International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI, Linguistics Abstracts, and ACL Anthology.

易繫辭曰上古結繩而
治後世聖人易之以書
契百官以治萬民以察
說文敘曰蓋文字者經
藝之本宣教明化之始
前人所以垂後後人所
以識古故曰本立而道
生知天下之至賾而不
可亂也教化既萌文心
雕龍則謂人之立言因
字而生句積句而成章
積章而成篇篇之彪炳

Vol.22 No.1 June 2017 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

International Journal of

Computational Linguistics & Chinese Language Processing

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP
Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.
This calligraphy honors the interaction and influence between text and language

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 22, no. 1, June 2017

Contents

Papers

Computational Linguistics and Chinese Language Processing
Vol. 22, No. 1, June 2017, pp. 1-26

當代非監督式方法之比較於節錄式語音摘要

An Empirical Comparison of Contemporary Unsupervised Approaches for Extractive Speech Summarization

劉士弘*、陳冠宇*、施凱文+、陳柏琳+、王新民*、許聞廉*

Shih-Hung Liu, Kuan-Yu Chen, Kai-Wun Shih, Berlin Chen,

Hsin-Min Wang, and Wen-Lian Hsu

摘要

由於網際網路的飛速發展，促成大資料時代的來臨，也因此自動摘要(Automatic Summarization)成為近年來一項熱門的研究議題。節錄式(Extractive)自動摘要是依據事先定義的摘要比例，從文字文件(Text Documents)或語音文件(Spoken Documents)中選取一些能夠代表原始文件主旨或主題的重要語句當作摘要。節錄式摘要可被視為一個資訊檢索(Information Retrieval, IR)的問題，在相關研究中，使用語言模型(Language Modeling)來挑選重要語句之方法，已初步地被驗證在文字與語音文件的自動摘要任務上有不錯的成果。本論文延續此項研究，進一步地提出三個主要的研究貢獻。首先，有鑑於關聯性(Relevance)資訊的概念在資訊檢索領域中已有不錯的發展成果，本論文嘗試結合關聯性資訊來重新估測並建立語句的語言模型，並嘗試使用三混合(Tri-Mixture Model, TriMM)模型，期待得以更精準地描述語句的語意內容，進而提升自動摘要之效能。第二，除了語言模型之外，本論文進一步地嘗試探究機率式檢索模型於語音文件

* 中央研究院資訊所
Institute of Information Science, Academia Sinica
E-mail: journey0621@gmail.com, {kychen, whm, hsu}@iis.sinica.edu.tw
+ 國立臺灣師範大學資訊工程系
Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: {60247065S, berlin}@csie.ntnu.edu.tw

摘要任務上之成效。最後，本論文亦探討不同的語言模型平滑化技術對於語音文件摘要任務之影響。本論文的語音文件摘要實驗語料是採用公視廣播新聞(MATBN)；實驗結果顯示，相較於其它現有的非監督式摘要方法，我們所應用的新穎式摘要方法能提供明顯的效能改善。

關鍵字: 最佳匹配、語言模型、虛擬關聯回饋、關聯模型、節錄式自動摘要。

Abstract

Due to the rapid-developed Internet and with the big data era coming, the automatic summarization research has been emerged a popular research topic. The aim of automatic summarization is in attempt to select important text or spoken sentence to represent the topic (theme) of original text or spoken document according to a predefined summarization ratio. In this study we frame automatic summarizaiton task as an ad-hoc information retrieval (IR) problem and employ the mathematical sound language modeling (LM) framework for extractive speech summarization, which can perform important sentence selection in an unsupervised manner and has shown its preliminary success. The main contribution of this paper is three-fold. First, by the virtue of relevance modeling, we explore several effective sentence modeling formulations to enhance the sentence models involved in the LM-based summarization framework and the first use of tri-mixture model to improve the performance of extractive speech summarization. Second, since the language modeling will suffer from data sparseness problem and the common solution is to adopt smoothing techniques, in this research we investigate three different smoothing approaches to evaluate how they influence the summarization performance. Third, we further apply the well-studied ranking model (BM25) and also its variants in IR community for ranking important sentence in extractive speech summarization. Experiments conducted on public avaiable dataset (MATBN) and the results show that our applied methods have effective summarization performance when compared to the other well-practiced and state-of-the-art unsupervised methods.

Keywords: BM25, Language Modeling, Pseudo-Relevance Feedback, Relevance Modeling, Extractive Automatic Summarization.

1. 緒論 (Introduction)

隨著大資料時代的來臨，眾多的文字及多媒體影音資訊被快速地傳遞並分享於全球各地，資訊超載(Information Overload)的問題也隨之產生。如何能讓人們快速且有效率地瀏覽與日俱增的文字資訊或多媒體影音資訊，已成為一個刻不容緩的研究課題。在眾多的研究

方法中，自動摘要(Automatic Summarization)被視為是一項不可或缺的關鍵技術(Lin & Chen, 2010; Liu & Hakkani-Tur, 2011)。自動摘要之目的在於擷取單一文件(Single-Document)或多重文件(Multi-Document)中的重要語意與主題資訊，藉此讓使用者能更有效率地瀏覽與理解文件的主旨，以便快速地獲得其所需的資訊，避免花費大量時間在審視文件內容。另一方面，語音是多媒體文件中最具資訊的成分之一；如何透過語音(文件)摘要技術來自動地、有效率地處理具時序性的多媒體影音內容，例如：電視新聞、廣播新聞、郵件、電子郵件、會議及演講錄音等(Ostendorf, 2008; Nenkova & McKeown, 2011)，更是顯得非常重要。其關鍵原因在於多媒體影音內容往往長達數分鐘或數小時，使用者不易於瀏覽和查詢，而必須耐心地閱讀或聽完整份多媒體影音內容，才能理解其中所描述的語意與主題，這違反人們講求方便、有效率的資訊獲取方式。

雖然對於含有語音訊號的多媒體影音，我們可透過自動語音辨識(Automatic Speech Recognition, ASR)技術自動地將其轉換成易於瀏覽的文字內容，再藉由文字文件摘要的技術來做處理，以達到摘要多媒體影音或其它語音文件之目的。但就現階段語音辨識技術的發展，語音文件經語音辨識後自動轉寫成文字的結果，不僅存在辨識錯誤的問題，也缺乏章節與標點符號，使得語句邊界定義不清楚而失去文件的結構資訊；除此之外，語音文件通常含有許多口語語助詞、遲疑、重覆等內容，這都使得語音文件摘要技術的發展面臨更多的挑戰。

一般來說，自動摘要研究可從許多不同面相來進行探討，包括了來源、需求、方式、用途以及模型技術，以下將簡述各個不同面相的相關議題(Mani & Maybury, 1999)：

(1) 來源：根據文件來源，可以分為單一文件摘要與多重文件摘要(Cai & Li, 2013)；單一文件摘要是依據事先定義好的摘要比例，選取能夠代表文件的句子當作摘要；而多重文件摘要是收集多篇相似的文件，需要移除文件間彼此冗餘性(Redundancy)的資訊(Carbonell & Goldstein, 1998)，考慮文件描述事件發生的先後順序(Causality)(Kuo & Chen, 2006)，並且確認文件之間的因果關係，經由這些資訊希望能產生有連貫性的文件摘要。

(2) 需求：依據使用者需求不同，摘要內容可區分為具有資訊性(Informative)、指示性(Indicative)、以及評論性(Critical)。具有資訊性的摘要是用來表達文件描述的主旨內容與核心資訊；具指示性的摘要是希望將文件中的主題內容做簡單的描述，並將文件分成不同的主題，例如：政治性、學術性、體育性和娛樂性文件，因此所產生的摘要不要求傳達詳細的原始文件內容；具評論性的摘要提供文件正面與反面的觀點(Positive and Negative Sentiments)(Galley, McKeown, Hirschberg & Shriberg, 2004)。

(3) 方式：可概分為二大類，節錄式(Extractive)摘要與抽象式(Abstractive)摘要(或重寫式摘要)。前者主要是依據特定的摘要比例，從最原始的文件中選取重要的語句來組成摘要；而後者是在完全理解文件內容之後，重新撰寫產生摘要來代表原始文件的內容，其所使用之語彙或慣用語不一定是全然地來自於原始文件，此種摘要方式是最為貼近人們日常撰寫摘要的形式。然而抽象式摘要需要複雜的自然語言處理(Natural Language Processing, NLP)技術，如資訊擷取(Information Extraction)、對話理解(Discourse Understanding)及自然語言

生成(Natural Language Generation)等(Paice, 1990; Witbrock & Mittal, 1999)，因此，近年來節錄式摘要之研究仍為主流。

(4) 用途：依摘要用途可分為一般性(Generic)摘要與以查詢為基礎(Query-focused)的摘要。前者是從整篇文件中萃取出能夠突顯整篇文件全面性主題資訊的語句，期望摘要產生的內容可以涵蓋整篇文件所有重要的主題；後者透過使用者或特定的查詢來產生與查詢相關的摘要。

(5) 模型技術：簡單分成三大類，(i)以簡單的語彙(Lexical)與結構(Structural)特徵做為判斷摘要語句的模型技術(Zhang, Chan & Fung, 2010)，(ii)監督式機器學習(Supervised Machine Learning)以及(iii)非監督式機器學習(Unsupervised Machine Learning)(Liu & Hakkani-Tur, 2011)之模型技術。雖然非監督式機器學習的方法在一般的情況下其效能沒有監督式機器學習方法來的好，但非監督式機器學習方法不需要事先準備大量人工標記的訓練資料，以及具有容易實作(Easy-to-Implement)的特性，仍吸引許多學者進行研究與發展，本論文主要也是探討且比較非監督式機器學習的方式於自動摘要之任務。

綜觀上述各個面向，本論文主要探究一般性、單一文件節錄式語音摘要問題，並比較各式非監督式機器學習模型技術。近年來，各式基於語言模型之非監督式模型技術運用在資訊檢索領域中已呈現卓越的研究成果(Zhai, 2008)，這些技術也初步地被應用於語音文件摘要之研究上(Lin, Yeh & Chen, 2011)，亦獲得一定的摘要成效。本論文將延續此一研究主軸，提出三個主要的研究貢獻。首先，有鑑於關聯性(Relevance)資訊的概念已被應用於資訊檢索領域之中(Zhai & Lafferty, 2001a; Lavrenko & Croft, 2001)，本論文嘗試結合關聯性資訊來重新估測並建立語句的語言模型，並首次使用三混合(Tri-Mixture Model, TriMM)模型，期待得以更精準地估測語句的語意內容，增進自動摘要之成效。當語言模型使用最大化相似度估測時，很可能會遭遇資料稀疏(Data Sparseness)的問題，而使得模型無法準確地估測每一個詞彙真正的機率分佈，也可能因為某些詞彙的條件機率值為零，導致無法準確地計算語句與文件間的相似度。為此，語言模型平滑化技術被提出來減輕上述的現象。過去的研究中顯示，各式平滑化技術的使用時常在各項任務中扮演關鍵的腳色。有鑑於此，本論文首次比較不同平滑化技術對於語音文件摘要任務之影響。最後，除了語言模型的探討之外，我們進一步地提出並使用多種機率式檢索模型於語音摘要任務上。本論文後續安排如下：第二章扼要地介紹現今自動摘要模型技術的相關研究與發展；第三章介紹使用語言模型於節錄式語音摘要任務之原理，然後闡述如何藉助語句關聯性資訊來改進語句模型之估測，使其得以更精準地代表語句的語意內容；第四章介紹多種機率式排序模型並將之應用至語音文件摘要任務中；第五章介紹實驗語料與設定以及摘要評估之方法；第六章說明實驗結果及其分析；最後，第七章為結論與未來研究方向。

2. 自動摘要模型技術 (Techniques of Automatic Summarization)

本論文將過去摘要研究所陸續發展出的自動摘要模型技術大略地歸納成三大類(Mani & Maybury, 1999)：

(1) 以簡單詞彙與結構特徵為基礎之自動摘要模型技術：在 1950 年代，有學者提出使用詞頻(Frequency)來評量每一個詞的重要性與計算文件中每一個語句的顯著性(Significance Factor)(Luhn, 1958)。在實作上，可以對每一個詞進行詞幹分析(Stemming)，將其還原成詞根(Root Form)，同時移除停用詞(Stop Word)的影響並計算實詞(Content Word)的重要性等，最後將語句依其顯著分數進行排序(由高至低)，再根據特定的摘要比例來進行節錄式摘要的產生。後來，有學者利用自然語言分析(Natural Language Analysis)技術對文件結構進行剖析，根據文法結構(Grammar Structure)與語言機制(Linguistic Devices)來決定不同語段的凝聚關係(Cohesion)，例如：首語重複(Anaphora)、省略(Ellipsis)、結合(Conjunction)，或同義詞(Synonymy)、上義詞(Hypernym)等語彙關係(Lexical Relation)，並以此結果進行文件自動摘要。相關研究包括使用語彙鏈(Lexical Chain)(Barzilay & Elhadad, 1997)、宏觀語段結構(Discourse Macro Structure)(Strzalkowski, Wand & Wise, 1998)、修辭結構(Rhetorical Structure)(Zhang *et al*., 2010)等。另有學者在審視 200 篇科技文件後，發現有 85%的重要語句出現在文件中的第一段，7%的重要語句出現在最後一段(Baxendale, 1958)。因此，提出了語句在文件中的位置(Position)資訊是進行摘要語句選取時的一項關鍵線索。

(2) 以非監督式機器學習為基礎之自動摘要模型技術：非監督式機器學習通常將自動摘要任務視為如何排序並挑選具代表性語句之問題，其方法通常是計算出一種摘要特徵供語句排序使用，常見的特徵有：語句與文件相關性(Gong & Liu, 2001)、語句所形成的語言模型生成文件之機率等(Chen, Chen & Wang, 2009)、語句間之相關性(Erkan & Radev, 2004; Mihalcea & Tarau, 2004; Wan & Yang, 2008)、或語句與文件在潛藏主題空間中的距離關係(Lin & Chen, 2009)等。

(3) 以監督式機器學習為基礎之自動摘要模型技術：監督式機器學習通常將自動摘要之任務視為二元分類問題(Binary Classification)，亦即將語句區分為摘要語句或非摘要語句。我們必須事先準備好一些訓練文件以及其對應的人工標註摘要資訊，然後透過各種分類器的學習機制，進行分類模型的訓練。對於尚未被摘要之文件，此類方法將文件裡的每個語句進行二元分類，即可依其結果產生出摘要。此類方法較著名的相關研究包括簡單貝氏分類器(Naïve-Bayes Classifier)(Kupiec, Pedersen & Chen, 1995)、高斯混合模型(Gaussian Mixture Model, GMM)(Murray, Renals & Carletta, 2005)、隱藏式馬可夫模型(Hidden Markov Model, HMM)(Conroy & O'Leary, 2001)、支持向量機(Support Vector Machines, SVM)(Kolcz, Prabakarmurthi & Kalita, 2001; Zhang & Fung, 2007)與條件隨機場域(Conditional Random Fields, CRF)(Shen, Sun, Li, Yang & Chen, 2007)等。監督式模型可同時結合多種摘要特徵來表示每一語句(通常是由上述以詞彙或結構為基礎之摘要方法、或是各式非監督式摘要模型針對語句所輸出的分數或機率值)，綜合各種摘要特徵所形成的特徵向量將被用來做為監督式摘要模型判斷語句是否屬於摘要語句的依據(Lin & Chen, 2009)。

此外，文字文件所要強調的是怎麼說(What-is-said)，而語音文件擁有許多純文字文件所沒有的資訊，通常除了怎麼說，更強調的是如何說(How-is-said)(Penn & Zhu, 2008)，明顯地，語音是多媒體內涵中最具資訊的成分之一，也因此語音文件摘要的相關研究通常從多媒體語音訊號中萃取豐富的韻律資訊(Prosodic Information)來判斷語句的重要性，

如：音調(Intonation)、音高(Pitch)、音強(Power)、語者發聲持續時間(Duration)、語者說話速率(Rate)、語者(Speaker)、情感(Emotion)和說話時場景(Environment)等資訊，這些都是從事語音文件摘要時可以善加利用的語句特徵資訊(Liu & Hakkani-Tur, 2011)。

3. 使用語言模型於語音文件摘要 (Language Modeling for Spoken Document Summarization)

語言模型的研究與發展最早是源自於語音辨識及自然語言處理。語言模型旨在描述語言中的所有詞彙之間共同出現與相鄰資訊的關係。其假設人類語言生成(Human Language Generation)是一個隨機過程，而語言模型就是在模擬如何由詞彙構成片語、語句、段落或者文件之過程的機率模型，故又稱為生成式語言模型(Generative Language Modeling)(Zhai, 2008)。最簡單的語言模型為單連語言模型(Unigram Language Model, ULM)，它不考慮詞彙之間的順序關係，只個別考慮每一個詞本身出現的機率。較為複雜且常被使用的語言模型為 *N*-連語言模型，通常 N 為 2 或 3（即二連或三連語言模型），其考慮兩個詞彙或三個詞彙之間的共同出現與緊連的順序關係。值得一提的是，單連語言模型和 *N*-連語言模型的主要優點之一是：它們僅需使用訓練語料來估測每一個詞本身出現的機率分佈，或者詞彙之間的共同出現與鄰近關係的機率分佈，並不需要額外的人工標記資訊，因此語言模型是屬於基於非監督式機器學習之模型技術。

在過去幾年中，語言模型在資訊檢索任務中已被廣泛地應用且有不錯的實務成效(Zhai, 2008)；但就我們所知，在語音文件摘要的任務上，關於使用語言模型的研究是相對較少的。本論文將藉由語言模型的使用來進行摘要語句選取，其基本方法為使用語句語言模型生成文件的文件相似度量值(Document Likelihood Measure, DLM)(Chen *et al.*, 2009)。此外，本章第 2 小節我們將闡述如何使用基於關聯性資訊來改進語句模型之估測，使其得以更精準的代表語句的語意內容。

3.1 文件相似度量值 (Document Likelihood Measure, DLM)

我們可以把語音文件摘要任務視為是資訊檢索的問題。一般來說，資訊檢索(Information Retrieval, IR)旨在尋找相關文件(Relevant Document)來回應使用者所送出的查詢(Query)或資訊需求(Information Need)。同樣地，在從事語音文件摘要時，我們可將每一篇被摘要文件視為是查詢，而文件中的語句(Sentence)視為候選資訊單元(Candidate Information Unit)；據此，我們可以假設在被摘要文件中，與其愈相關的語句愈有可能是可用來代表文件主旨或主題之摘要語句。

當給予一篇被摘要文件 D 時，文件中每一語句 S 的事後機率 $P(S|D)$ 可以用來表示語句 S 對於文件 D 的重要性。當使用語言模型來計算 $P(S|D)$ 時，我們透過貝氏定理(Bayes' Theorem)將 $P(S|D)$ 展開成()：

$$P(S \mid D) = \frac{P(D \mid S)P(S)}{P(D)} \tag{1}$$

其中 $P(D)$ 是文件 D 的事前機率，由於 $P(D)$ 不影響語句的排序結果，故可省略不討論；另一方面，$P(S)$ 是語句 S 的事前機率，可以使用各式非監督式方法或監督式方法來求得(Chen *et al*., 2009)。本論文的研究假設語句的事前機率為一個均勻分布(Uniform Distribution)，所以 $P(S)$ 亦可省略。最後，$P(D|S)$ 是語句 S 所形成的語言模型生成文件 D 之機率(或稱作文件相似度)，可以用來表示文件 D 與語句 S 之間的相似關係，如果語句 S 生成文件 D 的機率值愈高，代表語句 S 與文件 D 愈為相似(語句愈能代表文件 D)，即愈有可能是摘要語句。我們可以更進一步地假設文件 D 中詞與詞之間是獨立的，並且不考慮每一個詞在文件 D 中發生的順序關係(即詞袋假設(Bag-of-Word Assumption))，則語句 S 生成文件 D 的文件相似度量值(Document Likelihood Measure, DLM) $P(D|S)$ 可拆解成文件 D 中每一的詞 w 個別發生的條件機率之連乘積：

$$P(D|S)=\prod_{w\in D}P(w|S)^{C(w,D)} \tag{2}$$

此種方法是為語句 S 建立一個語句模型(Sentence Model) $P(w|S)$， w 是出現在文件 D 中的詞，$C(w,D)$ 是詞 w 出現在文件 D 中的次數。其中，我們可利用最大化相似度估測(Maximum Likelihood Estimation, MLE)的方式來建立每一個語句的語句模型：

$$P(w|S)=\frac{C(w,S)}{|S|} \tag{3}$$

在(3)中，$C(w,S)$ 表示詞 w 在語句 S 中出現的次數，$|S|$ 則表示語句 S 的總詞數。值得注意的是，由於語句 S 通常僅由少數字詞所組成，因此容易遭遇資料稀疏(Data Sparseness)的問題，這會使得語句模型使用最大化相似度估測時，不僅可能無法準確地估測每一個詞在語句中真正的機率分佈，也可能因為某些詞的條件機率值為零，導致語句 S 產生文件 D 的機率值為零。為了減輕上述的現象，可採用平滑化(Smoothing)技術來達成，常見的平滑化技術包含有 Jelinek-Mercer 平滑化、Dirichlet 平滑化、Add-delta 平滑化 (Zhai & Lafferty, 2001b)，本論文使用 Jelinek-Mercer 平滑化技術藉由使用以大量文字語料訓練而成的背景單連語言模型(Background Unigram Language Model)來調適語句模型(Zhai & Lafferty, 2001b)，故 $P(D|S)$ 可進一步地表示成：

$$P(D|S)=\prod_{w\in D}[\lambda\cdot P(w|S)+(1-\lambda)\cdot P(w|B)]^{C(w,D)} \tag{4}$$

其中，$P(w|B)$ 是詞 w 在背景單連語言模型 B 中之機率值。

3.2 虛擬相關回饋 (Pseudo-Relevance Feedback)

通常，文件中的語句僅由少許的詞彙所組成，當語句模型使用最大化相似度估測時，容易遭遇資料稀疏的問題；再者，由這語句 S 中些許的表面詞彙是遠不夠正確估算語句 S 與被摘要文件 D 之間的相似度(或低估了此相似度)，所以藉由背景語言模型進行語句模型之調適為最常見的方法之一(參照式(4))。

為了有效解決語句的資料稀疏及相似度被低估的問題，我們可利用在資訊檢索(Information Retrieval)領域被廣泛應用的虛擬相關回饋(Pseudo Relevant Feedback, PRF)技術來強化語句模型(重新估測或對其做調適)(Chen, Chen, Chen. Wang & Yu, 2014)。為此目的，當虛擬相關回饋運用於文件摘要領域中時，會將每一語句 *S* 當成是一個查詢(Query)，然後輸入到一個資訊檢索系統中，找出一些與語句 最可能相關的文件，而這些文件就稱之為虛擬相關文件(Pseudo Relevant Documents)；一個最簡單的方式即是選取排名最前面(檢索分數最高)的幾篇文件(Top-ranked Documents)。有了這些虛擬相關文件後，就可以利用它們來增進語句模型以解決語句資料稀疏及其相似度低估之問題，其虛擬關聯回饋示意圖如圖 1 所示。所以本論文針對語句模型調適進行初步研究，當我們透過資訊檢索系統已取得虛擬相關文件(最高排序文件)，接著就要做語句模型的調適估測，底下介紹常見的調適模型包含有關聯模型 (Relevance Model, RM)、簡單混合模型(Simple Mixture Model, SMM)以及三混合模型(Tri-Mixture Model, TriMM)。

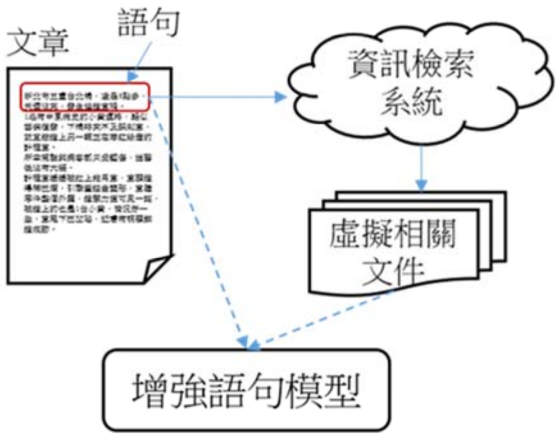


圖 1、虛擬關聯回饋示意圖
[Figure 1. Illustration of pseudo-relevance feedback.]

