cooperative work," *Human-Computer Interaction*, vol. 3, no. 1, pp. 3–30, 1987–1988.

[57] Wu, B., *Prescriptive Usage of Punctuation Marks.* Hong Kong: Joint Publishing (H.K.) Co., 1998.

[58] XTAG Research Group, "A lexicalized tree adjoining grammar for English," Tech. Rep. IRCS-01-03, IRCS, University of Pennsylvania, 2001.

[59] Xu, Z. and others, eds., *Unabridged Dictionary of Chinese Characters.* China: Sichuan Dictionary Publishing House and Hubei Dictionary Publishing House, 1995.

[60] Yen, Y.-S., ed., *Mass Modern Chinese-English Dictionary.* Taipei: Mass Publishing Company, 1988.

[61] Zhang, B., "Functional interpretation of Mandarin question sentences," in *Various Characteristics of Mandarin Grammar* (Xing, F., ed.), pp. 291–303, Beijing: Beijing Language and Culture University Publishing House, 1999.

[62] Zhang, Z., *Common Sense of Mandarin Grammar.* Shanghai: Shanghai Education Publishing House, 1959. Reprinted in traditional Chinese by Joint Publishing (H.K.) Co., Hong Kong, 1999.