3.2.1 關聯模型 (Relevance Model, RM)

關聯模型的基本假設是認為每一語句 *S* 皆是被用來描述一個概念、想法或主題，我們稱之為語句的關聯類別(Relevance Class)。在本論文中，我們的目標是想進一步地模型化關聯類別所代表的資訊，藉此來豐富語句模型所能傳達的語意內容或主題特性。然而，實際上每一語句 的關聯類別 是非常難以求得的；為此，我們透過虛擬相關回饋(Pseudo Relevant Feedback, PRF)來尋找與關聯類別可能相關的一些文件，並藉由這些文件來近似關聯類別。更明確地，在實作上我們將虛擬相關文件(最高排序文件)$\mathbf{D}_{\text{Top}}=\{D_1,D_2,\ldots,D_M\}$用以代表關聯類別 。接著，透過檢視詞彙 *w* 與語句 *S* 在這些虛擬相關文件中同時出現之關係，可計算出詞彙與語句的聯合機率(Lavrenko & Croft, 2001)：

$$P_{\mathbf{RM}}(w,S)=\sum\nolimits_{D_m\in\mathbf{D_{Top}}}P(w,S\,|\,D_m)P(D_m), \tag{5}$$

當我們進一步地假設在給定某一篇虛擬相關文件時，詞彙與語句是獨立的，並且語句內的詞彙也是獨立且不考慮其先後次序(即所謂的詞袋假設)，則透過虛擬相關回饋所估測的語句模型為：

$$P_{\mathbf{RM}}(w\,|\,S)=\frac{\sum\nolimits_{D_m\in\mathbf{D_{Top}}}\prod\nolimits_{w'\in S}P(w'\,|\,D_m)P(w\,|\,D_m)P(D_m)}{\sum\nolimits_{D_{m'}\in\mathbf{D_{Top}}}\prod\nolimits_{w''\in S}P(w''\,|\,D_{m'})P(D_{m'})}, \tag{6}$$

我們稱之為關聯模型(Relevance Model, RM)。關聯模型的優點在於藉由虛擬相關文件的資訊，可以更清楚地知道語句所蘊含的資訊、所欲表達的內涵，所以相較於傳統使用最大化相似度估測的語句模型，可更準確地表達語句的語意內容或主題特性，以提升摘要的成效。

3.2.2 簡單混合模型 (Simple Mixture Model, SMM)

簡單混合模型的基本想法是假設由虛擬相關回饋技術所得到的虛擬相關文件是相關的且能從最高排序文件中估測比較好的簡單混合模型 $P_{\mathrm{SMM}}(w|S)$，更明確地說，簡單混合模型是假設虛擬相關文件 $\mathbf{D}_{\mathrm{Top}}$ 裡的詞彙 w 是源自於二種成分混合模型(Two-Component Mixture Model)，其一為簡單混合模型 $P_{\mathrm{SMM}}(w|S)$，另一為背景語言模型 $P(w|BG)$。簡單混合模型的估測是藉由期望值最大化(Expectation Maximization, EM)演算法來最大化虛擬相關文件的對數相似度(Log-Likelihood)以進行模型的估測，其虛擬相關文件的對數相似度的定義如下(Zhai & Lafferty, 2001a)：

$$\begin{aligned}LL_{\mathbf{D_{Top}}} = &\sum\nolimits_{D_m\in\mathbf{D_{Top}}}\sum\nolimits_{w\in V}\ c(w,D_m)\\ &\cdot\log[(1-\alpha)\cdot P_{\mathbf{SMM}}(w\,|\,S)+\alpha\cdot P(w\,|\,BG)],\end{aligned} \tag{7}$$

其中 α 為平衡參數，用來控制模型估測時是要比較偏好簡單混合模型或是背景語言模型，$c(w,D_m)$為詞彙 w 在虛擬相關文件 D_m 的次數，式(7)的最大化可透過期望值最大化迭代更新式來達成：

期望值步驟：

$$\tau_w^{(l)}=\frac{\alpha\cdot P_{\mathrm{SMM}}^{(l)}(w\,|\,S)}{\alpha\cdot P_{\mathrm{SMM}}^{(l)}(w\,|\,S)+(1-\alpha)\cdot P(w\,|\,BG)}, \tag{8}$$

最大化步驟：

$$P_{\mathbf{SMM}}^{(l+1)}(w|S)=\frac{\sum_{D_m\in\mathbf{D_{Top}}}c(w,D_m)\cdot\tau_w^{(l)}}{\sum_{w'\in V}\sum_{D'_m\in\mathbf{D_{Top}}}c(w',D'_m)\cdot\tau_w^{(l)}},\tag{9}$$

其中 l 表示期望值最大化的第 l 次迭代。這個簡單混合模型的估測會加強具有獨特性(Specificity)的詞彙之機率，例如某詞彙沒有在背景語言模型中有好解釋(Well-Explained)則會被加強其機率，這樣使得此模型為更具有鑑別(Discriminant)能力的語句模型；反之，若是沒有獨特性的詞彙，則其機率就會被背景語言模型所吸收。

3.2.3 三混合模型 (Tri-Mixture Model)

另一方面，本論文嘗試將三混合模型(Tri-Mixture Model)(Hiemstra, Robertson & Zaragoza, 2004)用於語音摘要任務。三混合模型可視為是複雜化後的簡單混合模型；它更進一步的假設虛擬相關文件 $\mathbf{D}_{\text{Top}}$ 裡的詞彙 w 是源自於三種成分模型(Component Models)，其一為文件模型 $P(w|D_m)$，其二為三混合模型 $P_{\text{TriMM}}(w|S)$，最後為背景語言模型 $P(w|BG)$。三混合模型的估測也是藉由期望值最大化演算法來最大化虛擬相關文件的對數相似度以進行模型的估測，其虛擬相關文件的對數相似度的定義如下(Hiemstra *et al*., 2004)：

$$\begin{aligned}LL_{\mathbf{D_{Top}}} =& \sum_{D_m\in\mathbf{D_{Top}}}\sum_{w\in V}c(w,D_m)\\ &\cdot\log[(1-\lambda-\mu)\cdot P_{\mathbf{TriMM}}(w|S)+\lambda\cdot P(w|D_m)+\mu\cdot P(w|BG)],\end{aligned}\tag{10}$$

其中 λ 和 μ 為平衡參數，用來控制模型估測時是要比較偏好三混合模型或文件模型亦或是背景語言模型，$c(w,D_m)$為詞彙 w 在虛擬相關文件 D_m 的次數，式(10)的最大化可透過期望值最大化迭代更新式來達成：

期望值步驟：

$$\begin{cases}r_{w,D_m}=\dfrac{c(w,D_m)\cdot(1-\lambda-\mu)\cdot P_{\text{TriMM}}(w|S)}{(1-\lambda-\mu)\cdot P_{\text{TriMM}}(w|S)+\mu\cdot P(w|BG)+\lambda\cdot P(w|D_m)}\\ e_{w,D_m}=\dfrac{c(w,D_r)\cdot\lambda\cdot P(w|D_m)}{(1-\lambda-\mu)\cdot P_{\text{TriMM}}(w|S)+\mu\cdot P(w|BG)+\lambda\cdot P(w|D_m)}\end{cases},\tag{11}$$

最大化步驟：

$$\begin{cases}\hat{P}_{\mathbf{TriMM}}(w|S)=\displaystyle\sum_{D_m\in\mathbf{D_{Top}}}\frac{r_{w,D_m}}{\sum_w r_{w,D_m}}\\ \hat{P}(w|D_m)=\dfrac{e_{w,D_m}}{\sum_w e_{w,D_m}}\end{cases},\tag{12}$$

運用此三混合模型來調適語句模型時，可取代原本的語句模型或與之線性結合(linearly interpolation)：

$$\hat{P}(w \mid S) = \gamma \cdot P(w \mid S) + (1-\gamma) \cdot P_{\mathbf{TriMM}}(w \mid S), \tag{13}$$

其中 $0 \le \gamma < 1$，當 $\gamma = 0$ 代表使用三混合模型取代原本的語句模型。

關聯模型、簡單混合模型及三混合模型在資訊檢索領域中已被廣泛應用(Zhai & Lafferty, 2001a; Lavrenko & Croft, 2001; Hiemstra *et al*., 2004)，但在摘要任務中卻是相對較少研究的，值得一提的是，雖然關聯模型、簡單混合模型已初步被應用在摘要任務上(Chen, Chang & Chen, 2013; Liu *et al*., 2014)，但三混合模型卻是本論文首次引入到(語音)文字摘要任務中。

4. 機率式排序模型 (Probabilistic Ranking Model)

在資訊檢索領域(Information Retrieval, IR)中，主要的概念就是設計一個排序模型並利用此模型來將文件做排序。同樣地，我們將節錄式語音文件摘要視為設計一個排序模型，用來排序一篇文件中的每一語句之問題，因此便可應用一些已在資訊檢索領域中發展良好的排序模型於語音摘要任務中，其中最著名的機率式模型即為最佳匹配(Best Matching, BM25)排序模型，我們將陸續介紹最佳匹配排序模型及其延伸。

4.1 BM25

在各式的排序系統中，有學者由機率模型的角度出發，發展出一套簡單且有效地排序計算公式，稱之為 BM25 (Jones, Walker & Robertson, 2000; Robertson & Zaragoza, 2008)：

$$BM25(S, D, B) = \sum_{w \in S} F(w, D) \cdot Sim(w, S) \cdot IDF(w, B) \tag{14}$$

$$F(w, D) = \frac{c(w, D)(k_2 + 1)}{c(w, D) + k_2} \tag{15}$$

$$Sim(w, S) = \frac{c(w, S)(k_1 + 1)}{c(w, S) + k_1(1 - b + b\frac{|S|}{avgsl})} \tag{16}$$

$$IDF(w, B) = \log\frac{B - n(w) + 0.5}{n(w) + 0.5} \tag{17}$$

其中，$c(w,B)$為 w 在文件 D 中的出現次數，B 為背景資訊中所有文件數目，$n(w)$為 w 在背景資訊中出現的文件數目，$|S|$為語句長度，$avgsl$ 為文件 D 中語句的平均長度，k_1、k_2 和 b 為可調的模型參數。

BM25是一個融合語句的詞頻資訊、文件相似度以及反文件頻函數之排序計算公式。在 BM25 的計算公式中，字詞出現在文件 D 的頻率資訊會經由權重函數 $F(w,D)$進行適當

的調整：當參數 k_2 設定為 0 時，則表示 BM25 僅考慮字詞是否有出現於文件當中，而不考慮其出現的頻率，若參數 k_2 的設定不為 0，BM25 將不僅考慮字詞的出現與否，並且進一步地將字詞於文件中出現的頻率資訊做適當的加權；文件相似度 $Sim(w,S)$則用於計算候選文件中與查詢共同出現的詞彙於文件中的重要性，查詢的詞彙在候選文件中亦扮演舉足輕重的角色，若查詢的詞彙共同出現較多次且參數 k_1 的設定不為 0，則表示此篇候選文件應被賦予較高的排序分數；反文件頻函數 $IDF(w,B)$是用於決定每一個詞彙的重要性，也就是加強內容字詞(Content word)的權重，並削弱功能字詞(Function word)的貢獻度。

近年來，有學者將 BM25 運用於意見摘要(Opinion Summarization)研究中，為了符合意見摘要所偏好的語句特性，他們進一步地將 BM25 修改為(Kim, Castellanos, Hsu, Zhai, Dayal & Ghosh, 2013)：

$$BM25_E(S,D,B)=\sum_{w\in S}Sim_E(w,S)\cdot IDF_E(w,B) \tag{18}$$

$$Sim_E(w,S)=\frac{c(w,S)(k_1+1)}{c(w,S)+k_1(1-b+b\frac{|S|}{avgsl})} \tag{19}$$

$$IDF_E(w,B)=\log\frac{|B|-c(w,B)+0.5}{c(w,B)+0.5} \tag{20}$$

其中，$c(w,B)$為 w 在背景資訊 B 中的出現次數，$|B|$為背景資訊所有字詞的次數。比較式(14)與(18)，$BM25_E$ 在對語句進行排序時，省略了考慮查詢詞彙出現頻率的資訊，僅考慮詞彙是否有出現於查詢中；另外，其 $IDF_E(w,B)$的算法是使用字詞 w 在背景資訊 B 中出現的次數，而不是使用字詞 w 在背景資訊 B 中出現的文件數目。

4.2 BM25L and BM25+

當語句很長的時候，文件相似度 $Sim(w,S)$在傳統的 BM25 排序公式(參照式(16))中會變得很小，意即傳統的 BM25 計算公式容易偏好短語句。有鑑於此，有學者提出一個解決方法來平衡語句長度的影響。為了方便解釋此方法，我們將式(16)重新改寫如下：

$$Sim(w,S)=\frac{c'(w,S)(k_1+1)}{c'(w,S)+k_1} \tag{21}$$

其中 $c'(w,S)$為

$$c'(w,S)=\frac{c(w,S)}{1-b+b\frac{|S|}{avgsl}} \tag{22}$$

當重新改寫為式(21)後，有學者提出使用新的文件相似度 *Sim'*(*w*,*S*)，其定義如下：

$$Sim'(w,S)=\begin{cases}\dfrac{(c'(w,S)+\delta)(k_1+1)}{(c'(w,S)+\delta)+k_1} & if\ c'(w,S)>0\\ 0 & c'(w,S)=0\end{cases} \tag{23}$$

其中 δ 為一定值，則新的排序公式為(BM25L)：

$$BM25L(S,D,B)=\sum_{w\in S}F(w,D)\cdot Sim'(w,S)\cdot IDF(w,B) \tag{24}$$

此新的文件相似度不僅保留原有BM25的兩點良好特性，(即當 $c'(w,S)=0$ 時 $Sim'(w,S)=0$；另外，$c'(w,S)$與 *BM25L* 皆呈單調遞增，並且 *BM25L* 有漸進最大值(Asymptotic maximal))，同時也因此有了一個正下界(positive lower bound)的特性(即對於 $c'(w,S)>0$，至少都會有 $(k_1+1)\delta/(k_1+\delta)$)，此特性可以平衡語句長度之影響，不會因為語句過長而影響變大且不會特別容易偏好短語句。

一方面，Lv & Zhai (2011a)發現不只原始 BM25 排序公式會過度懲罰長語句，就連其他的排序公式都會有一樣的情形，因此他們更進一步地提出一般化的方法來解決此問題，也就是要保證只出現一次的詞彙在長語句中至少會有一定的貢獻度，為了達到此目的，他們就在原始 BM25 公式中的文件相似度 *Sim*(*w*,*S*)裡加入一個常數值，且反文件頻函數 *IDF*(*w*,*B*)也有小修改，則新的排序公式為(BM25+，(Lv & Zhai, 2011b))：

$$BM25+(S,D,B)=\sum_{w\in S}F(w,D)\cdot Sim^{+}(w,S)\cdot IDF^{+}(w,B) \tag{25}$$

$$Sim^{+}(w,S)=\frac{c(w,S)(k_1+1)}{c(w,S)+k_1(1-b+b\dfrac{|S|}{avgsl})}+\delta \tag{26}$$

$$IDF^{+}(w,B)=\frac{|B|+1}{n(w)} \tag{27}$$

其中 δ 為一個固定值。

4.3 BM25T

在 4.1 小節所介紹的 BM25 排序公式中有三個需要設定的參數(k_1, k_2, b)，且所有的詞彙共享同一組設定，但其實每個詞彙應該要根據不同的重要性而設計不同的參數值。由於文件相似度 *Sim*(*w*,*S*) 是 BM25 公式中最重要的排序因子，所以參數 k_1 的設計就更顯重要。Lv & Zhai (2012) 認為經長度正規化的詞頻貢獻度應該要與有較高的長度正規化詞頻的文章數成正比，因此他們使用對數邏輯 (Log-logistic)方法來計算每個詞彙所對應不同的

參數 k_1，首先定義一個菁英集(Elite set)C_w，意即所有包含詞彙 w 的語句集合，則詞彙 w 的參數 k_1'之定義如下：

$$k_1' = \arg\min_{k_1} \left(g_{k_1} - \frac{\sum_{S' \in C_w} \left(\log(c'(w,S')+1) \right)}{n(w)} \right)^2 \tag{28}$$

$$g_{k_1} = \begin{cases} \frac{k_1}{k_1 - 1} \log(k_1) & if \ k_1 \neq 1 \\ 1 & k_1 = 1 \end{cases} \tag{29}$$

其中 $c'(w,S')$與式(22)相同，我們將 k_1 的範圍設定在 0.1 到 10 之間(每次增加 0.1)，透過式(28)我們可找到每一個詞彙 w 的最佳參數 k_1'，將式(28)所求得的參數帶回原始的 BM25 排序公式，便可得到新的排序公式(BM25T，(Lv & Zhai, 2012))：

$$BM25T(S,D,B) = \sum_{w \in S} F(w,D) \cdot Sim^T(w,S) \cdot IDF(w,B) \tag{30}$$

$$Sim^T(w,S) = \frac{c(w,S)(k_1'+1)}{c(w,S) + k_1'(1 - b + b\frac{|S|}{avgsl})} \tag{31}$$

表 1. 實驗語料統計資訊
[Table 1. The statistics of the dataset.]

	訓練集	測試集
語料時間	2001/11/07-2002/01/22	2002/01/23-2002/08/22
文件個數	185	20
文件平均持續幾秒	129.4	141.2
文件平均詞個數	326.0	290.3
文件平均語句個數	20.0	23.3
文件平均字錯誤率 (Character Error Rate, CER)	28.8%	29.8%
文件平均詞錯誤率 (Word Error Rate, WER)	38.0%	39.4%

5. 實驗語料及評估方法 (Dataset and Evaluation Method)

5.1 實驗語料 (Dataset)

本論文實驗語料庫為公視新聞語料(Mandarin Chinese Broadcast News Corpus, MATBN)(Wang, Chen, Kuo & Cheng, 2005)，是由中央研究院資訊科學研究所耗時三年與公共電視台合作錄製並整理的中文新聞語料，其錄製內容為每天一個小時的公視晚間新聞深度報導。我們抽取其中由 2001 年 11 月到 2002 年 8 月總共 205 則新聞報導，區分成訓練集(共 185 則新聞)以及測試集(共 20 則新聞)兩部分，其詳細的統計資訊如表 1 所示。全部 205 則語音文件長度約為 7.5 小時，我們先做人工切音，切出真正含有講話內容的音訊段落，再經由語音辨識器自動產生出的語音辨識結果稱之為語音文件(Spoken Document, SD)，因此語音文件中只包含有語音辨識錯誤之雜訊；另一方面，我們將此 205 則語音文件藉由人工聽寫的方式，產生出沒有辨識錯誤的正確文字語料，我們稱之為文字文件(Text Document, TD)，每則文字文件再經由三位專家標記摘要語句，我們將此標記的人工摘要做為語音文件與文字文件的正確摘要答案。藉由比較語音文件和文字文件的摘要效能，我們可以觀察語音辨識錯誤對於各種摘要方法之影響。本研究的背景語言模型訓練語料取材自 2001 到 2002 年的中央社新聞文字語料(Central News Agency, CNA)，並且以 SRI 語言模型工具訓練出經平滑化的單連語言模型，我們假設此單連語言模型為明確度中的非相關資訊之來源。另外，本論文蒐集 2002 年中央通訊社的約十萬則同時期新聞文字文件做為建立關聯模型時的檢索標的(Chen *et al.*, 2013)，關於語句 S 的虛擬相關文件(最高排序文件)篇數為 15(也就是$|\mathbf{D}_{\mathbf{Top}}|$=15)。

5.2 評估方法 (Evaluation Method)

自動摘要的評估方法主要有兩種，一為主觀人為評估，另一為客觀自動評估；前者為請幾位測試人員來為系統所產生的摘要做評估，給分的範圍為 1-5 分，後者則是預先請幾位測試者依據事先定義好的摘要比例挑選出適合的摘要語句，系統所產生的摘要句子將與測試者所挑選出的句子計算召回率導向的要點評估(Recall-Oriented Understudy for Gisting Evaluation, ROUGE)(Lin, 2003)。由於主觀人為評估非常耗時耗力，所以目前多數自動摘要方法皆採用召回率導向的要點評估做為文件摘要的評估方式，本論文亦採用此種評估方式。ROUGE 方法是計算自動摘要結果與人工摘要之間的重疊單位元(Units)數目占參考摘要(Reference Summary)長度(單位元總個數)的比例。估計的單位可以是 *N*-連詞(*N*-gram)、詞序列(Word Sequences)，如：最長相同詞序列或詞成對(Word Pairs)。由於此方法是採用單位元比對的方式，不會產生語句邊界定義的問題，並且適合於多份人工摘要的評估。其評估的分數有三種，ROUGE-1 (單連詞 Unigram)、ROUGE-2 (雙連詞 Bigram)和 ROUGE-L (最長共同片段 Longest Common Subsequence) 分數，ROUGE-1 是評估自動摘要的訊息量，ROUGE-2 是評估自動摘要的流暢性，ROUGE-L 是最長共同字串，本論文希望觀察摘要的流暢性，因此，實驗數據主要是以 ROUGE-2 分數為主。本論文所設定的摘要比例為 10%，其定義為摘要所含詞彙數占整篇文件詞彙數的比例，也就是以詞

彙做為判斷摘要比例的單元。在挑選摘要語句過程中，若選到某語句中的某個詞彙時就已經剛好達到摘要比例，為了保持語句語意完整性，此語句剩下的詞彙也會被挑選成為摘要。

6. 實驗結果 (Experimental Results)

6.1 基礎實驗結果 (Baseline Experiments)

首先，我們比較文件相似度量值(DLM)與數個非監督式摘要方法之摘要成效，包含有最長語句摘要(Longest Sentence, LS)、首句摘要(LEAD)(Penn & Zhu, 2008)、向量空間模型(Vector Space Model, VSM)(Gong &. Liu, 2001)、潛藏語意分析(Latent Semantic Analysis, LSA)(Gong &. Liu, 2001)、最大邊際關聯(Maximal Marginal Relevance, MMR)(Carbonell & Goldstein, 1998)、馬可夫隨機漫步(Markov Random Walk, MRW)(Wan & Yang, 2008)、次模(Submodularity)(Lin & Bilmes, 2010)以及整數線性規劃(Integer Linear Programming, ILP)(McDonald, 2007)。一般來說，文件中長句可能蘊含有較豐富的主題資訊，因此依據文件中語句長度做排序後，依序選取最長語句做為摘要結果是一種簡單的摘要方法。除此之外，也有學者研究發現，文件常以開門見山法的方式來提點出主題，因此文件開頭的前幾個語句經常是具代表性的語句，首句摘要即是以此概念出發，選取前幾句語句來形成整個文件的摘要。最長語句摘要(LS)及首句摘要(LEAD)都僅適用在一部分具有特殊結構的文件上，因此它們的缺點就是有其侷限性。另外，向量空間模型是把文件和語句分別視為一個向量，並使用詞頻-反文件頻(TF-IDF)特徵來計算每一維度的權重值，文件與語句間的關聯性是藉由餘弦相似度量值來估測，當語句分數較高時，則越有機會成為此文件的摘要。潛藏語意分析是在向量空間的假設下更進一步地使用奇異值分解(Singular Value Decomposition, SVD)來找到可能的潛藏語意空間，使之能在考量潛藏語意的情況下進行文件與語句的關聯性量測。最大邊際關聯可視為是向量空間模型的一個延伸，在做語句排序時考量了冗餘性以達到更好的摘要結果。馬可夫隨機漫步(MRW)的概念是把文件中的語句視為一個網際網路，文件中的語句代表網路中的節點，節邊權重值是兩個節點之間的語彙相似度，通常是透過節點的內分支度(Indegree)與外分支度(Outdegree)並採用餘弦(Cosine)估測法求得，所以馬可夫隨機漫步主要是依賴較一般化的資訊，例如：有概念性的網際網路，而不是考慮區域性的特徵(例如：每個語句)，因此如果有一個語句跟其他語句很相似的話，則可以代表摘要使之來描述文件中的主旨(Wan & Yang, 2008)。次模是一個貪婪(Greedy) 的語句選取方法，因其滿足次模的特性，意即每選取一語句就會有回報減少(Diminished Return)的效應，因此次模具有一個近似最佳解(Near-Optimal)(Lin & Bilmes, 2010)。整數線性規劃是一個全域(Global)的限制性最佳化(Constraint Optimization)的語句選取方法(McDonald, 2007)。

表 2 為本論文之基礎實驗結果。首先，在 TD 的實驗中，DLM 的摘要效果比 LS、LEAD、VSM、LSA、MMR 等非監督式摘要方法來得好些；因 LS 與 LEAD 僅適用於特殊文章結構上，所以若被摘要文件不具有某種特殊的文章結構，其摘要效能就會有限。

表 2. 基礎實驗結果
[Table 2. Baseline experiments.]

		F-score		
		ROUGE-1	ROUGE-2	ROUGE-L
TD	LS	0.225	0.098	0.183
	LEAD	0.310	0.194	0.276
	VSM	0.347	0.228	0.290
	LSA	0.362	0.233	0.316
	MMR	0.368	0.248	0.322
	MRW	0.412	0.282	0.358
	Submodularity	0.414	0.286	0.363
	ILP	0.442	0.337	0.401
	DLM	0.411	0.298	0.361
SD	LS	0.181	0.044	0.138
	LEAD	0.255	0.117	0.221
	VSM	0.342	0.189	0.287
	LSA	0.345	0.201	0.301
	MMR	0.366	0.215	0.315
	MRW	0.332	0.191	0.291
	Submodularity	0.332	0.204	0.303
	ILP	0.348	0.209	0.306
	DLM	0.364	0.210	0.307

相較之下，DLM 是較具一般性的摘要方法，因此比較不會受限於文章的結構之影響，故摘要效能比 LS 以及 LEAD 來得彰顯。DLM 與 VSM 皆使用淺層的詞彙(詞頻)資訊，但由於 DLM 是計算語句模型與文件模型之間的距離關係，對於代表語句與文件的語言模型，我們較容易透過各種技術來進行模型的估計與調適，進而獲得較好的摘要成果。整數線性規劃是一個全域選擇方法，所以在 TD 上可以得到最好的摘要效能。

另一方面，在 SD 的實驗中，DLM 同樣較優於 LS、LEAD、VSM、LSA 等之摘要方法，但 MMR 的結果則稍微較 DLM 好一點，我們認為這可能是因為 MMR 比較不受到語音辨認錯誤的影響。但 MRW 及次模也可能是受到語音辨識錯誤的影響而造成摘要效能減低，甚至比 DLM 來得差。出乎意料的是原以為 ILP 也會在 SD 中得到最好的摘要效

能，結果反而是 MMR 得到最好的摘要效能，可能的原因是 ILP 受到語音辨識錯誤的影響比較大，造成其摘要結果不彰。

通常語音文件主要會有語音辨識錯誤和語句邊界偵測錯誤的問題，但我們有先經人工切音，因此摒除了語句邊界偵測錯誤的問題，藉由比較 TD 與 SD 之實驗結果，我們可以觀察語音辨識錯誤率對摘要結果的影響性。比較各式方法，SD 比 TD 下降了 1.9%~8.8%的 ROUGE-2 摘要效能，由此可知語音辨識錯誤率對摘要效能是有顯著的影響性。為了減緩語音辨認錯誤的問題，在未來我們將嘗試使用音節(Syllable)為單位來建立語句以及文件模型；或利用詞圖(Word Graph)、混淆網路(Confusion Network)來含括更多的可能正確候選詞彙以裨益模型估測；更可利用韻律資訊(Prosodic Information)等聲學線索來輔助減緩語音辨認錯誤對摘要效能的影響。

6.2 關聯模型之實驗結果 (Experiments of Relevance Model)

使用關聯模型於語句模型之建立時，需要做一次的資訊檢索來為每個語句找出虛擬相關文件，由同時期的新聞文字文件(共 101,268 篇)中為每一語句選取出 15 篇虛擬相關文件。由於文件中的語句通常相對簡短，因此當使用最大化相似度估測建立語句模型時，容易遭遇資料稀疏的問題，不容易獲得精準的模型，故我們期望考慮額外的關聯資訊於語音文件摘要，亦即藉由虛擬相關文件來重新估測並建立語句的語言模型，能獲得進一步地摘要成效。重新估測後的關聯模型則可與原本的語句模型相結合或取代之，相結合的參數調整在本實驗中是採用經驗設定(Empirical Setting)。實驗結果如表 3 所示，在 TD 與 SD 之摘要成效上，使用關聯模型(RM)、簡單混合模型(SMM)及三混合模型(TriMM)皆能比基礎的 DLM 實驗較好，尤其是三混合模型(TriMM)相較於 DLM 在 TD 及 SD 的 ROUGE-2 結果上能有 5.2%與 1.8%的改進。接著，我們比較不同關聯模型的摘要成效，首先是關聯模型(RM)與簡單混合模型(SMM)的比較，從表 3 的實驗結果得知關聯模型在 TD 上表現比簡單混合模型來得好，但在 SD 似乎在 ROUGE-1 就沒比簡單混合模型好，不過 SD 的 ROUGE-2 跟 ROUGE-L 都還是比簡單混合模型的效果好。關聯模型的假設是強調詞彙 w 與語句 S 在這些虛擬相關文件中同時出現之關係(參照式(5))來估測模型，而簡單混合模型是強調訓練好的模型能讓有獨特性的詞彙得到更多的機率值因而讓模型具有鑑別能力，兩者皆有其好處。最後，三混合模型(TriMM)因複雜化了簡單混合模型(SMM)，額外多考量文件模型的影響力，因此相較於關聯模型及簡單混合模型能得到更佳的摘要效能，三混合模型相較於關聯模型在 TD 上有明顯的進步，於 ROUGE-2 結果能有 1.4%的改進，但在 SD 上，於 ROUGE-2 結果只有微量的 0.2%改善。

在關聯模型的相關實驗中，語音辨識錯誤也是影響摘要效能非常嚴重，在三混合模型的數據中，SD 比 TD 劇烈下降了 12.2%的 ROUGE-2 摘要效能，在未來研究中，我們認為可以以次詞索引(Subword Indexing)的方式來建立關聯模型以減緩語音辨識錯誤之影響。

表 3、關聯模型之實驗結果
[Table 3. Experimental results of different relevance models.]

		F-score		
		ROUGE-1	ROUGE-2	ROUGE-L
TD	DLM	0.411	0.298	0.361
	RM	0.450	0.336	0.400
	SMM	0.436	0.325	0.385
	TriMM	**0.457**	**0.350**	**0.404**
SD	DLM	0.364	0.210	0.307
	RM	0.374	0.226	0.321
	SMM	0.375	0.221	0.314
	TriMM	**0.379**	**0.228**	**0.325**

6.3 平滑化技術於關聯模型之實驗結果 (Experiments of Smoothing Methods for Relevance Model)

語言模型在使用時會遇到資料稀疏的問題，通常的解決方法為替語言模型做平滑化(Smoothing)，我們將探討平滑化技術於語言模型在語音(文字)摘要結果上的影響，在本小節中我們以關聯模型(RM，參考 3.2.1 小節)為例[1]，採用三種不同的平滑化技術於關聯模型中，第一為 Jelinek-Mercer 平滑化，第二為 Dirichlet 平滑化，第三為 Add-delta 平滑化，茲分別如下(Zhai & Lafferty, 2001b)：(i) Jelinek-Mercer 平滑化為最簡單的與背景模型 $P(\mathrm{w}|B)$線性結合的平滑化技術，其公式為：

$$P_{JM}(w|S) = \lambda P(w|S) + (1-\lambda)P(w|B) \tag{32}$$

其中 λ 為線性結合參數，在實驗設定中是從 0.1 到 0.9(每次增加 0.1)。(ii) Dirichlet 平滑化主要是根源於貝式平滑(Bayesian Smoothing)而來的，它假設語言模型有個事前(Prior)機率，而此事前機率的分布剛好就是 Dirichlet 分布，因此 Dirichlet 平滑化公式可定義如下(Zhai & Lafferty, 2001b)：

$$P_{Dir}(w|S) = \frac{c(w|S) + \mu \cdot P(w|B)}{|S| + \mu} \tag{33}$$

[1] 我們實驗發現不同的平滑化技術都會對這三種不同的關聯模型有幫助，在摘要成效上也都有明顯的進步，且關聯模型(RM)會有最大的進步。

表 4. ***平滑化技術於關聯模型(RM)之實驗結果***
[Table 4. Experimental results of various smoothing methods for relevance model.]

Relevance Model (RM)		*F*-score		
		ROUGE-1	ROUGE-2	ROUGE-L
TD	Jelinek-Mercer	0.450	0.336	0.400
	Dirichlet	0.472	0.365	0.428
	Add-delta	0.493	0.386	0.441
SD	Jelinek-Mercer	0.374	0.226	0.321
	Dirichlet	0.401	0.254	0.349
	Add-delta	0.402	0.255	0.347

其中 μ 為 Dirichlet 參數，在實驗設定中的範圍為 1 到 100(每次增加 1)。(iii) Add-delta 平滑化是一個簡單平滑化技術，其原理就是加入一點點值，使沒有出現過的詞彙之機率不為零，其公式定義如下(Lv & Zhai, 2014)：

$$P_{Delta}(w\,|\,S)=\frac{c(w\,|\,S)+\delta}{|\,S\,|+\delta\cdot|\,V_F\,|} \tag{34}$$

其中 δ 為可調參數，在實驗設定範圍為 0.1 到 1(每次增加 0.1)，而$|V_F|$為虛擬相關回饋文件(在此為 15 篇)中不同詞彙的個數。三種平滑化技術於關聯模型(RM)的語音(文字)摘要結果如表 4 所示，無論在 TD 或 SD 的情況下，其中表現最佳為 Add-delta 平滑化，其次是 Dirichlet 平滑化，最差的是 Jelinek-Mercer 平滑化。Add-delta 平滑化表現比較好的原因是因為利用到相關回饋文件中不同詞彙的個數($|V_F|$)的資訊，使之能讓共同出現在語句與相關回饋文件中的詞彙 w 有比較高的機率(相較於沒有共同出現的詞彙)，因此在估測關聯模型時能更具有鑑別能力(區分出重要且共同出現在語句與相關回饋文件的詞彙與一般性且不重要的詞彙)，而讓摘要效能變得更好，尤其是 TD 的情況下相較於 Jelinek-Mercer 平滑化在 ROUGE-2 的絕對進步率有 5%之多，這是相當顯著的。但在 SD 的情況下，雖然 Add-delta 平滑化還是會比 Dirichlet 平滑化及 Jelinek-Mercer 平滑化來得好，但與 Dirichlet 平滑化相比，其摘要效能其實已經相差無幾，可能的原因之一還是因為語音辨識錯誤的影響所造成，在未來研究中，我們將以次詞索引(Subword Indexing)的方式來建立關聯模型以減緩此問題。

6.4 機率式排序模型之實驗結果 (Experiments of Probabilistic Ranking Model)

接著我們將焦點轉移到機率式排序模型上，在本小節中我們將比較各種不同的 BM25 排序模型，包含有原始 BM25(參照式(14))、$BM25_E$(參照式(18))、BM25L(參照式(24))、BM25+(參照式(25))及 BM25T(參照式(30))。其實驗結果如表 5 所示，在 TD 的部分，BM25 的摘要表現已經很不錯，甚至都比其他良好發展的非監督式摘要方法來得好(與表 2 之結果比較)，$BM25_E$ 因少了一個重要因子來做語句排序，所以可預期它的摘要效能比 BM25 來得差，但超乎預期的是在 ROUGE-2 將近有 16%的差距。BM25L 的提出是為了解決過度懲罰長語句的問題，在本實驗中的 TD 情況下，可看出 BM25L 的摘要效能會沒有比原始 BM25 來的好，其可能的原因是在資訊檢索領域中，確實會有很長文章(Long Document)的出現，懲罰長文章會有一定的效果，但在語音摘要任務中，很長語句(Long Sentence)的出現幾乎是不太可能，因此懲罰長語句就不一定會得到好的摘要效能。BM25+也是為了解決過度懲罰長語句的問題，但更一般化地保證只出現一次的詞彙至少要有個下界，因此 BM25+的摘要效能比 BM25L 能更進一步的提升，與原始 BM25 及 BM25L 相比較，在 ROUGE-2 上分別能有 0.2%及 1.1%的絕對進步。BM25T 從訓練資料中自動學習與詞彙相關的參數 k_1，在原始的文獻裡用於資訊檢索領域中的實驗是相對排序公式來得好(Lv & Zhai, 2012)，但 BM25T 的摘要效能有點出乎意料之外，沒有比 BM25L 與 BM25+好，甚至也會比原始 BM25 排序公式來的差，我們認為此種自動學習參數的方法可能是與資料相關的，或許可以替換另一套訓練資料集來重新學習詞彙相關的參數，但這也是未來的工作之一。另一方面，在 SD 的實驗部分，原始的 BM25 排序公式還是維持一定的水準，與其他非監督式摘要方法相比還是會比較好(參照表 2)，BM25L、BM25+及 BM25T 的優勢就沒那麼大，其摘要效能在 ROUGE-2 上比原始 BM25 都要來得差，確實語音辨識錯誤的影響還蠻大的，原本 BM25L、BM25+及 BM25T 的優點都被錯誤辨識的詞彙所消彌，甚至 BM25L、BM25+與 BM25T 的摘要結果幾乎相差無幾。

表 5. BM25 其及變形之相關實驗結果
[Table 5. Experimental results of BM25 and its variants.]

		F-score		
		ROUGE-1	ROUGE-2	ROUGE-L
TD	BM25	0.484	0.374	0.442
	$BM25_E$	0.352	0.210	0.294
	BM25L	0.480	0.365	0.434
	BM25+	0.486	0.376	0.444
	BM25T	0.463	0.352	0.419
SD	BM25	0.390	0.247	0.338
	$BM25_E$	0.279	0.151	0.250
	BM25L	0.384	0.246	0.337
	BM25+	0.383	0.242	0.335
	BM25T	0.382	0.238	0.332

7. 結論與未來展望 (Conclusions and Future work)

本論文主要有三個研究貢獻，其一為有鑑於關聯性(Relevance)的概念在資訊檢索領域中已有不錯的發展成果，本論文嘗試結合關聯性資訊來重新估測並建立語句的語言模型，並首次使用三混合(Tri-Mixture Model, TriMM)模型，使其得以更精準地代表語句的語意內容，期望可增進自動摘要之效能，實驗結果顯示三混合模型可以有最佳的摘要效能。其二為有鑑於語言模型著重依賴平滑化技術，本論文也是首次比較研究不同平滑化技術所估測得的語言模型對語音文件摘要任務之影響，根據實驗結果 Add-delta 平滑化可以達到最佳摘要效果，所以我們建議關聯模型的平滑化技術應當使用 Add-delta 平滑化來達成。最後為我們首次提出並應用多種機率式資訊檢索排序模型於語音摘要任務上，並且從實驗結果中得知與其他常見的非監督式摘要方法相比較能有不錯的摘要效能。

未來，我們的研究將有三個主要的方向：首先，多種機率式檢索排序模型還是需要經驗去調整不確定的參數，我們將進一步的研究是否可以針對不同的文件或不同的語句給予適當的權重調整，以期獲得更好的摘要成效；第二，目前關聯模型僅運用於重建語句的語言模型，我們將嘗試使用被摘要文件的關聯資訊來重新估測並建立文件的語言模型；最後，我們希望將非監督式摘要方法所產生的分數視為一種具代表性的摘要特徵資訊並結合於監督式機器學習方法(如條件隨機場域(Conditional Random Fields, CRFs)或深度類神經網絡(Deep Neural Network Learning, DNN)等)中，期望訓練後的模型能夠在文字文件摘要或語音文件摘要上獲得更好的表現。

Reference

Barzilay, R., & Elhadad, M. (1997). Using Lexical Chains for Text Summarization. In *Proceedings of ACL Workshop on Intelligent Scalable Text Summarization*, 10-17.

Baxendale, P. (1958). Machine-made Index for Technical Literature - an Experiment. *IBM Journal of Research and Development*, *2*(4), 354-361.

Cai, X.-Y., & Li, W.-J. (2013). Ranking through Clustering: An Integrated Approach to Multi-Document Summarization. *IEEE Transactions on Audio, Speech and Language Processing*, *21*(7), 1424-1433.

Carbonell, J., & Goldstein, J. (1998). The Use of MMR Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 335-336.

Chen, Y.-T., Chen, B., & Wang, H.-M. (2009). A Probabilistic Generative Framework for Extractive Broadcast News Speech Summarization. *IEEE Transactions on Audio, Speech and Language Processing*, *17*(1), 95-106.

Chen, B., Chang, H.-C., & Chen, K.-Y. (2013). Sentence Modeling for Extractive Speech Summarization. In *Proceedings of the International Conference on Multimedia & Expo (ICME)*. doi: 10.1109/ICME.2013.6607518

Chen, B., Chen, Y.-W., Chen, K.-Y., Wang, H.-M., & Yu, K.-T. (2014). Enhancing Query Formulation for Spoken Document Retrieval. *Journal of Information Science and Engineering*, *30*(3), 553-569.

Conroy, J.-M., & O'Leary, D.-P. (2001). Text Summarization via Hidden Markov Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 406-407. doi: 10.1145/383952.384042

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligent Research*, *22*(1), 457-479.

Galley, M., McKeown, K., Hirschberg, J., & Shriberg, E. (2004). Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 669-676. doi: 10.3115/1218955.1219040

Gong, Y., & Liu, X. (2001). Generic Text Summarization using Relevance Measure and Latent Semantic Analysis. In *Proceedings of the 24th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 19-25. doi: 10.1145/383952.383955

Hiemstra, D., Robertson, S., & Zaragoza, H. (2004). Parsimonious Language Models for Information Retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 178-185. doi: 10.1145/1008992.1009025

Jones, K. S., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, *36*(6), 779-808. doi: 10.1016/S0306-4573(00)00015-7

Kim, H.-D., Castellanos, M. G., Hsu, M., Zhai, C., Dayal, U., & Ghosh, R. (2013). Ranking Explanatory Sentences for Opinion Summarization. In *Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1069-1072. doi: 10.1145/2484028.2484172

Kolcz, A., Prabakarmurthi, V., & Kalita, J. (2001). Summarization as Feature Selection for Text Categorization. In *Proceedings of the tenth International Conference on Information and Knowledge Management*, 365-370. doi: 10.1145/502585.502647

Kuo, J.-J., & Chen, H.-H. (2006). Multi-document Summary Generation using Informative and Event Words. *Journal of ACM Transactions on Asian Language Information Processing, 7*(1), 550-557. doi: 10.1145/1330291.1330294

Kupiec, J., Pedersen, J., & Chen, F. (1995). A Trainable Document Summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 68-73. doi: 10.1145/215206.215333

Lavrenko, V., & Croft, W.-B. (2001). Relevance-based Language Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 120-127. doi: 10.1145/383952.383972

Lin, S.-H., & Chen, B. (2009). Improved Speech Summarization with Multiple-hypothesis Representations and Kullback-Leibler Divergence Measures. In *Proceeding of the 10th Annual Conference of the International Speech Communication Association (Interspeech)*, 1847-1850.

Lin, S.-H., & Chen, B. (2010). A Survey on Speech Summarization Techniques. *The Association for Computational Linguistics and Chinese Language Processing Newsletter*, *21*(1), 4-16.

Lin, H., & Bilmes, J. (2010). Multi-document Summarization via Budgeted Maximization of Submodular Functions. In *Proceeding of NAACL HLT*, 912-920.

Lin, S.-H., Yeh, Y.-M., & Chen, B. (2011). Leveraging Kullback-Leibler Divergence Measures and Information-Rich Cues for Speech Summarization. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(4), 871-882. doi: 10.1109/TASL.2010.2066268

Lin, C.-Y. (2003). *ROUGE: Recall-oriented Understudy for Gisting Evaluation*. [Online]. Retrieved from http://haydn.isi.edu/ROUGE/.

Liu, Y., & Hakkani-Tur, D. (2011). Speech Summarization. In G. Turand & R. D. Mori (Eds), *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. West Sussex, U.K.: Wiley. doi: 10.1002/9781119992691.ch13

Liu, S.-H., Chen, K.-Y., Hsieh, Y.-L., Chen, B., Wang, H.-M., Yen, H.-C., & Hsu, W.-L. (2014). Effective Pseudo-relevance Feedback for Language Modeling in Extractive Speech Summarization. In *Proceedings of the IEEE International Conference on*

Acoustics, Speech, and Signal Processing, 3226-3230. doi: 10.1109/ICASSP.2014.6854196

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, *2*(2), 159-165.

Lv, Y., & Zhai, C.-X. (2011a). When Documents Are Very Long, BM25 Fails! In *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1103-1104. doi: 10.1145/2009916.2010070

Lv, Y., & Zhai, C.-X. (2011b). Lower-bounding Term Frequency Normalization, In *Proceeding of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*, 7-16. doi: 10.1145/2063576.2063584

Lv, Y., & Zhai, C.-X. (2012). A Log-logistic Model-based Interpretation of TF Normalization of BM25. In *Proceedings of European Conference on Information Retrieval (ECIR)*, 244-255.

Lv, Y., & Zhai, C.-X. (2014). Revisiting the Divergence Minimization Feedback Model, In *Proceeding of the 23rd ACM International Conference on Information and Knowledge Management (CIKM)*, 1863-1866. doi: 10.1145/2661829.2661900

Mani, I., & Maybury, M.-T. (1999). *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press.

McDonald, R. (2007). A Study of Global Inference Algorithms in Multi-document Summarization, In *Proceedings of the 29th European Conference on Information Retrieval*, 557-564.

Mihalcea, R., & Tarau, P. (2004). TextRank Bringing Order into Texts. In *Proceedings of Empirical Method in Natural Language Processing (EMNLP 2004)*, 404-411.

Murray, G., Renals, S., & Carletta, J. (2005). Extractive Summarization of Meeting Recordings. In *Proceedings of the Conference of the International Speech Communication Association (Interspeech)*, 593-596.

Nenkova, A., & McKeown, K. (2011). Automatic Summarization. *Foundations and Trends in Information Retrieval*, *5*(2-3), 103-233. doi : 10.1561/1500000015

Ostendorf, M. (2008) Speech Technology and Information Access. *IEEE Signal Processing Magazine*, *25*(3), 150-152. doi: 10.1109/MSP.2008.918685

Paice, C.-D. (1990). Constructing Literature Abstracts by Computer Techniques and Prospects. *Journal of Information Processing and Management*, *26*(1), 171-186. doi: 10.1016/0306-4573(90)90014-S

Penn, G., & Zhu, X. (2008). A Critical Reassessment of Evaluation Baselines for Speech Summarization. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 470-478.

Robertson, S. & Zaragoza, H. (2008). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, *3*(4), 333-389. doi: 10.1561/1500000019

Shen, D., Sun, J.-T., Li, H., Yang, Q., & Chen, Z. (2007). Document Summarization using Conditional Random Fields. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2862-2867.

Strzalkowski, T., Wand, J., & Wise, B. (1998). A Robust Practical Text Summarization. In *Proceedings of AAAI Conference on Artificial Intelligence Spring Symposium on Intelligent Text Summarization*, 26-33.

Wan, X., & Yang, J. (2008). Multi-document Summarization using Cluster-based Link Analysis. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 299-306. doi: 10.1145/1390334.1390386

Wang, H.-M., Chen, B., Kuo, J.-w., & Cheng, S.-S. (2005). MATBN: A Mandarin Chinese broadcast news corpus. *International Journal of Computational Linguistics and Chinese Language Processing*, *10*(2), 219-236.

Witbrock, M., & Mittal, V. (1999). Ultra Summarization: a Statistical Approach to Generating Highly Condensed Non-extractive Summaries. In *Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 315-316. doi: 10.1145/312624.312748

Zhai, C.-X., & Lafferty, J. (2001a). Model-based feedback in the language modeling approach to information retrieval. In *Proceeding of the tenth International Conference on Information and Knowledge Management (CIKM)*, 403-410. doi: 10.1145/502585.502654

Zhai, C.-X., & Lafferty, J. (2001b). A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 334-342. doi: 10.1145/383952.384019

Zhai, C.-X. (2008). Statistical Language Models for Information Retrieval: A Critical Review. *Foundations and Trends in Information Retrieval*, *2*(3), 137-213. doi: 10.1561/1500000008

Zhang, J., & Fung, P. (2007). Speech Summarization without Lexical Features for Mandarin Broadcast News. In *Proceedings of NAACL HLT*, Companion Volume, 213-216.

Zhang, J.-J., Chan, H.-Y., & Fung, P. (2010). Extractive Speech Summarization using Shallow Rhetorical Structure Modeling. *IEEE Transactions on Audio, Speech and Language Processing*, *18*(6), 1147-1157. doi: 10.1109/TASL.2009.2030951

Computational Linguistics and Chinese Language Processing
Vol. 22, No. 1, June 2017, pp. 27-52

反義詞「多」和「少」在數量名結構中的不對稱現象——以語料庫為本的分析[1]

The Asymmetric Occurences of *Dou1* and *Shao3* in the [Numeral + Measure Word/Classifier + Noun] Construction: A Corpus-based Analysis

陳威佑*、鍾曉芳+

Wei-Yu Chen and Siaw-Fong Chung

摘要

反義詞組「多」和「少」在許多語法環境，尤以「數詞+量詞/分類詞+名詞(以下簡稱「數量名」)」構式中所呈現的不對稱現象最為明顯；僅採用「標記理論」無法有效解釋兩者的使用差異。本文假設現代漢語「多」具有「數詞」用法，相反地，「少」則無此用法，本文並使用「中研院現代漢語平衡語料庫」為工具，檢視「多」於「數量名」結構中，適合歸類於何種句法位置。本文將「多」歸為「數詞定詞(Neu)」，並認為處於數詞位置的「多」具有「計數」和「表量」兩種功能，且經歷和「很多」、「許多」不同的發展路徑：首先由「數詞」發展出表「約量」和「鄰近增量」的「數量定詞」功能，最後發展出作為動詞和形容詞「補語」的功能。

[1] The authors would like to thank the Taiwan Ministry of Science and Technology Project [104-2420-H-004-003-MY2] for supporting the research herein.

* 國立政治大學華語文教學學位學程碩士
Master's Program in TCSL, National Chengchi University
E-mail: lancelot2925@gmail.com

+ 國立政治大學英國語文學系
Department of English, National Chengchi University
E-mail: sfchung@nccu.edu.tw

Abstract

As two words with opposite meanings, *dou1* and *shao3* are expected to be similar and different in various environments. In this work, we looked into the construction of [numeral + measure word/classifier + noun] (hereafter [Num + MW/CL + N]) and explored in what ways both words present an asymmetric phenomenon. We found that *dou1* carries a numeral meaning while *shao3* lacks this use. Based on the analysis of these two words in the Sinica Corpus, this paper argued that *dou1* is better categorized as ‘Neu’ and suggested that *dou1* in [Num + MW/CL + N] serves two functions: one is for counting numbers; the other is for the expression of quantities. These findings can be related to the use of *dou1* as complement and as numeral.

關鍵詞：「多」和「少」、數量名結構、反義詞、數量概念

Keywords: *Dou1* and *Shao3* , [Num + MW/CL + N], Antonyms/opposites, Numeral Concept

1. 前言 (Introduction)

「多」和「少」為一組常見的反義詞，在相同的語法形式下，兩者具有「相對」的概念，用以表達意義上的相反，如：「我平常吃得很**多**/我平常吃得很**少**」，分別指稱吃食量的多和少。

然而，「多」和「少」在語言的使用中卻存在不對稱的現象。某些可用「多」的語言形式，無法直接使用「少」表達相對的反義概念，例如：「簡單**多**了/*簡單**少**了」。此搭配限制並非形容詞「簡單」所引起的，其反義詞「困難」，也受此限制：「困難**多**了/*困難**少**了」。按照語意上的邏輯，「簡單多了」等於「困難少了」、「困難多了」等於「簡單少了」，但是兩個形容詞，即便是反義概念，所選擇的搭配項目都傾向「多」而不與「少」搭配。

2. 反義形容詞與標記理論 (Opposite Adjectives and Case Theory)

關於「多」和「少」的研究，學者們以不同的角度出發，著重的面向略有不同。陸儉明(1985)、施一昕(1988)按照詞性的角度，詳細描寫「多」和「少」作為動詞和形容詞時的不對稱表現；歷時方面，方一新、曾丹(2007)、陳昌來、占云芬(2009)探討「多少」的語法化過程和多、少之間的關係；文獻上亦可見以「標記理論」解釋語意相反的形容詞組特性(沈家煊，1999；石毓智，2011)。本文認為「多」和「少」出現在「數量名」結構中的不對稱現象，僅以標記理論解釋具有侷限性。

下面先簡述反義形容詞的特性，並引用石毓智先生(2011)以「標記的有無」處理反義詞的方法，指出標記理論無法解釋「數量名」結構中「多」和「少」的不對稱現象。

2.1 語意相反的形容詞特性 (The Attributes of Opposite Adjectives)

湯廷池(1979)指出，華語中語義相反的形容詞，是可以出現在「既不……也不……」中空白處的一對形容詞，例如：「大、小；長、短；強、弱；多、少」等。通常擺在前面的形容詞表示正面或積極的意義，後面的形容詞則表示反面或消極的意義。這類的反義形容詞組合，其共同語義特點為「肯定其中一個必定否定另外一個，但否定其中一個並不一定肯定另外一個。」

(1a) 張三很高。

(1b) 張三不矮。(湯廷池，1979)

例句(1a)若成立，則例句(1b)必然成立；亦即肯定「高」，則必定否定「矮」。

(2a) 張三不高。

(2b) 張三很矮。(湯廷池，1979)

假設例句(2a)成立，則(2b)不必然成立；亦即否定「高」，並不一定就肯定「矮」。

表達中立意義時，則以正面積極意義代表，反面意義不能代表中立的意義，如下面例句：

(3) 你有多高？

(4) 水有多深？(湯廷池，1979)

例句(3)和(4)，皆以正面積極意義的形容詞作為中性意義的問句形式；例句(3)的答句，可能為：120 公分(實際身高很矮)、例句(4)的答句可能為：20 公分 (實際水深很淺)，但反過來的情況，以反面意義提問時，不包括中性意義。

(5) 你有多矮？

(6) 水有多淺？

例句(5)和(6)的答句，必須是符合提問者的預設條件，即被提問者的身高很矮、水必須很淺，否則形成答非所問的情形：

(5') 問：你有多矮？答：??我身高 180 公分。

(6') 問：水有多淺？答：??100 公尺。

上述的語言現象，可由「標記理論」獲得解釋。標記理論最初用於解釋語音的不對稱現象，現已擴及至語言研究的各個層面，用以解釋一個範疇內部存在的某種不對稱的現象。沈家煊(1999)指出，無標記的使用頻率高於有標記項。由此，可以解釋出現在「既不……也不……」中的形容詞，擺在前面表正面意義的詞，可兼用以表示中性意義，使用頻率較高，因此是無標記項，與之搭配的相對詞則是有標記項。沈家煊(1999)更進一步指出，否定是標記顛倒的一種手段，各種反義詞，無標記項傾向和無標記項組配，有標記項傾向和有標記項組配。石毓智(2011)則使用量級幅度的概念，具體分析出反義詞組有標記和無標記的差別。

2.2 標記理論 (Case Theory)

石毓智(2011)指出，有標記(marked)和無標記(unmarked)是指一對成分中是否帶有區別性特徵，這種區別性特徵可以將一個成分和另一個成分區分開來，有標記的成分有明顯的傾向性，無標記的成分則是一種沒有傾向性的客觀詢問。

石毓智(2011)並以詢問域的包含範圍為依據，劃分出反義概念形容詞的量級差別，積極成分的量級數目一般為消極成分數目的兩倍，所謂積極成分，是表示事物肯定、正面、如意的性質的詞，消極成分則是表示物否定、反面、不如意的詞，石毓智將積極成分稱為全量幅詞，消極成分稱為半量幅詞。

下面以石毓智(2011)所列舉的其中一組反義形容詞「乾淨/髒」為代表，說明其概念：

表 1.「乾淨/髒」類反義詞分布區間
[Table 1. Distributed Interval of Antonym Type Gan1 Jing4/Zang1]

	積極成分										
	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	L_9	L_{10}	L_{11}
乾淨	+	+	+	+	+	+	+	+	+	+	+
	消極成分										
	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	L_9	L_{10}	L_{11}
髒	-	-	-	-	-	+	+	+	+	+	+

註：L_i 表示量級，i 為量級的大小，量級大小分為：L_1 最不、L_2 十分不、L_3 太不、L_4 很不、L_5 有點不、L_6 不、L_7 有點/比較、L_8 很、L_9 太、L_{10} 十分、L_{11} 最

將 L1 視為區間值 0，L6 為區間值 0.5，L11 為區間值 1，則「乾淨/髒」類的反義形容詞，積極成分的量幅區間為[0, 1]。若將消極成分轉換為積極成分的對應，如將「最不

乾淨」對應轉換至「最髒」，則可以將消極成分和積極成分的區間黏合，消極成分的區間為[0, 0.5]，得到結果為積極成分的量級數目一般為消極成分數目的兩倍，積極成分範圍大，為無標記項；消極成分範圍小，為有標記項。

本文討論的「多/少」則是另一種狀況，應歸類於石毓智(2011)提出的「大、小類詞」。

表 2.「多/少」類反義詞分布區間
[Table 2. Distributed Interval of Antonym Type Duo1/Shao3]

	積極成分										
	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	L_9	L_{10}	L_{11}
多	-	-	-	-	-	+	+	+	+	+	+
	消極成分										
	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	L_9	L_{10}	L_{11}
少	-	-	-	-	-	+	+	+	+	+	+

註：L_i 表示量級，i 為量級的大小，量級大小分為：L_1 最不、L_2 十分不、L_3 太不、L_4 很不、L_5 有點不、L_6 不、L_7 有點/比較、L_8 很、L_9 太、L_{10} 十分、L_{11} 最

此類形容詞，兩者皆是半量幅詞，才是典型的有無標記的問題，而且無法以量幅的大小解釋有無標記的現象。石毓智(2011)認為任何事物都具有長、寬、高等三維性質和質量，此物體存在必須有一個假定的預設量，否則此事物不會存在，因此，「大、小類詞」用於問句時，都具有預設量。

兩個半量幅詞剛好組成區間值[0, 1](即積極成分[0.5, 1]；消極成分[0, 0.5])，加上預設量後，積極成分的量幅變為[0, 1]，消極成分的量幅還是[0, 0.5]，因此積極成分能夠涵蓋至整個詢問範圍，所以為無標形式。

採取量幅級別搭配的解釋，可依據層級標準，具體說明兩類不同特性的反義形容詞的標記特性，但無法解釋為何「數量名」結構中，只允許無標形式「多」進入此構式，而不與有標形式「少」搭配。例如：「十多位選手/*十少位選手」，但「三大盤/三小盤」皆可說。按石毓智(2011)的分類標準，「多」和「少」和「大」、「小」類詞為同一類詞，但由上例可觀察到「多」和「少」於「數量名」結構中的表現，和同一類的「大/小」的語法表現不同，本文認為，雖然「多」和「少」皆兼有多種詞性，但「多」具有作為「數詞」的功能，而「少」則不具備此詞性。

特別說明本文針對「多」做為「數詞」用法的觀點。理論上而言，數詞「多」可視為和其他用法的「多」不同的詞項(entry)，因此「多」和「少」於數詞用法的差異，可能是由於詞彙空缺產生的情形，但本文認為，反義詞組「多」和「少」是由於在「數量名」結構上表現的差異，發展出後續一系列的不對稱現象。

於下面章節，本文將先由文獻對於「多」和「少」的辭典定義和相關語意出發，並

使用「數量名」結構的語序作為語料庫搜尋關鍵設定，以語料庫「數量名」中的各詞類「共現率(Collocation)」作為比較依據，討論「多」和「少」具有「數詞」語法身分的合理性與可能性。

3. 以語料庫經由「數量名結構」探討「多」、「少」的詞類標記 (Occurrences of *Duo1* and *Shao3* in [Numeral + Measure Word/Classifier + Noun] in Sinica Corpus)

如先前討論，本文認為「多」和「少」皆可兼多種詞類，但僅「多」可做「數詞」，「少」則無此用法。下面將於比較「多」和「少」的詞典定義後，使用語料庫的設定，將「數量名」結構作為觀察環境，比較「多」和「少」出現於此語法環境時的表現差異，再透過各詞類在語料庫中的典型搭配，分析「多」的詞類歸屬問題。

本研究內容的語料來源主要來自：

1. 現代漢語平衡語料庫
2. 近代漢語語料庫
3. 作者通過自編自省的語料
4. Google 搜尋引擎

3.1 「多」和「少」的詞典定義 (Dictionary Definitions of *Duo1* and *Shao3*)

程龍飛(2014)考察多本辭典中「多」和「少」的基本義，並且討論兩者的基本義由不同引申路徑發展出的引申義。本文將其多和少的基本義資料整理製表如下。

表 3.「多」的辭典定義比較
[Table 3. The Definitions of Duo1 in Dictionaries]

項目	辭典出處	義項
多	甲骨文、金文	數量大
	《說文解字》	“多，重也。從重夕。夕者，相繹也，故爲多。重夕爲多，重日爲曡。凡多之屬皆從多。”指數量大。
	《現代漢語辭典》(第六版)	1.數量大 2.表示相差的程度大 3.過分的、不必要的
	《現代漢語八百詞》	1.數量大 2.比原來數目有所增加 3.形+多+了；形+得+多。表示相差的程度大、用於比較。
	《現代漢語大辭典》	1.數量大 2.超出原有或應有數目；比原來的數目有所增加。

項目	辭典出處	義項
多	《現代漢語大辭典》	3.過分的、不必要的 4.表示相差的程度大

表 4.「少」的辭典定義比較
[Table 4. Definitions of Shao3 in Different Dictionaries]

項目	辭典出處	義項
少	甲骨卜辭	少、小為一字。
	《正字通》	古小少同。
	《說文解字》	不多也。從小丿聲。
	《現代漢語辭典》(第六版)	數量小。
	《現代漢語八百詞》	1.數量小 2.比原來的數目有所減少；數量尚不足
	《現代漢語大辭典》	數量小；不多

綜合各詞典的定義，程龍飛(2014)認為在現代漢語中，「多」的基本義為[數量大]，作為語素的「多」帶有[+數量大]的語意特徵；「少」的基本義為[數量小]，含有[+數量小]的語意特徵。從各辭典的定義觀察，「多」和「少」通常被視為一組相對的概念。

至於引伸義方面，程龍飛(2014)認為「多」透過隱喻投射的作用，由表示具體事物數量大的功能，發展出表達抽象事物量大的引申義，並由於此種認知模式易於掌握與運用，基本義與引申義產生的時間不會相差很遠，兩者相伴而生，在具體和抽象兩個認知域中分工明確。最後，再由表示抽象事物量大的「多」進一步衍生出表示「程度高」的「多」。程龍飛(2014)認為，「少」也和「多」一樣透過隱喻投射的作用，由表示「具體的事物數量小」發展出表達「抽象事物量小」的功能，但並無再進一步發展出表示「程度低」的用法，原因在於，數量的增大，並沒有最終的限制，可視為無上限，反之，數量的減少，到一定程度便趨近於零，因此無法發展出表達「程度低」的用法。

由「多」和「少」的基本義和引申義，可以歸納出兩項重點：一、「多」和「少」的基本義，一般被視為表達數量大小概念的一對詞。二、「多」和「少」在引申義的發展路徑上，產生了不對稱的情形；「多」的語意發展，比起受到語意限制的「少」，更為廣泛。

下面將使用中研院的現代漢語平衡語料庫，作為比較「多」和「少」語法發展和所出現之語言環境差異的工具。首先，先整理出語料庫中，「多」和「少」的詞類標記，其次，將焦點鎖定於本文所欲檢視的「數量名」結構，透過詞類典型搭配的詞頻高低比

較概念，分析「多」和「少」的語法特性，並且分析「多」應歸入「數量名」結構中的哪一部分。

3.2「多」和「少」在語料庫中的詞類標記 (Part-of-Speech Tagging of *Duo1* and *Shao3* in Sinica Corpus)

以「數量名」結構為檢視標準，「多」和「少」的語法特性差異為：「多」可以進入「數量名」結構，如「三十多人」；而「少」則不被允許進入此結構，如「*三十少人」。經常被視為相對概念詞組的反義詞「多」和「少」，在「數量名」結構中，並不構成一組反義詞。

本研究先鎖定「多」和「少」在語料庫中和「數量概念」有關的語法標記，並進一步以「數量名」的語法排序作為一個檢視框架，論證「多」應為數詞部分(即語料標記的數詞定詞)，「多」和「少」的一系列不對稱現象即是因為兩者是否具有「數詞」身分所引發。

根據中研院的現代漢語平衡語料庫，「多」有九種詞類標記，「少」則僅有四種詞類標記(見表 5)。「多」和「少」用法的不對稱性，可由「多」比「少」多出了五種詞類標記得到初步反映，換言之，由語料庫的標記觀察，詞類標記種類的數量差異(即代表「多」和「少」的不同語法表現)，為「多」和「少」最外顯的用法不一致現象。

表 5.「多」和「少」的詞類標記[2]
[Table 5. The POS Tagging List of Duo1 and Shao3]

「多」的詞類標記[3]	「少」的詞類標記
副詞(D)	副詞(D)
狀態不及物動詞 (VH)	狀態不及物動詞 (VH)
狀態及物動詞 (VJ)	狀態及物動詞 (VJ)
數量定詞 (Neqa)	**數量定詞 (Neqa)**
後置數量定詞 (Neqb)	
非謂形容詞 (A)	
動詞前程度副詞(Dfa)	
動詞後程度副詞(Dfb)	
動作不及物動詞 (VA)	

上文提到，「多」和「少」的基本語意為「±數量」的多寡，作為一組概念上的反

[2] 本文將討論到的詞類標記，以粗黑體標示。其餘詞類不在本文中探討。

[3] 語料庫中詞類標記，數字相關形式標記為「數詞定詞(Neu)」(見表 8)，和表示數量的「數量定詞(Neqa)」不同，語料庫中，沒有將「多」歸為「數詞定詞(Neu)」的用法。

義詞，根據語料庫的資料初步顯示，兩者皆具有「數量定詞」的用法。

概念上為反義詞的「多」和「少」，若在語料庫中能夠以相同的「詞類標記」身分出現，兩者的語法環境應是可以交互替換、用以表達概念相對的語意，然而，觀察語料庫中，「多」和「少」作為「數量定詞」的語法表現，可以發現「多」和「少」即使作為同一種詞類使用，兩者仍呈現出用法上的不對稱現象。

首先，透過語料庫，限定「詞類候選單」為「數量定詞(Neqa)」，並將「多」設為關鍵詞，共得語料筆數 1957 筆。進階處理後，可歸納出「關鍵詞右一搭配項目」，共 29 項；接著，以同樣的「詞類候選單」設定，輸入「少」為關鍵詞，共得語料 26 筆，「關鍵詞右一搭配項目」共三項，見表 6[4]。

表 6.「多/少(數量定詞 Neqa)」之右一搭配前三位詞類項目比較
[Table 6. Comparison of the Top 3 Collocates of Duo1/ Shao3(Neqa) in the Right 1 Position]

詞項	詞類	freq(y)	freq(x,y)	MI	詞項	詞類	freq(y)	freq(x,y)	MI
多	Nf	53923	1374	4.328	少	V_2	4630	18	6.769
	Na	27321	251	3.308		Na	27321	7	4.050
	Nc	4951	85	3.934		VD	11120	1	3.003

由表 6 可得第一點觀察：「多(數量定詞 Neqa)」的語料數量，共有 1957 筆，和「少(數量定詞 Neqa)」的 26 筆用例，差距懸殊，此為兩者在使用頻率上的不對稱。

除此之外，關鍵詞「多(數量定詞用法)」其後接「量詞」的用法共有 1374 筆，佔了「多(數量定詞 Neqa)」約 70%的用法；「少」後則無接「量詞」的用法。換言之，右一搭配項目為「量詞」的用法，只能與「多(數量定詞 Neqa)」搭配出現，此為第二個不對稱現象。

「多(數量定詞 Neqa)」和「少(數量定詞 Neqa)」後接「普通名詞」的用法皆排在第二位，但「多(數量定詞 Neqa)」右一搭配項目第一位「量詞」和第二位「普通名詞」的語料筆數相差五倍之多，至此，可以歸納「多+量詞」為「多(數量定詞 Neqa)」典型的用法，用例比數最高，然而語料庫中卻無「少+量詞」的用例。此現象反映出高度的不對稱，而此現象並非語料庫所收錄之語料的範圍限制所導致，對照實際上的語言使用，(7a)可以成立，(7b)則不合語法。

(7a) 本院歷史語言研究所**多位**研究人員獲獎。

(7b) *本院歷史語言研究所少位研究人員獲獎。

[4] 其他搭配項目和討論無關，表格僅取搭配詞超過 10 次的詞類項目。

(7b)若要成為合法句子，可以於「少位」前加上「僅有」，如(7c)。但句子接受度不及(7d)高。

(7c) ??本院歷史語言研究所僅有少位研究人員獲獎。

(7d) 本院歷史語言研究所僅有一、兩位研究人員獲獎。

王燦龍(1995)提及，一般認為「多」兼有三種詞性，數詞、形容詞和副詞；「少」則有副詞、形容詞兩種用法，但不能作為「數詞」使用。裘榮棠(1999)也針對「多」和「少」的語法差異性，提出十二項差異點，其中一點為：「多+量(物量)+名」，可構成偏正結構，如多項冠軍、多塊金牌，「少」則不行。關於此點，王燦龍(1995)認為，在「多」和「少」前加上「很」，用以取代數量結構時，兩者呈現很大的對立，如：「很多蘋果/*很少蘋果」。裘榮棠(1999)則認為「多」就是「許多」，能跟後面的量詞組合，是數詞；「少」雖然也有幾種詞性，但不能是數詞。尚國文(2012)比較新加坡華語和中國普通話的差別，也將「多」視為「數詞」處理。

本文透過語料庫分析「多」和「少」的數詞身分。操作方法為使用語料庫針對**與「多(數量定詞)」搭配的「量詞」和「多/少+量詞+名詞」**之構式進行分析檢視。語料庫中搭配「量詞」的語例只有「多(數量定詞)」，為確定「少」在此結構中，不能和「多」替換，本研究採取人工檢核是否所有搭配皆無法使用「少(數量定詞)」替代[5]。

首先檢視和「多(數量定詞 Neqa)」搭配的「量詞」用法。將關鍵詞「多(數量定詞用法)」的右一搭配「量詞」去除重複後，可得 62 種量詞搭配項，見表 7。

表 7.數量定詞「多」右一位置之搭配「量詞」
[Table 7. Measure Words with Duo1(Neqa) in Right 1 Position]

分	天	戶	支	日	口	本	件	台
次	位	名	年	些	卷	屆	所	宗
門	枚	則	美元[6]	架	段	面	首	個
時	株	套	家	座	隻	起	條	張
票	組	部	場	處	通	期	棟	款
幅	筆	項	葉	遍	種	臺	樣	層
篇	艘	檔	餐	點	類	齣	元	

[5] 此階段暫且不討論「多」應視為「數量定詞 Neqa」或是「數詞定詞 Neu」的問題。

[6] 此筆語料為「七塊多美元」，「多」應為後置數量定詞，與討論中的「數詞+多+量詞」結構不同。

以上「量詞」，經過人工檢視，排除非正確的量詞用法後，可大致區分為兩大類，即「多」加「量詞」(如：多時、多年、多次)與「多」加「量詞」後再接「普通名詞」(如：多名投手、多筆紀錄、多項大獎)，兩大類情況，都無法使用「少」作為直接替代。

第一類：「多+量詞」

(8a) 等候多時、多年仇恨、中風多次

(8b) *等候少時、*少年仇恨、*中風少次

第二類：「多+量詞+名詞」

(9a) 多名投手、多筆紀錄、多項大獎

(9b) *少名投手、*少筆紀錄、*少項大獎

由語料庫以及語感測試所得結果，「少」無法出現在「數量名結構」當中，和「多」在此結構中的表現，呈現不對稱的現象。若進一步再以「數量名結構」：「十瓶酒」為測試依據，「多」若進入構式中，形成「十多瓶酒」，邏輯上有三種可能情況：

1.「多」為數詞的一部分，「十多」應分析為數詞部分(Neu)，而非「數量定詞 Neqa」。

2.此構式允許「數詞」和「量詞」之間插入「數量定詞」，形成「數詞+數量定詞+量詞」的形式。

3.「多」為量詞的一部分，「多瓶」應分析為量詞部分，而非「數量定詞」。

下面使用現代漢語平衡語料庫分析「數詞[7]」、「數量定詞」以及「量詞」三者間排列組合的關係，分析「多」最適合的語法定位。

3.3 「多」的詞類標記之歸屬 (*Duo1*'s Part-of-Speech Tagging)

3.3.1 「多」為數詞的一部分 (*Duo1* as Numeral)

首先設定語料庫的詞類標記「數詞定詞(Neu)」為檢視項目，經過去除重複後，共得 427 筆項目。經人工檢視並整理出可出現於此位置的形式，其代表形式如表 8。

其中，「約量數」經過「去除重複」處理後，含有「多」的「數詞定詞(Neu)」計有：十多、一百七十多、二十多、五十多、二百五十多、一千多、七千多、一仟多、八仟多、一百多、九百多、三百多、一千三百多等 13 筆項目；若以「*多」為關鍵詞項目，並設

[7] 現代漢語平衡語料庫中的「數詞」為「數詞定詞」，包含基數、序數和「約量數」(如：十幾)，標記為(Neu)。

定「詞類候選單」為「數詞定詞(Neu)」，則可於現代漢語平衡語料庫中蒐得語料 1496 筆；根據語料顯示，語料庫中將「數字+多」視為一個語法單位處理的用例並不罕見。「數字」加「多」表達的是不精確的數目。

表 8.數詞定詞數目字表達形式
[Table 8. The Forms of Neu in Sinica Corpus]

阿拉伯數字形式		中文字數目形式	
小數	4.3	整數	六
整數	3500	整數後加進位單位	七十
整數後加進位單位	33 萬	序數	第七十
序數	第 470	約量數[8]	一千多
電話號碼	365-72945	後加文字	九百多
信箱號碼	1-87	後加單位	三十億
其他數字標示	20-20-20		

尚國文(2012)指出，華語中表達不精確、大概的數目，可使用「概數」，表達方式有「數詞連用(如兩、三天)」或是使用「表約數的詞語(如上下、左右)」，他將「多」視為「數詞」處理[9]。劉月華(1996)也採取將「多」歸類為「概數」的表示法之一，用以表示比前面的數詞所表示的數目略多，語法位置可以歸結為兩種：

A 數詞(以 0 結尾)+多+量詞(各種量詞)(+名詞)，如：三百多斤。

B 數詞(以 1...9 結尾及 10)+量詞(表示連續量)+多+名詞，如：三年多時間。

Yip(2004)則將「多」歸類於「約數(imprecise numbers)」，表達「不確定的超額(indeterminate excess)」概念，用於「十」以上之數目，並指出當「多」和「數字」或「量詞加名詞」共現時，「多」出現在數字之後，在量詞或名詞之前。十以下的數字，可以和「度量衡量詞(standard measure)」或「時間單位詞(temporal noun)」等表示「標準規格量詞(de facto measures)」的詞項搭配，在此用法中，「多」出現在量詞之後，名詞之前(若名詞有出現時)，如：五年多(時間)。

由文獻中可以發現「多」在「數量名」結構中，主要出現的位置有二：

一、「數詞+多+量詞+名詞」

二、「數詞+量詞+多+(名詞)」

8 不精確的整數，如一千多、二十餘、六十幾等，下文以「約量數」稱之，亦包含單位數目在後之形式，如數十萬、幾千萬。

9 尚國文(2012)討論新加坡華語「多」的用法，除了一般緊隨數量結構並至於時間詞之前的用法(如四個多小時)，也可將「多」置於「月」之後，形成「短短一個月多之內」的用法。

上述文獻與語料庫的語料，可以提供將「數字+多」視為「數詞定詞(Neu)」的證據，並說明「多」和「許多」並不能直接替換，如：三十多位老師、*三十許多位老師；五年多、*五年許多。因此，雖然認知概念上，「多」和「很多」、「許多」皆表「量多」，在實際使用上具有不同的語法特性。

主張「許多」是數詞的學者，如裘榮棠(1999)，主要是根據「許多」可以和「量詞」結合的特點(如：現代漢語八百詞、邢福義、馮志純)。唐愛華(2000)則認為「許多」由歷史脈絡、文獻及方言來看，具有作為「形容詞」的句法功能，其重疊方法亦符合形容詞重疊特性，「許多」雖然能跟量詞結合，但也有其他形容詞能夠與量詞結合，考量到「許多」具有數量詞的性質，唐愛華(2000)將其稱為「數量形容詞」。

根據語料庫的標記顯示，「許多」共有三種標記：「動詞後程度副詞(Dfb)」、「數量定詞(Neqa)」及「後置數量定詞(Neqb)」。若「許多」為「數詞」，在語料庫中應有「許多」作為「數詞定詞(Neu)」的用法，但語料庫中的「許多」多為「數量定詞」用法，和「多」的語法表現也不盡相同，將「多」和「許多」皆解釋為「數詞」會衍生許多無法合理解釋的語言現象。

「多」和「許多」在下面的例句中可以交替使用：三十多位老師、許多位老師。若「許多」前不接數詞，又可以出現在「許多位老師」構式中，第二個可能就是「多」和「許多」都應該分析為「數量定詞(Neqa)」。

3.3.2 「數詞」和「量詞」之間插入「數量定詞」 (The Insertion of Neqa between Numeral and Measure Words)

第二個可能分析為「多」是可以允許插入「數詞」和「量詞」之間的「數量定詞」。首先，放寬語料庫範圍，檢視以「數量定詞」為關鍵詞，再將其前後範圍分別限定為「數詞」及「量詞」，搜尋語料庫中，此語言形式的語料數目。

設定過濾條件之先後順序可能影響語料數目，因此操作兩次，第一次以「數量定詞」為關鍵詞，使用「進階處理」功能，限定關鍵詞左一位置為「數詞定詞」，得出經過過濾之第二層語料(為了便於討論，稱此層語料為 Neu 限定)，再進入「進階功能」設定關鍵詞右一位置為「量詞」，可得第三層語料，共有四種形式：一整圈、一整年、一整罐、一整天。

第二次操作則以「數量定詞」為關鍵詞，使用「進階處理」功能，先限定關鍵詞右一位置為「量詞」，得出第二層語料(稱此層語料為 Nf 限定)，再進入「進階功能」設定關鍵詞左一位置為「數詞定詞」，可得第三層語料，依舊得到四種形式：一整圈、一整年、一整罐、一整天。

「Neu 層限定」可得以下形式：

全案整理[10]、一步步、一部份、六遍、一鞭

---

[10] 語料庫將「全」分析為「數量定詞」，但「全案」應為一完整詞彙，故不列入分析。

「Nf 層限定」可得以下形式：

不少次、二分之一個、十數粒、全株、半年、多年、多少年、好多天、好些年、那些個、那麼多年、很多次、若干年、許多組、這些年、這麼多年、無數次、滿籮滿筐、整圈

以上可以整理出「數量定詞 Neqa」的三種典型用法：

典型一：數詞定詞「一」+數量定詞「整」+量詞。

典型二：數詞定詞「一」+數量定詞

典型三：數量定詞+量詞 (以時間量詞「年」較多)

表 9.「數詞定詞+數量定詞+量詞」之構式典型
[Table 9. Typical Construction Type of [Numeral + Measure Word/Classifier + Noun]]

Neu	Neqa	Nf
一	整	圈、年、罐、天
一、六	步步、部分、遍、鞭	
	不少、二分之一、十數、全、半、多、多少、好多、好些、那些、那麼多、很多、若干、許多、這些、這麼多、無數、整	次、個、粒、株、年、天、組、圈

由「Neu 層限定」和「Nf 層限定」進一層檢視得出的結果相同，「數詞+數量定詞+量詞」的結構，最典型的用法是「一+整+量詞」。「多」可以出現於「數量定詞」的位置，其他含有「多」的詞還有「好多」、「那麼多」、「很多」、「許多」和「這麼多」，將「多」歸納為「數量定詞 Neqa」也是可能的分析之一。需要特別注意的是，只有「多」可以在「數詞定詞(Neu)」部分加上「大於十」的數字，其他則不具備此特性，如：二十多次、*二十那麼多次，若將「多」分析為「數量定詞 Neqa」就必須找出特殊的語言規則，說明為何僅有「多」可以允許和帶有「大於十」的數字共現，其他具有相同語法特性的成分則不行。

3.3.3 「多」為量詞的一部分 (*Duo1* as Measure Word)

邏輯上最後一個可能假設為：「多」屬於量詞的一部分。使用語料庫檢視關鍵詞項為「量詞」之左一位置為「多」的語料，搜尋語料檢視數目，分析「多+量詞」是否為典型用法。設定條件後可得語料 51 筆，經過去除重複處理，可得以下項目：

多元、多日、多次、多位、多年、多個、多場、多項、多種。

除了「多元」前接數詞「八仟」之外，其他語料在「多」之前皆無「數詞」。此說明「多+量詞」結構可能有兩種可能：

一、「多」被涵蓋於數詞部分，可以兩種形式出現，即「數目字+多」或「多」單獨出現的形式，換言之，「多」可視為一種「數詞」，能夠單獨佔據「數詞+量詞+名詞」結構中的「數詞」位置。

二、「多+量詞」可視為一個單位，即佔據「數詞+量詞+名詞」結構中的「量詞」位置。

第一種分析可能性高於第二種。首先，以帶入「數詞」作為測試，「X 項大獎」可以形成「三」項大獎、「三十」項大獎、「三十多」項大獎和「多」項大獎。使用語料庫觀察「數詞定詞(Neu)」的特性，「多」前必須加上「大於十」的整數，並無「多」單獨作為「數詞定詞」[11]的用法。但表 9 的語料，基於用法上的相似性，也可作為將「多」視為「數詞定詞」的佐證。

表 10.　「幾」、「數」和「餘」的語法表現
[Table 10. Syntactic behaviors of Ji3, Shu4 and Yu2]

單獨形式	幾	數	餘(不單獨使用)
前加數目形式	十幾、二十幾、六十幾	十數	十餘、七十餘、二十餘、一百二十餘、六十餘、四十餘、千餘、二千餘、三千餘、五千餘、四十五餘、一四餘、百餘、一百餘、八百餘、十萬餘
其他形式	好幾		

雖然「多」與此三者的語法特性不盡相同，但使用上的規則大體一致，即為可單獨使用或前加整數，且整數數值必須不小於十，差異在於搭配整數的範圍，「多」和「幾」較無限制、「數」則僅和十組成「十數」，用法較受限制。

另一方面，若將「**多項**」視為一個共同單位，則「X 多項大獎」前，僅可使用大於十之整數，如「三十」，其他形式皆不合語法，如「*三**多項**大獎」、「*三十多**多項**大獎」，產出之正確語言形式較受限制。

其次，部分量詞允許於「數詞」和「量詞」之間插入形容詞「大」，如「三十大盤」，能以此形式出現的「數詞」及「量詞」都有限制。「數詞」部分，進入此結構頻率最高的數詞為「一」，其他數目詞必須為「整數」形式才進入此構式，如「**三十**大盤」，其他數目詞皆無法進入此構式，如「***二分之一**大盤」「***兩百多**大盤」、「***多**大盤」；「量詞」部分的限制和其語意有關，如「一大盤」可說，「*一大位」則不可說，但無論量詞其原本形式是否受到語意限制，「多盤」和「多位」皆不能進入「數詞+大+量詞+名詞」構式中，「*三十大**多盤**」和「*三十大**多位**」皆無法成立。

上述之「數詞+大+量詞+名詞」之構式，若將「多」視為「數詞」，即可說明語言

[11]「多」於語料庫中的標記處理為「數量定詞」

現象，亦即「整數」才得以進入此構式，因此不允許帶有「多」之數詞形式；反之，將「多加量詞」視為一個單位則無法解釋「*三十大**多盤**」和「*三十大**多位**」為何不合語法。

表 11.「多」的詞類歸屬比較
[Table 11. Comparison of Duo3 as Different POS Tagging]

將「多」視為數詞部分	將「多」視為量詞部分
語言規則 1：精確數目字可進入「數詞+大+量詞+名詞」構式。 如：三大盤蘋果、十大盤蘋果。	**語言規則 1**：精確數目字可進入「數詞+**(多+量詞)**+名詞」構式。 **語言規則 2**：但數目必須「大於十」。 如：*七**多盤**蘋果、三十**多盤**蘋果。
語言規則 2：約數形式不能進入「數詞+大+量詞+名詞」構式。 如：***多**大盤蘋果、***十多**大盤蘋果。	**語言規則 3**：精確數目字不能進入「數詞+大+**(多+量詞)**+名詞」構式。 如：*三十大**多盤**蘋果。

註：根據筆者語感，帶入構式中進行合語法度測試。

綜合上述討論，若將「多+量詞」視為一個單位，所需要的語言規則與限制較多，根據語言使用的經濟原則，是較差的分析方式。

三種可能分析中，將「多」視為「數詞」，可由語料庫中找到支持的語料並涵蓋最多語言事實，解釋力最高。將「多」解釋為「數詞」，也可解釋為何不存在「*三十少盤」的說法，因為「三十多」為數詞形式，「多」在此結構中，並非為與「少」成反義的形容詞，也非概念上能夠與「少」相對的數量定詞，而是表達「約量概數」的數詞定詞，它既可搭配十以上之位數詞構成一個約數單位，亦可單獨使用，表達量多之概念。

然而，將「多」視為「數詞定詞」分析，又會出現以下「多」和「數目字」語法表現不同的矛盾現象，例如：「多位老師」和「十位老師」皆可說，但「很多位老師」可說，卻不能說「*很十位老師」。本文認為，「很多」的「多」和數詞定詞的「多」不必然有相同的語法表現，如「*十老師」和「*多老師」等結構，若不與量詞搭配則不合語法，「很多老師」或是「許多老師」卻是可說的形式，「多」需和「量詞」搭配才合語法的特性和一般數詞相同，本文認為「多」和「很多」、「許多」不同的語法表現，說明「很多」、「許多」或許和「多」具有某種程度上的發展關聯，但「多」在「數量名」結構中，具有和數詞較接近的語法表現。此外，「數量名」結構的合語法度，可能還受到其他的語意限制影響，舉例來說，「幾」在現代漢語平衡語料庫中，是以「數詞定詞」的標記處理，「幾」的語法特性和其他的數字形式更為接近，「幾位老師」、「十幾位老師」和「十位老師」等皆可說，在一般的「數量名」結構下，「多位老師」、「很多老師」也可以說，當「數量名」結構前加上指示詞「這」或「那」時，「這幾位老師」和「那幾位老師」都可以說，但「*這很多老師/*這很多位老師/*這多位老師」和「*那很

多老師/*那很多位老師/*那多位老師」卻不合語法，將「這」或「那」改為「這麼/那麼」時，「這麼多位老師/那麼多位老師」就又可以說，一般而言，「這十位老師」可說，「*這麼十位老師」不能說，出現在相同的結構中卻有不同表現，本文認為是語意搭配的問題而產生的配詞限制，暫不深入討論。

4. 「很多很少」與「不多不少」(*Bu4 Duo1 Bu4 Shao3* and *Hen3 Duo1 Hen3 Shao3*)

將「多」視為「數詞」部分，能夠有效解釋為何「少」無法進入「數量名」結構，就此結構更進一步觀察，可以發現語言現象如下：

(a)「很多」和「很少」皆能進入此結構，但「很少」為較罕見的「有標」用法。

(b)原本無法在語言中使用的「*少位老師」，前面加上「不」，成為「不少位老師」便可說；原本可以單獨使用的「多位老師」，加上「不」後，合語法度大幅降低「??不多位老師」。

裘榮棠(1999)也曾觀察到此現象，他提出『「很少+動」可以，「很多+動」』不行，如：很少休息、*很多休息。裘榮棠(1999)認為，「很少」相當於「不經常」，但頻率上比「不經常」更少。「不少」和「經常」是相對的，如：很少休息/經常休息。但「很多」不能指「經常」，所以和「很少」不能替換。

另外，他也提出『「不少+名」可以，「不多+名」不可以』，他認為「不多」是短語，為「不」加上單音節形容詞「多」，如同*不高山、*不大樹，原本可以存在的單音節形容詞加名詞的形式，如高山、大樹，加上「不」以後便不合語法，並且單音節形容詞「多」能直接修飾名詞的現象為極少數，因此「不+多+名」就更不合語法。而「不少」可以成立，則是因為其本身便為一個詞，相當於「許多」、「很多」的用法。

裘榮棠(1999)的分析具有一定的解釋力，但仍能找到明顯的反例，例如：

「他多話/他不多話」、「他多事/他不多事」。即便兩者皆分析為短語，「多話」和「多事」也可以進入「不+多+名」的形式。

朱德熙(1989)認為「很多」的語法功能相當於數量詞，並指出「很多」可直接修飾後面出現之名詞；或其出現於主語及賓語位置上用以指代名詞，功能與數量詞相當。

另外，朱德熙先生於〈語法講義〉中也提到，「很多」可以作為「謂語」、「定語」、「賓語」等帶有體詞性的成分使用；「很少」則不帶體詞性，但可做定語，用法和「很多」相同。朱德熙先生對於為何「很多」、「很少」於「體詞性」上具有語法差異性，但「很少」卻可以作為「定語」使用的原因並無解釋。

張維耿(1993)則認為，認定「很少」不帶體詞性不夠全面。他舉了兩組例子：

(10a) 我買了十斤鴨梨，很多是壞的。

(10b) 我買了十斤鴨梨，很少是壞的。

(11a) 這筐蘋果很多壞的。

(11b) 這筐蘋果很少壞的。

他認為，「很多」、「很少」的語意對應，皆表事物數量，並處在主語位置，因此「很少」也具有體詞性。張維耿(1993)認為「很多」、「很少」修飾名詞時，後面可以不用帶「的」，這點和數量詞很像，如「三個人」不說「三個的人」，「很多人、很少人」，所以說明「很多、很少」具有體詞性。

值得注意的是，張維耿(1993)提到，作賓語修飾語的「很多」，其相同位置的反義替換為「一點」、「幾個」、「一兩座」之類的，可以說明「很多」相當於數量詞，但他舉的例子，表達數量少的結構，卻不能使用「很少」直接對應，而必須採取其他語言形式，因此「很少」是否也能視為其所謂之數量詞，仍有討論空間。

另外，張維耿(1993)也提到，作數量詞的「很多」、「很少」不能單說「多、少」，如「很多壞的/*多壞的」、「見過很多/*見過多」、「很少人/*少人」，但原因為何並無具體說明。

王燦龍(1995)持不同觀點，他認為「我買了十斤鴨梨，很少是壞的。」一句中，「很少」不具有指代作用，而是「範圍副詞」；並且認為「這筐蘋果很少壞的。」在語感上有些牽強，此句應說為「這筐蘋果沒幾個壞的。」他並以「很多」與「很少」的句法表現差異作說明：「很多+名詞」可以作主語和賓語，但不能作謂語；「很少」不能直接置於名詞前限制名詞，因此作為主語及謂語的情況不存在。王燦龍(1995)舉例句說明，「廣場上很少人」中的「很少」和「人」之間，必須加上動詞「有」，「廣場上很少有人」才是符合現代漢語的使用習慣，換言之，「很少」修飾的成分是「有」，而非限制名詞。

崔顯軍(1996)認為，「很多」既屬於狀態形容詞，又兼有數量詞的功能，因此可以自由地修飾名詞或名詞性結構；「很少」則是有形容詞性，但不具備數量詞功能，所以不能修飾名詞或是名詞性結構。他認為「很多」、「很少」都有謂詞性，但「很多」有體詞性，「很少」則無體詞性，和張維耿(1993)認為「很少」也具有體詞性的觀點不同；此外，「很多」沒有副詞性，「很少」則有副詞性。

回顧文獻後，我們由語言事實觀察，並且使用語料庫中所標記的用法進行討論：「很多」和「很少」可以進入「數量名」結構，並有兩種用法，一為偏向「數詞定詞」，如：「很多位老師/??很少位老師」；另一種用法為省略「量詞」部分，形成：「很多老師/很少老師」，此為「數量定詞」的用法。「數詞定詞 Neu」和「數量定詞 Neqa」兩種用法都能使用，因此將「很多」和「很少」設定為關鍵詞，觀察兩者於語料庫的語法表現，檢視兩者較偏向哪一種類型的使用。

「很多」在語料庫的標記為「數量定詞(Neqa)」及「狀態不及物動詞(VH)」兩種，共有 2995 筆語料；「很少」和「很多」一樣有「數量定詞(Neqa)」及「狀態不及物動詞(VH)」兩種標記，並多了「副詞(D)」與「動詞前程度副詞(Dfa)」兩種用法，共 381 筆

語料。[12]

以「很多(Neqa)」和「很少(Neqa)」為關鍵詞，檢視其右一共現搭配項目(collocation)，將項目限定為詞類，並以詞頻排序，各自前三項目排行如下表:

表 12.「很多/很少」右一共現搭配詞類項目
[Table 12. POS of Collocates of Hen3 Duo1/Hen3 Shao3 in Right 1 Position]

「很多」(Neqa)右一詞類				**「很少」(Neqa)右一詞類**			
y:詞/詞類	**freq(y)**	**freq(x,y)**	**MI**	**y:詞/詞類**	**freq(y)**	**freq(x,y)**	**MI**
Na	27321	1555	4.834	Na	27321	29	5.092
VH	17279	145	2.92	V_2	4630	4	4.886
Nf	53923	69	1.039	P	149034	1	0.028

「很多」和「很少」的右一搭配項目都以「普通名詞(Na)」的用法最多；「量詞」在與「很多」的搭配用法，排在第三位[13]，「量詞」和「很少」則沒有搭配，比較先前表格(重製為表 13)「多(Neqa)」為關鍵詞的右一搭配第一位則是「量詞」，其次才是「普通名詞」。

表 13.「很多/多」右一共現搭配詞類項目
[Table 13. POS of Collocates of Hen3 duo1/ Duo1 in Right 1 Position]

「很多」(Neqa)右一詞類				**「多」(Neqa)右一詞類**			
y:詞/詞類	**freq(y)**	**freq(x,y)**	**MI**	**y:詞/詞類**	**freq(y)**	**freq(x,y)**	**MI**
Na	27321	1555	4.834	Nf	53923	1374	4.328
VH	17279	145	2.92	Na	27321	251	3.308
Nf	53923	69	1.039	Nc	4951	85	3.934

對比語料庫中，典型的「數量定詞(Neqa)」和「數詞定詞(Neu)」用法，兩者右一搭配的前三排行項目如表 14：

[12] 裘榮棠(1999)提到的「很少+動詞」為「很少」的「副詞」或是「動詞前程度副詞」用法，和數量概念上的討論較無相關，下文並不就此現象深入討論。
[13] 「普通名詞(Na)」以虛線外框標示，「量詞(Nf)」以灰底標示；下面表 13、表 14 標示方法相同。

表 14.「數量定詞/數詞定詞」右一共現搭配詞類項目
[Table 14. POS of Collocates of Neqa /Neu in Right 1 Position]

「數量定詞 (Neqa)」右一詞類				「數詞定詞(Neu)」右一詞類			
y:詞/詞類	**freq(y)**	**freq(x,y)**	**MI**	**y:詞/詞類**	**freq(y)**	**freq(x,y)**	**MI**
Na	27321	2072	4.481	Nf	53923	3326	4.274
Nf	53923	393	2.139	Na	27321	547	3.149
VH	17279	354	3.172	Nc	4951	87	3.019

由表格上的對照，「多」應該劃分為「數詞定詞」，而非「數量定詞」，而「很多」和「很少」則和「數量定詞」的用法較一致，其後不接量詞的用法頻率較高，如：**一點**咖啡、**這些**條件、**有的**人。典型的「數量定詞」後一般不接「量詞」，如：「*一點杯咖啡」、「*這些項條件」、「*有的位人」，皆不能成立。

此外，即便部分「很多」+「普通名詞」的結構中間允許插入「量詞」，如：「很多老師」、「很多位老師」。「很多」+「普通名詞」的結構是否有「量詞」明顯有語義上的差異：

(12a) 生活上有「很多事」並不簡單。(語意無特定範圍)

(12b) 生活上有「很多件事」並不簡單。(語意有特定範圍)

「多」加「量詞」加「名詞」的結構，以「很多」替代「多」，語意基本上不受影響，但仍有些例子替換後接受度較降低，大體上說，「多」加「量詞」加「名詞」的結構可以替換為「很多」加「量詞」加「名詞」的結構。反之，「多」與「數詞定詞」則不能進入「很多」+「普通名詞」的結構，取代「很多」的位置，如：「*多老師」、「*三十老師」。換言之，數量定詞「很多」可進入「數詞定詞」的位置，但「多」和其他「數詞定詞」不能作為典型「數量定詞」使用。

由以上討論，「數詞定詞」的使用較「數量定詞」嚴格，也就是表達「量」的形式可以涵蓋表達「數字」，反之不能成立。

李英哲(2001)指出，有些位數詞可以出現在位數「億」、「萬」前面，充當系數，例如《詩經》中的「萬億及秭」，此種用法也出現於現代漢語，如「百萬雄師」，而唐宋詩歌中，有「百、千」做系數表概數的形式，如白居易的詩句「霜竹百千竿」。換言之，位數可以和位數結合，充當一個表概量的系數，也就是說，精確數目字可以和其他的精確數目字搭配使用，表達「概量」的數字觀念。因此本文認為精確數目可以發展出表達模糊概量的用法。下面考察「多」與精確數目字的關係至加入「多」表達約量數的

用法，最後表達「模糊量」的過程。

考察語言中使用數字的形式，整數一到九，皆表示精確的數目，自「十」以後的數目字，語法特性便有不同，如可以在表精確整數的十後頭加上「數」、「餘」和「多」，用以表達「約量數」；由於必須在超過九以後，達到進位基數，才能表達「約量數」的用法，因此，「約量數」需涵蓋最小進位基數的位數詞「十」；此後，更進一步發展出省略數字部分的「約量數」形式，如「多」位老師。

一旦進入表達「模糊量」的形式，量的界線模糊後，可以不必再遵循「數字」形式，因此可以不必大於「十」；所要表達的「量」完全由說話者決定，範圍界線模糊，原來不合語法的「*少位老師」，變成「表量」的「很少位老師」之後，接受度便提高，原因可能為此用法已離典型的「數詞」+「量詞」形式較遠，加上「很多」可以類推為相對概念「很少」，因此合語法度增高。

華語中表達「數量少」的形式，多使用「數詞連用」，即相鄰數目的並列形式，如一、兩位老師。若使用「很少」位老師，表達上為「有標」形式：

(13a) 這次會議只有一、兩位老師參加。

(13b) 這次會議只有很少位老師參加。

(13a)表達的意思很明確，即參加的老師人數很少，(13b)若為可接受的形式，由於使用較為模糊的表達，多了一種「人數極少」的語用效果，日常語言使用不容易出現此形式。

崔顯軍(1996)認為，「很少」與客觀事物或現象之間反映的不是一種「數量關係」，而是一種反映主體的動作或行為的「頻度關係」，出現的形式為「很少+有+名詞或名詞短語」，(13b)不常使用的原因，也有可能是「很少+有+名詞或名詞短語」的形式在語言的使用中較凸顯，使(13b)的解讀較難獲得。

先前討論曾提到，劉月華(1996)、Yip(2004)針對「多」均提出兩種用法，現重複如下：

用法一、「數詞+多+量詞+名詞」。

用法二、「數詞+量詞+多+(名詞)」。

「用法一」的「多」在上述的討論中，將其視為「數詞定詞」，所能解釋的語言現象最多，可用以表達「約量數」的概念，如三十多位老師；「用法二」則為不同的概念，「用法二」中「多」的語法特性和「用法一」則又有些許不同，首先，「用法二」中的「多」必須出現在量詞後，而且量詞普遍是具有參考標準的「時間量詞」或是「度量衡」量詞，如：三天多、二十公里多。「用法二」中的量詞皆是可作為「參照標準」的量詞，並受數詞部分限制，被精確的限定範圍，其後加「多」便是表示鄰近「參照標準」的「增量」的概念，因此不能超出參照標準過多。

以實際上的語感作為測試，「十多公斤」和「十公斤多」，前者的「多」處於表「約數」的位置，後者表示鄰近「參照標準」的「增量」的概念，前者的範圍較模糊，後者的範圍較受限制，「十多公斤」可以指涉「十一公斤」或「十二公斤」；「十公斤多」指涉範圍較小，如「十公斤三十公克」或「十公斤一百公克」，「十一公斤」便超過「十公斤多」的指涉範圍。「用法一」表示「約量」，「用法二」則要求受精確數目限制的參照標準，由於兩種用法具有表達上的互斥性，所以兩者不能共現，亦即「三十多公里」表示「約量」、「三十公里多」表示鄰近參照點的「增量」，表達參照點的數目不能是不精確的「約量」用法，因此「*三十多公里多」不合語法。

「用法二」中的「多」在語料庫中被歸納為「**後置數量定詞 (Neqb)**」，其出現的語言環境可以歸納為下表。

表 15.「多」出現的語言環境及語例
[Table 15. Possible usage of Duo1 and Examples]

語言環境	代表語例
於「數詞定詞+量詞」之後	三點多、一年多、兩歲多
於「數詞定詞+量詞」和「形容詞」之間	兩尺多長、一丈多遠
於「數詞定詞+個」和「量詞或名詞」之間	一個多小時、兩個多星期時間
於「形容詞」和「了」之間	優秀多了、踏實多了、簡單多了

「多」的「後置數量定詞 (Neqb)」用法，可以使用於「形容詞」和「了」之間，表達具有隱含「比較增量」或是「程度增量」的語意。

本文使用「近代漢語平衡語料庫」，觀察近代漢語「多」和「少」在「數量名」結構中的表現，發現「多」和「少」在近代語料中，同樣呈現不對稱的分佈，「多」用數量名結構的用例共 282 筆，佔全部用例約 6%(282/4372)[14]，而「少」出現在「數量名」條件設定下的語料僅有一筆，並且不是真正的數量名結構[15]，因此人工比對排除後，「少」一樣不出現在「數量名」結構當中。由近代漢語語料庫中「多」的右一搭配項目觀察，取搭配次數超過十次的項目如下表。

[14] 近代漢語語料庫針對「多」的標記處理之一，採取分析為「表約數或餘數的依附詞(T7)」的方式。

[15] 語料為「時有一少師普化」，「少」為狀態不及物動詞用法。

表 16. 近代漢語語料庫「多(T7)」的右一搭配項目
[Table 16. Concordance candidates with Duo1(T7) in Academia Sinica Tagged Corpus of Early Mandarin Chinese in Right 1 Position]

y:詞/詞類	freq(y)	freq(x,y)	MI
歲(Nf)	1279	77	6.980
兩(Nf)	3867	40	5.218
年(Nd)	3390	20	4.657
銀子(Na)	3921	19	4.460
日(Nd)	7708	14	3.479
里(Nf)	1592	13	4.982

現代漢語的「多」，除了少數凝固的用法和臨時借用(三十多「車」遊客)外，通常需和量詞搭配以後，才能再帶一個名詞，但由上表可觀察到，同樣的數量名結構：「x 多兩銀子」和「x 多銀子」，在近代漢語都可以說，但與量詞搭配的用例比直接帶名詞的用例多了一倍(40:19。量詞是一種較晚發展成熟的詞類，「多」和「數詞」由近代漢語往現代漢語的發展趨勢是和「量詞」關係更為緊密，而「多」和「許多」、「很多」原本類似的用法(許多銀子、兩千多銀子)卻消失不用[16]，顯示現代漢語中，「多」和「數詞」與「量詞」之間有更緊密牢固的搭配傾向。近代漢語語料庫中，也可搜尋到位於「形容詞」和「了」之間的「多」，如：「無理多了」、「花草樹木也少多了」和「你比他大多了」等用例，但使用頻率不高。綜合比較「近代漢語語料庫」和「現代漢語平衡語料庫」的資料，「多」和「少」於「數量名」結構中，存在不平衡的現象，「少」無法進入「數量名」結構之中；由近代向現代發展，可以發現「多」和「量詞」的關係越趨緊密，「多」和「很多」、「許多」的用法也開始出現分歧。其他形式的數字依附詞，如：幾、餘、數等，於「數量名」結構中，也有不同的特性，「幾位老師」、「數位老師」和「多位老師」可以說，但「*餘位老師」不能說；「二十幾位老師」、「二十多位老師」和「二十餘位老師」可以說，「*二十數位老師」又不能說，語法的合語法度或許受到其他語義搭配的因素影響，我們可以比較保守的說，這些和數字高度相關的「依附詞」，或許正在經歷能否單獨作為「數詞」使用的階段，「幾」和「多」發展比較快速，因此具有可以獨立使用的形式，但兩者仍有發展程度上的差別。

洪舶堯(2012)以共時角度討論「一點」的語法化過程，他認為「一點」最開始的用法是作為不定量詞，用以說明名詞的量，之後再演變為動詞和形容詞的補語。由字面上來看，「一點」最初始的本意，即是「一個點」，數字意義即為「一」。

[16] 類似的用法還有「錢」。

洪舶堯(2012)將「一點」的語法化過程分析為：

表精確數目「一」→作為「不定量詞」→作為動詞和形容詞「補語」

本文認為「多」也是以相同的路徑發展中，可能的發展路徑如下：

數字依附詞「多」(表約量數) >數詞定詞「多」(可獨立使用) >

數量定詞「多」(表量或鄰近增量) >作為動詞和形容詞「補語」

其中，第三個階段，可能和「很多」、「許多」的發展重疊。由近代語料看來，「多」曾經可以不接量詞，並直接和名詞搭配，形成「兩千多銀子」，和現代的「許多銀子」、「很多銀子」相似，但近代漢語語料庫中，僅能找到「名詞」後搭配「甚少」，表達數量少的概念，沒有可以和近代的「許多銀兩」對稱的用法，也沒有類似今日的「很少銀子」的結構。

本文認為，若將「多」具有數詞特性，並且經過一系列的用法發展，就可以解釋為何反義詞「多」和「少」在「數量名」結構中具有如此失衡的不對稱現象：「三十多 vs. *三十少」、「三個星期多 vs. *三個星期少」和「舒服多了 vs. *舒服少了」。原因在於上述的「多」是由數詞形式逐漸發展得來，而「少」則沒有經過此過程，語法的分歧性便可獲得較好的解釋。

最後一項觀察為：原來合語法的「多位老師」在前頭加上「不」，形成「??不多位老師」語法度降低；原來不合語法的「*少位老師」在前頭加上「不」，形成「不少位老師」反而合語法。裘榮棠(1999)認為「不多」是短語，「不少」是詞，造成語法特性上的差異原因為兩者的句法身分不同，但前已談過反例如「不多話」，句法角度的解釋需要更精確的檢視。我們單純就語意面看，「不少」的語義相當於「多」，而「不多」即相當於「少」。「不多」和「不少」都是表達「模糊量」的用法，但概念上表達「少量」的「不多」在構式中受到較高的限制，也就是認知概念上的「多」(也就是「不少」)，才得以進入「數量名」結構，差異在於使用「不」的形式，表達的「量」更模糊。

由上述討論，「標記理論」的概念並不僅僅顯現於語言層面，也存在於認知概念的層級，也就是華語使用者對於認知概念上的「數量多」關注高於「數量少」。

5. 結論 (Conclusion)

總結討論，反義概念「多」和「少」的語法上不對稱現象，不能僅以標記理論解釋。本文以語料庫作為檢視工具，論證「多」應該視為數詞，若語料庫原先將「多」標記為「數量定詞」的處理方式，以「數詞定詞」的標記分析，更能夠精準的反映出「多」和「少」之間的語法殊性。「多」經由「數字依附詞」開始發展出獨立的「數詞定詞」用法，再發展出「表量或鄰近增量」的功能，最後可作為動詞和形容詞「補語」。

「少」並無由數詞路徑發展出來的此一系列用法，因此和「多」有一系列的不對稱現象。而「多」在「數量名」結構中出現的位置，也有不同的語法特性，出現在「數量名」結構中，數和量之間的數詞「多」處於表「約數」的位置；出現在「數量名」結構

中，名詞後的「多」表示鄰近「參照標準」的「增量」的概念，兩者不能共現；最後，「多」和「少」與「不多」和「不少」的特性，反映出華語使用者對數量概念「多」的關注性高於「少」，此概念可以由語言形式展現。

參考資料 (References)

Yip ,P. -C., & Rimmington, D. (2004). Chapter 2. *Chinese: A Comprehensive Grammar* (pp. 17-46). New York, NY: Routledge.

方一新、曾丹(2007)。“多少”的語法化過程及其認知分析。*語言研究*，*3*，76-81。[Fang, Y.- x., & Ceng, D. (2007). “Duo shao ” de yu fa hua guò cheng ji qi ren zhi fen xi. *Yu yan yan jiu, 3*, 76-81.]

王燦龍(1995)。也談“很多”與“很少”。*世界漢語教學*，*2*，28-30。[Wang, C.-l. (1995). Ye tan “hen duo ”yu “hen shao ”. *Shi jie han yu jiao xue, 2*, 28-30.]

石毓智(2011)。*肯定和否定的對稱與不對稱(增訂本)*。北京:北京語言文化大學出版社。[Shi, Y.-z. (2011). *Symmetry and asymmetry between affirmation and negation* (*2nd edition*). Beijing, China: Beijing Language and Culture University Press.]

朱德熙(1982)。*語法講義*。香港:商務印書館。[Zhu, D.-x. (1982). *Yu fa jiang yi*. Hong Kong, China: Shang wu yin shu guan.]

朱德熙(1989)。很久、很長、很多。*漢語學習*，*1*，1-2。[Zhu, D.-x. (1989). Hen jiu, hen zhang, hen duo. *Han yu xue xi (Chinese Language Learning), 1*, 1-2.]

呂叔湘(2000)。*現代漢語八百詞*。北京:商務印書館。[Lü, S.-x. (2000). *Xiandai hanyu babai ci*. Beijing, China: Shang wu yin shu guan.]

李英哲(2001)。*漢語歷時共時語法論集*。北京:北京語言大學出版社。[Li, Y.-z. (2001). *Chinese Grammar Then and Now: Writings on Chinese Diachronic and Synchronic Syntax*. Beijing, Chnia: Beijing language & culture university press.]

沈家煊(1999)。*不對稱和標記論*。南昌:江西教育出版社。[Shen, J.-x. (1999). *Bu dui cheng he biao ji lun*. Nan chang, China: Jiang xi jiao yu chu ban she.]

邢福義(1993)。*現代漢語(修訂版)*。北京:北京高等教育出版社。[Xing, F.-y. (1993). *Xian dai han yu (xiu ding ban)*. Beijing, China: Bei jing gao deng jiao yu chu ban she.]

尚國文(2012)。新加坡華語中的數詞及其相關表達。*華文教學與研究*，*4*，67-75。[Shang, G.-w. (2012). Xin jia po hua yu zhong de shu ci ji qi xiang guan biao da. *Hua wen jiao xue yu yan jiu, 4*, 67-75.]

施一昕(1988)。“多”和“少”的不對稱性。*語文論文集* 3。[Shi, Y.-x. (1988). “duo ” he “shao ” de bu dui cheng xing. *Yu wen lun wen ji 3*.]

洪舶堯(2012)。*漢語動詞後「一點」的語法化*(碩士論文)。取自 http://140.113.39.130/cgi-bin/gs32/tugsweb.cgi?o=dnctucdr&s=id=%22GT079845520%22.&searchmode=basic [Hung, P.-Y. (2012). *The Grammaticalization of Post-verbal yidian 一點’ in Mandarin Chinese* (Master’s thesis). Retrieved from http://140.113.39.130/cgi-bin/gs32/tugsweb.cgi?o=dnctucdr&s=id=%22GT079845520%22.&searchmode=basic]

唐愛華(2000)。“許多”是數詞，還是形容詞。*淮北煤師院學報*，*21*(2)，77-78。[Tang, A.-h. (2000). “xu duo ”shi shu ci, hai shi xing róng ci. *Huai bei mei shi yuan xue bao, 21(2),* 77-78.]

崔顯軍(1996)。再談“很多”與“很少”。*湛江師範學院學報*，*17*(3)，90-92。[Cui, X.-j. (1996). Zai tan “hen duo ”yu “hen shao”. *Zhan jiang shi fan xue yuan xue bao*, *17*(3), 90-92.]

張維耿(1993)。“很多”與“很少”。*漢語學習*，*6*，12-14。[Zhang, W.-g. (1993). “Hen duo ”yu “hen shao ”. *Han yu xue xi (Chinese Language Learning), 6*, 12-14.]

陳昌來、占云芬(2009)。“多少”的詞彙化、虛化及其主觀量。*漢語學報*，*3*，8-15。[Chen, C.-l., & Yun, F.-Z. (2009). “Duo shao ”de ci hui hua, xu hua ji qi zhu guan liang. *Han yu xue bao, 3*, 8-15.]

陸儉明(1985)。“多”和“少”作定語。*中國語文*，*1*。[Lu, J.-m. (1985). “duo ” he “shao ”zuo ding yu. *Zhong guó yu wen, 1*.]

湯廷池(1979)。*國語語法研究論集*。台北:台灣學生書局。[Tang, T.-c. (1979). *Guo yu yu fa yan jiu lun ji*. Taipei, Taiwan: Tai wan xue sheng shu ju.]

程龍飛(2014)。*量度形容詞“多”和“少”的不對稱現象及其認知解釋(碩士論文)*。上海師範大學，上海。[Cheng, F.-l. (2014). *Cognitive explanation and asymmetric behaviors for quantitative adjectives “dou1” and “shao3”*. (Master’s thesis, Shanghai Normal University)]

馮志純、周行健(1995)。*新編現代漢語多功能詞典*。北京:當代中國出版社。[Feng, Z.-c. & Zhou, H.-j. (1995). *Xin bian xian dai han yu duo gong neng ci dian*. Beijing, China: Dang dai zhong guó chu ban she.]

裘榮棠(1999)。“多”與“少”語法功能上的差異性。*中國語文*，*6*，23-25。[Qiu, R.-T. (1999). “Duo” yu “shao”yu fa gong neng shang de cha yi xing. *Zhong guo yu wen, 6*, 23-25.]

劉月華(1996)。*實用現代漢語語法*。台北:師大書苑。[Liu, Y.-h. (1996). *Shi yong xian dai Han yu yu fa* (*Modern Chinese grammar for teachers of Chinese as a second language & advanced learners of Chinese*). Taipei, Taiwan: Shi da shu yuan.]

Computational Linguistics and Chinese Language Processing
Vol. 22, No. 1, June 2017, pp. 53-94

An Approach to Extract Product Features from Chinese Consumer Reviews and Establish Product Feature Structure Tree

Xinsheng Xu[*], Jing Lin[*], Ying Xiao[*] and Jianzhe Yu[*]

Abstract

With the progress of e-commerce and web technology, a large volume of consumer reviews for products are generated from time to time, which contain rich information regarding consumer requirements and preferences. Although China has the largest e-commerce market in the world, but few of researchers investigated how to extract product feature from Chinese consumer reviews effectively, not to analyze the relations among product features which are very significant to implement comprehensive applications. In this research, a framework is proposed to extract product features from Chinese consumer reviews and construct product feature structure tree. Through three filtering algorithms and two-stage optimizing word segmantation process, phrases are identified from consumer reviews. And the expanded rule template, which consists of elements: phrase, POS, dependency relation, governing word, and opinion, is constructed to train the model of conditional random filed (CRF). Then the product features are extracted based on CRF. Besides, two index are defined to describe product feature quantitatively such as frequency and sentiment score. Based on these, product feature structure tree is established through a potential parent node searching process. Furthermore, categories of extensive experiments are conducted based on 5,806 experimental corpuses from *taobao.com*, *suning.com*, and *zhongguancun.com*. The results from these experiments provide evidences to guide product feature extraction process. Finally, an application of analyzing the influences among product features is conducted based on product feature structure tree. It provides valuable management connotations for designer, manufacturer, or retailer.

[*] China Jiliang University
E-mail: {lionkingxxs, linjing, xiaoying, yujianzhe}@cjlu.edu.cn
The author for Correspondence is Xinsheng Xu.

Keywords：Chinese Consumer Review, Product Feature Extraction, Rule Template, Sentiment Analysis, Product Feature Structure Tree

1. Introduction

With the rapid expansion of e-commerce business, the Web has become an excellent source for gathering consumer reviews about products (Turney, 2002; Dave, Lawrence & Pennock, 2003; Dellarocas, 2003; Godes & Mayzlin, 2004; Hu & Liu, 2004a, b; Liu, Hu & Cheng, 2005; Duan, Gu & Whinston, 2008; Forman, Ghose & Wiesenfeld, 2008). Many product review websites (e.g., Amazon.com, Taobao.com) have been established to collect consumer opinions about products. Consumers also comment on products in their blogs, which are then aggregated by Blogstreet.com and AllConsuming.net etc. In addition, it has become a common practice for retailers (e.g., Amazon.com, taobao.com, jd.com) or manufacturers to provide online forums that allow consumers to express their opinions about products they have purchased or in which they are interested. Consumer reviews are essential for both retailers and product manufacturers to understand the general responses of consumers to their products. Proper analysis and summarization of consumer reviews can further enable retailers or product manufacturers to insight consumers' opinions about specific features of products (Liu *et al*., 2005). Consumer reviews also offer retailers a better understanding of the specific preferences of individual customers. Furthermore, from a consumer perspective, consumer reviews provide valuable information for purchasing decisions.

As the number of consumer reviews expands, however, it becomes more difficult for users (e.g., product designer & manufacturers, consumers) to obtain a comprehensive view of consumer opinions pertaining to the products through a manual analysis. Consequently, an efficient and effective analysis technique that is capable of extracting the product features stated by consumers and summarizing the sentiments pertaining to specific product features automatically becomes desirable. This analysis essentially consists of two main tasks: product feature extraction from consumer reviews and opinion orientation identification for these product features (Hu & Liu, 2004a, b; Popescu & Etzioni, 2005; Jindal & Liu, 2006; Wei, Chen, Yang & Yang, 2010).

Product feature extraction is crucial to sentiment analysis, because its effectiveness significantly affects the performance of opinion orientation identification. Several product feature extraction techniques have been proposed in the literatures (Hu & Liu, 2004a, b; Kobayashi, Inui, Matsumoto, Tateishi & Fukushima, 2004; Kobayashi, Iida, Inui & Matsumotto, 2005; Popescu & Etzioni, 2005; Wong & Lam, 2005, 2008; Bahu & Das, 2015). However, product feature extraction and opinion orientation identification suffer huge challenges for Chinese consumer reviews because of the natural complexity of Chinese language (Zhang, Yu, Xu & Shi, 2011; Song, Yan & Liu, 2012; Li, 2013; Zhou, Wan & Xiao, 2013; Liu, Song, Wang, Li & Lu, 2014; Wang, Liu, Song & Lu, 2014). First, there is always no interval between the words of Chinese sentences. It leads to the difficulty of distinguishing Chinese phrases. Besides, some Chinese phrases have synonyms e.g. "电板" (electroplax), which exactly appears at Chinese consumer reviews although it is rare, is a synonym of "电池" (Battery). This kind of product features cannot be recognized and extracted based on frequency item method. Moreover, the syntactic and grammar of Chinese sentences are very complex as well as their structures, e.g. the consumer review "电池/noun 还/adverb 可以/verb" (The battery is good) always expresses the positive evaluation of consumers for the "电池"(Battery). The phrase "可以" (can)[1] is a verb but it acts as an opinion word that modifies the phrase "电池" (Battery). That means, on the context of Chinese language, verbs may also modify nouns or noun phrases and express opinion orientation. Thus, the existing methods that find product features based on adjective are also not enough for Chinese consumer reviews. In addition, there are some specific correlations among product features according to our observations. Some product features extracted from consumer reviews are the attributes of the product, components, or parts such as function, performance, quality, material, and service while some product features are product, components or parts itself. For example, "摄像头(cameral)" and "像素(pixel)" are two product features. The "摄像头(cameral)" is a component of intelligent mobile phone while the "像素(pixel)" is the attribute of "摄像头(cameral)". There is a description relation between the "像素(pixel)" and the "摄像头(cameral)". Therefore, it is always interrelated among product features. How to extract product features effectively from Chinese consumer reviews and establish the interrelations among product features are difficult tasks and huge challenges. This paper focuses on such a text mining issue of Chinese consumer reviews. More specifically, we will establish a structure tree of product features and infer the key factors of influencing the sentiment scores of product features from consumers. The goal is to provide evidences for the designer & manufacturer to improve and update their products effectively.

With these considerations, a technique framework of extracting product features from Chinese consumer reviews and its applications are proposed in which a two-stages optimizing word segmentation solution is proposed to improve the correct rate of word segmentation for supporting product feature extraction from Chinese consumer reviews, and an expanded rule template for CRF, in which two new elements namely governing word and opinion word are added, is developed to deal with complex syntaxes and grammars of Chinese language and implicit opinion words. This increases the precision of product feature extraction and is also helpful for the sentiment analysis for product features. Furthermore, product feature structure

[1]The phrase "可以" expresses a positive opinion that means "good" in Chinese language. However, its literal meaning corresponds to the word "can" in English language. The POS of it defines as verb at word segmanetation process.

tree is constructed considering the natural internal correlations among product features, and an application of inferring the key factors, that influence the preference of consumers for a product feature, is proposed based on Bayes theory whose results can be used as evidences for designers, manufacturers, or retailers to product improvement, market management, etc. Finally, 5,806 consumer reviews from *taobao.com*, *suning.com*, and *zhongguancun.com* are retrieved and used as corpus to explain the applications of these principles and methods proposed in this work. It is innovative method of implementing comprehensive applications based on Chinese consumer reviews at product feature level.

The remainder of this article is organized as follows: In Sect. 2, we review existing product feature extraction techniques and discuss their fundamental limitations to highlight our research motivation. Subsequently, a technique framework of extracting product features from Chinese consumer reviews and its applications are proposed in Sects. 3. Sects. 4 investigate the methods of extracting product features based on CRF. The quantitative characters of product feature including frequency and sentiment score are explored in Sects. 5. On the basis of these, product feature structure tree is constructed in Sects. 6. Categories of extensive experiments are conducted in Sects. 7. Sects. 8 give an example to illustrate the applications of the methods mentioned in this work. Secs.9 discuss our research works. Finally, we conclude with a summary and some future research directions in Sect. 10.

2. Literature Review

Some researchers have devoted to analyzing consumer reviews for valuable information and implementing applications based on it. These analyses and applications essentially consist of two aspects: product feature extraction and opinion orientation identification. Product feature extraction is the foundation of opinion orientation identification, and opinion orientation identification is the application based on product features.

2.1 Product Feature Extraction

Hu and Liu (2004a, b) assumes that product features must be nouns or noun phrases and employs the association rule mining algorithm (Agrawal & Srikant, 1994; Srikant & Agrawal, 1995) to discover all frequent itemsets (i.e., frequently occurring nouns or noun phrases) within a target set of consumer reviews. In addition to association rule mining, other information-extraction-based product feature extraction techniques have also been proposed (Kobayashi, Inui, Matsumoto, Tateishi & Fukushima, 2004; Kobayashi, Iida, Inui & Matsumotto, 2005). Popescu and Etzioni employ KnowItAll and propose OPINE to extract product features from consumer reviews automatically (Popescu & Etzioni, 2005; Etzioni *et al*., 2005). Using a set of domain-independent extraction patterns predefined in KnowItAll, OPINE instantiates specific extraction rules for each product class under examination and then uses these rules to extract possible product features from the input consumer reviews. Wong & Lam (2005, 2008) employ Hidden Markov Models and CRF, respectively, as the underlying learning method to extract product features from auction websites. Liu, Wu & Yao (2006) adopted supervised method to extract product features and compare variety of products for consumers based on them. Choi and Cardie (2009) presented the methods of recognizing the product feature from consumer reviews based on CRF.

2.2 Opinion Orientation Identification

Opinion orientation identification is to determine the sentiments of consumers for product features. Therefore, product feature extraction and opinion orientation identification cannot be separated in practice. Li *et al*. (2010) researched the extraction methods of opinion words for product features by integrating two CRF variables such as Skip-CRF and Tree-CRF. Htay and Lynn (2013) extracted product features and opinion words using pattern knowledge in customer reviews. Yi and Niblack (2005) worked on identifying the specific product features and opinion sentences by extracting noun phrases of specific patterns. Zhuang, Feng and Zhu (2006) proposed a supervised learning method based on dependency grammatical graph to extract product feature and opinion information. Yin and Peng (2009) studied the sentiment analysis for product features in Chinese reviews based on semantic association. Ouyang, Liu, Zhang and Yang (2015) investigated features-level sentiment analysis of movie reviews. And Chen, Qi and Wang (2012) extracted multiple types of feature-level information from consumer reviews.

In addition, topic/opinion summary is also an important aspect based on product feature extraction and opinion orientation identification. For example, Miao, Li and Zeng (2010) executed the topic extraction from movie reviews based on CRF. Turney (2002) investigated the unsupervised classification of reviews based on semantic orientation.

However, the existing product feature extraction and application techniques for English language cannot be used to deal with Chinese language directly because of the natural complexity of Chinese language mentioned above. Then some experts explore the product feature extractions and applications from Chinese consumer reviews. Li, Ye, Li and Law (2009) and Zu and Wang (2014) researched product feature extracting methods from Chinese customer online reviews. Liu and Wang (2013) proposed a keywords extraction method based on semantic dictionary and lexical chain. Ma and Yan (2014) presented the product features extraction of online reviews based on LDA model. In order to process Chinese language sentences effectively, Liu and Ma (2009) investigated the Chinese automatic syntactic parsing issues. Similarly, Li (2013) researched the Chinese Dependency Parsing for product feature extraction. Jiang *et al*. (2012) also proposed a method to enhance the feature engineering for CRF by using unlabeled data. From the perspective of applications, Chang, Chu, Chen and

Hsu (2016) investigated the linguistic template extraction for reader-emotion features based on Chinese text. Wang and Meng (2011) studied the opinion object extraction based on the syntax analysis and dependency analysis. Lv, Zhong, Cai and Wu (2014) investigated the task of aspect-level opinion mining including the extraction of product entities from Chinese consumer reviews. Besides, Hu, Zheng, Wu and Chen (2013) developed a method of extracting product characteristic from consumer reviews to provide users with accurate product recommendation. Dai, Tsai and Hsu (2014) presented a joint learning method of entity linking constraints from Chinese consumer reviews based on markov-logic network. Wang and Wang (2016) investigated comparative network for product competition in feature-levels through sentiment analysis. These literatures exactly proposed some effective methods of extracting product features from Chinese text, and used them at specific research tasks. These methods can be classified into two major approaches: supervised and unsupervised.

Supervised product feature extraction techniques require a set of preannotated review sentences as training examples while unsupervised product feature extraction approach automatically extracts product features from consumer reviewers without involving training examples. Generally, the supervised methods have better results at the precision, recall or *F*-score than those of the unsupervised methods because it can set the training samples according to specific research or application goals (Li *et al*., 2009; Zu & Wang, 2014; Ma & Yan, 2014). This work focuses on supervised product feature extraction issues and its applications.

3. Product Feature Extraction Technique for Chinese Consumer Reviews

Aiming at Chinese consumer reviews, a technique framework of product feature extraction is proposed that consists of three key phases: word segmentation and optimization, product feature extraction based on CRF, and the quantitative descriptions of product features. The proposed technique begins with the preprocessing of the inputting consumer reviews, where the preprocessing task includes word segmenting & POS tagging, reconstructing noun phrase based on N-gram, filtering and optimizing. Subsequently, product feature extraction process employs CRF to identify product features in which a train set and a rule template for constructing the *model* of CRF are developed. Based on the extracted product features and the results of word segmentation, the quantitative descriptions for product features including the frequency of product feature and the sentiment score of product feature, are constructed. On the basis of these, product feature structure tree is established based on the fact that product features are interrelated. **Figure 1** presents the framework of product feature extraction techniques for Chinese consumer reviews. In the following subsections, we will depict the detailed design and implementation of each phase.

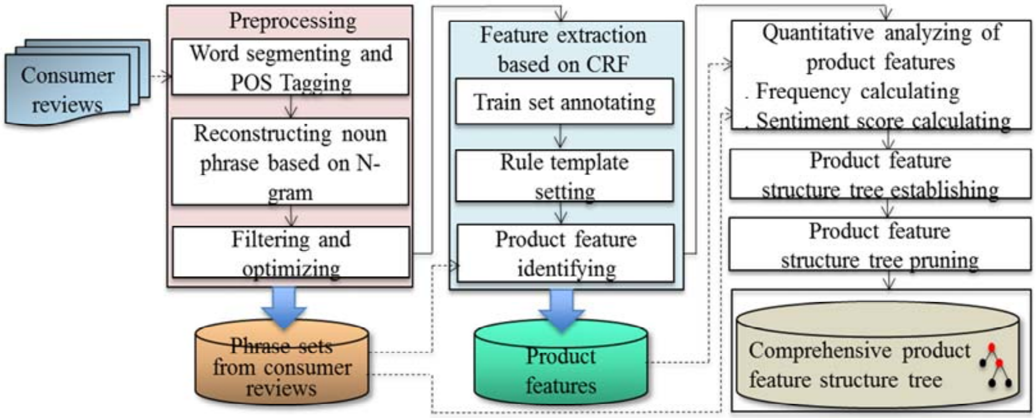


Figure 1. Overview of product feature extraction techniques for Chinese consumer reviews

3.1 Preprocessing Techniques

Preprocessing techniques consist of word segmenting and POS tagging, reconstructing noun phrase based on N-gram, filtering and optimizing.

Phase A *Word Segmenting and POS Tagging*

Word segmenting and POS tagging start with the inputting review sentence S, and end with the pairs ($wordi$, $POSi$), where $wordi$ is the ith word contained in sentence S, and $POSi$ is the POS tagging result of the $wordi$. For the convenience of presentation and measure, phrase (word), sentence, and consumer review are defined respectively as **Figure 2**. In this work, the word refers to phrase in general unless there are specific instructions. For the review sentence S “手机的屏幕很模糊(The screen of this phone is very indistinct)”, the word segmenting and its POS tagging are as follows: (手机(phone), n), (的[2], ude1) , (屏幕(screen), n), (很(very), d), (模糊(indistinct), a) illustrated in **Figure 2**. At the same time, the dependency relations among these words and their governing words are also identified through syntactic parsing process based on consumer review (Liu & Ma, 2009; Wang & Meng, 2011; Li, 2013; Dai *et al.*, 2014). The objective of this phase is to divide the review sentences into discrete phrases and annotate its POS tag, and provide the data resource for the next analysis phases.

[2] It is an auxiliary word in Chinese language. There is no word corresponding to it in English language.

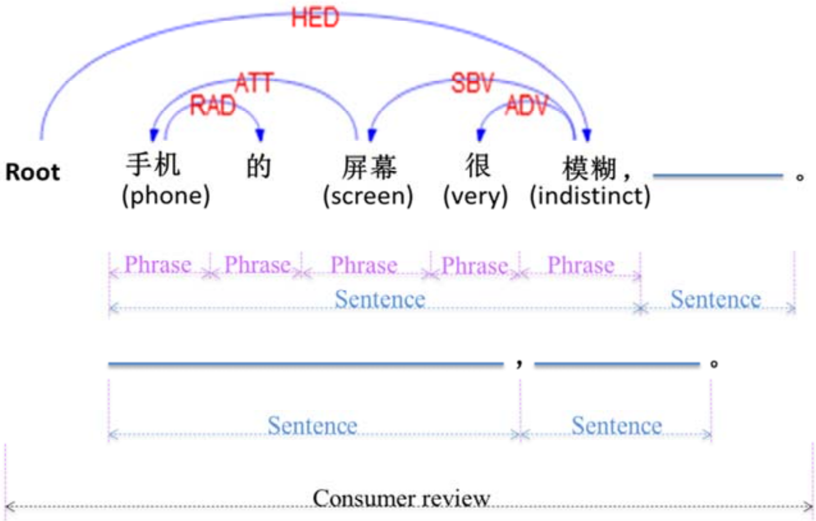


Figure 2. Word segmenting and its POS tagging for a case

Phase B *Reconstructing Noun Phrase based on N-gram*

Word segmenting process may generate some incorrect results sometimes. For example, the phrase ”分辨率(resolution)” always be divided into three kinds of independent phrases “分(divide)”, “分辨(distinguish)”, and “率(rate)”. However, the phrase “分辨率(resolution)” should be a complete phrase for digital product e.g. intelligent mobile phone. Obviously, it is an incorrect result of word segmenting. In order to deal with this problem, it is necessary to recombine these fragmental phrases into its correct form. A reconstruct method based on *n*-gram is introduced which consists of two steps: (a) Identifying the number *n* of the *n*-gram method reasonably; (b) Constructing new phrases according to giving number *n*. Using $wordi$ as an example and assuming *n*=3, then new phrases can be generated by recombining it with adjacent words from left and right directions, respectively. For example, the reconstructing new phrases based on *3*-gram method are as follows ($word_{(i-1)}word_i$, $word_{(i)}word_{(i+1)}$, $word_{(i-2)}word_{(i-1)}word_i$,$word_{(i)}word_{(i+1)}word_{(i+2)}$, $word_{(i-3)}word_{(i-2)}word_{(i-1)}word_i$, $word_{(i)}word_{(i+1)}word_{(i+2)}word_{(i+3)}$. After this reconstructing process, the phrase “分辨率(resolution)” that was incorrect segmented will be restored to its correct from. Likeness, all other incorrect segmented phrases can also be restored to their correct forms through this kind of reconstructing process.

Unfortunately, this phase may also lead to other error phrases due to over-combination. Thus we also need to optimize the results generated from reconstructing phase.

Phase C *Filtering Algorithms*

In order to remove the over-combination phrases from ***Phase*** B, a series of filtering algorithms are employed.

(Ⅰ) Frequency filtering. In general, some new combination phrases which are incorrect such as "屏幕很(screen very)" or "的屏幕('s screen)" seldom occurrence at consumer reviews. Therefore, we can remove them through frequency filtering process by setting a reasonable threshold. An expression for frequency filtering is generalized as follows:

$$\text{If } Number(word_i') \leq Q_1 \text{ then remove it from } \Omega \quad (1)$$

where $word_i'$ is a phrase generated from *Phase B*. And Ω is the phrase group of $word_i'$'s. $Number(word_i')$ is the function that calculates the number of the $word_i'$ appearing at consumer reviews. Q_1 is the threshold of frequency filtering process.

This filtering rule means that the $word_i'$ whose frequency appearing at consumer reviews less than Q_1 will be removed from Ω.

(Ⅱ) Cohesive filtering. However, there are another kind of phrases such as "就这样(That's it)" which consist of two frequency words "就[3]" and "这样(this/it)", and is also a frequency phrase because of the expression habit of Chinese. But it is not a valid phrase. This kind of phrases still cannot be removed through frequency filtering process only.

According to our observation, the constitute elements of a phrase, for example "分辨(distinguish)" and "率(rate)" are two constitute elements of the phrase "分辨率(resolution)", are always strongly coupled among them. That means the cohesive among them is very strong. However, the cohesive among the constitute elements of the over-combination phrases generated from *Phase* B is weak because the combination form of these elements is seldom or may not exist at consumer reviews at all. Therefore, we can use cohesive to remove these phrases from the results of *Phase B*. The cohesive among the constitute elements of a phrase is generalized as follows (Li *et al*., 2009):

$$Coh(word_i') = {Fre(word_j'')_{(word_i')}} \Big/ {(Fre(word_i') + 1)} \text{ and } word_j'' \subset word_i' \quad (2)$$

where $Fre(word_j'')$ is the frequency of phrase $word_i'$ occurring at the results of original word segmentation. $word_j''$ is one of the constitute elements of phrase $word_i'$. $Fre(word_j'')_{(word_i')}$ is the frequency of the constitute elements $word_j''$ of phrase $word_i'$ occurring at the results of original word segmentation.

Then, the expression of cohesive filtering is generalized as follows:

[3] It is an auxiliary word in Chinese language. There is no word corresponding to it in English language.

If $Coh(word_i') \geq Q_2$ then $word_i'$ is not a correct phrase (3)

Through cohesive filtering process, the over-combined frequency phrases that consist of two frequency words can be removed from phrase set $word_i'$.

(Ⅲ) Left entropy and right entropy filtering. In addition, a complete phrase always has various neighbors including left neighbors and right neighbors. If a phrase has a fixed neighbor either left neighbor or right neighbor, it is always not a complete phrase. For example, phrase "诺基亚" (Nokia: a band of mobile phone) should be a complete phrase. But it is always divided into two separated words "诺基[4]" and "亚[5]". Although the process of reconstructing phrase can generate its complete form "诺基亚(Nokia)", but some incorrect word segmentation results such as "诺基" and "亚" still exist at the original word segmentation results. Therefore, it is necessary to remove these phrases from the original word segmentation results to keep the accuracy of word segmentation results.

The calculation models of the left entropy and the right entropy are defined as follows, respectively (Li *et al*., 2009):

Left entropy: $H_L(U) = \sum_i \frac{C_{Li}}{n} \times log \frac{C_{Li}}{n}$ (4)

where C_{Li} is the number of the ith left neighbor appearing at the results of the original word segmentation. n is the number of the current phrase appearing at the results of the original word segmentation. $H_L(U)$ is the left entropy of the current phrase.

Right entropy: $H_R(U) = \sum_i \frac{C_{Ri}}{n} \times log \frac{C_{Ri}}{n}$ (5)

where C_{Ri} is the number of the ith right neighbor appearing at the results of the original word segmentation. $H_R(U)$ is the right entropy of the current phrase.

On the basis of these, an expression of the left entropy and right entropy filtering is generalized as follows:

If $H_L(U) \geq Q_3$ or $H_R(U) \geq Q_3$ then $word_i'$ is not a complete phrase (6)

where Q_3 is the threshold of the left entropy and right entropy filtering process.

3.2 Optimizing Word Segmentation Process

Through reconstructing phrase and three filtering processes, some incorrect word segmentations are removed from the results of original word segmentation and some fragmented phrases are restored also. Besides, some valuable new phrases corresponding to specific research object can also be found during these processes. By adding these new

---

[4] 诺基亚 is a transliteration word of brand name of mobile phone in Chinese language. There is no word corresponding to "诺基" in English language.

[5] Likeness, there is no word corresponding to "亚" in English language.

phrases into the user dictionary which is the important evidences of word segmentation process, then the word segmentation process will restart again based on this extended user dictionary. Thus, the process of word segmentation in this work contains two stages which is presented in **Figure 3**. These two stages can optimize the results of word segmentation to provide valid data resources for the next product feature extraction process.

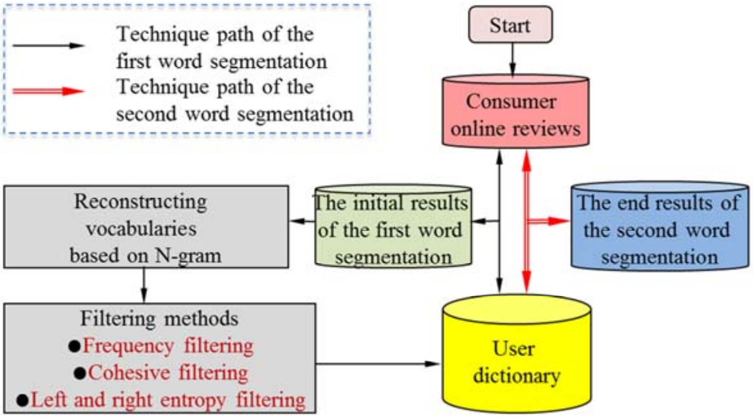


Figure 3. Two-stages word segmentation process

4. Product Feature Extraction based on CRF

The CRF (Lafferty, McCallum & Pereira, 2001; Jakob & Gurevych, 2010) is a sequence modeling framework that can solve the label bias problem in a principled way. CRF has a single exponential model for the joint probability of the entire label sequence given the observation sequence which assign a well-defined probability distribution over possible labeling, trained by maximum likelihood or MAP estimation. Therefore, the weights of features at different states can be traded off against each other. CRF perform better than HMMs and MEMMs when the true data distribution has higher-order dependencies than the model, as is often the case in practice (Zheng, Lei, Liao & Chen, 2013; Zhang & Li, 2015). With these considerations, CRF is employed to extract product features from Chinese consumer reviews in this work. The principles of CRF can be described as follows:

Let X is a random variable over data sequences to be labeled. Y is a random variable over corresponding label sequences. And $X = (x_1, x_2, \cdots, x_t)$ might range over natural language sentences, and x_i denotes the ith phrase in X. $Y = (y_1, y_2, \cdots, y_t)$ range over POS taggings of those sentence Xs, and y_i is the POS tag of the phrase x_i. It is illustrated in **Figure 4**. The random variables X and Y are jointly distributed. CRF, with the known observation data sequence X, calculate the conditional probability $p(Y|X)$. As a result, the POS tag sequence Y that corresponds to the maximum value of the conditional probability $p(Y|X)$ will be label sequence of the X.

The conditional probability $p(Y|X)$ can be calculated as follows:

$$p(Y|X) = \frac{1}{Z(x)} \exp(\sum_i \sum_k \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_i \sum_k u_k s_k(y_i, x, i)) \quad (7)$$

where $t_k(y_{i-1}, y_i, x, i)$ is the transfer character function. It denotes that the label corresponding to the $(i-1)$th element in the observation sequence X is y_{i-1}, and the label corresponding to the i th element in the observation sequence X is y_i. $s_k(y_i, x, i)$ is the status character function. It denotes that the label corresponding to the ith element in the observation sequence X is y_i. λ_k and u_k are the weights for the transfer character function and the status character function, respectively.

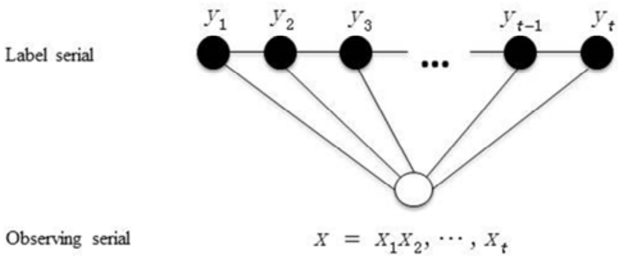


Figure 4. Undirected graph of conditional random fields

According to the principle of CRF, the process of extracting product feature from the results of word segmentation mainly contains two tasks: annotating train set and designing rule template.

4.1 Annotating Train Set

Annotating train set, based on the results of the preprocessing phase including POS tag, dependency relations, and governing words, is to identify the opinion words, product features and their types that is presented in **Figure 5**.

4.1.1 Opinion Word Identifying

For Chinese language, opinion words may also be other kinds of POSs, not just adjective. For example, Chinese phrase “可以(can)[6]” is a verb but it may express a positive opinion of consumer sometimes. This is one of the notable differences between Chinese language and English language. However, these phrases are usually not included in traditional opinion word set. This leads to the inaccuracy of the sentiment analysis for product features inevitably,

[6] The phrase “可以” expresses a positive opinion that means “good” in Chinese language. However, its literal meaning corresponds to the word “can” in English language. Therefore, the POS of it defines as verb at word segmanetation process.

especially for Chinese product features. In order to analyze Chinese product features effectively, it is necessary to identify this kind of opinion words. Table 1 presents these unusual opinion words (partial) based on the analysis for Chinese language at preprocessing phase. Using them, many nouns or noun phrases can be identified and evaluated. This is high significant for product feature extraction from Chinese consumer reviews and its sentiment analysis.

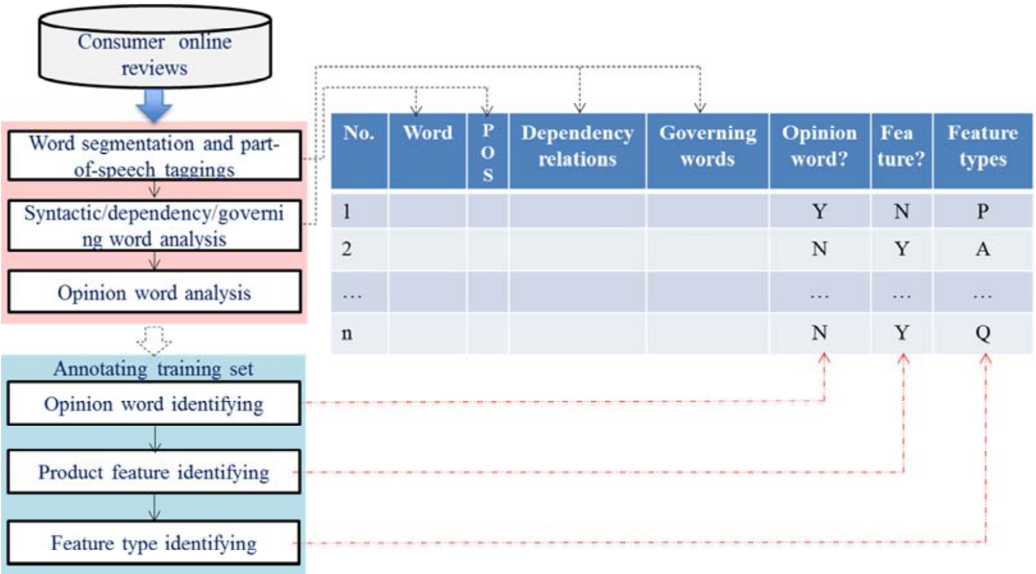


Figure 5. Annotating train set process

Table 1. Unusual opinion words at Chinese consumer reviews

No	POS	Phrases	Sentences
1	v	可以(can), *et al*.	“手机的分辨率还可以” (The resolution of this phone is good)
2	n	战斗机(fighter), *et al*.	“手机中的战斗机” (It is a **fighter** among phones)
…	…	…	…

4.1.2 Product Feature Identifying

Product feature identifying is a crucial step for supervised feature extraction method. It will affect the validity of product feature extraction directly. A reasonable size of train set is necessary to keep the accuracy of product feature extraction. Therefore, it is a time-consuming manual annotating process.

4.1.3 Feature Type Identifying

In general, the product features extracted from consumer reviews include contents and types. For example, some product features refer to the product, and some product features refer to the components/parts constituting this product while some product features refer to the attributes of the product or the components/parts. Furthermore, these attributes can be grouped into the function, performance, quality, and service and so on. Distinguishing these product features carefully can help designer, manufacturer, or retailer to insight into the correlation and influence characters among them. It provides evidences for deep comprehensive applications based on product features. Therefore, identifying feature type is very necessary.

Considering the types of product features and their classifications as well as the interrelations among them, a hierarchical structure for product features can be constructed which is presented in **Figure 6**. This hierarchical structure consists of two parts: basic product structure and the product features describing the attributes of the nodes in basic product structure such as function, quality, and (or) service. Basic product structure consists of root node (product), components, and parts which may also be extracted from consumer reviews and are product features. And the attributed product features including function, quality, and service are the expanded descriptions to corresponding product, component, or part. This product feature structure tree connects the attributed product features with corresponding product, component or parts. It is the foundation for implementing deep comprehensive applications based on product features.

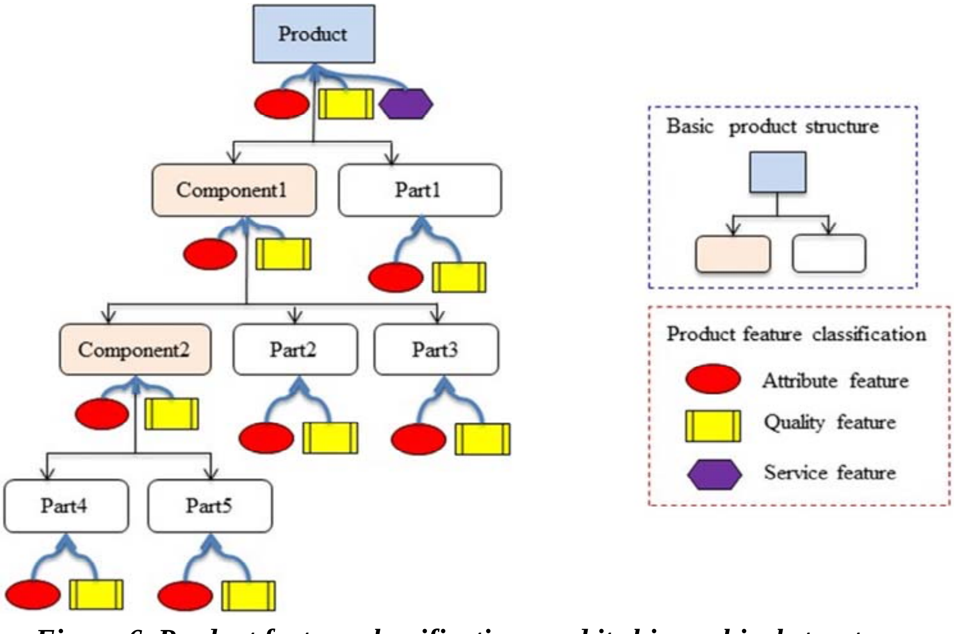


Figure 6. Product feature classifications and its hierarchical structure

4.2 Rule Template Designing

Product feature extraction based on CRF need a rule template to train its *model* which is the core module of CRF to guide product feature extraction process. According to the requirements of our research works, an approach of designing the rule template for Chinese product feature extraction is proposed. It mainly includes three aspects of works such as the core elements of rule template, the unit structure of rule template, and the organization form of rule template. Considering the characters of Chinese language, the core elements that consist of rule template are presented in **Table 2** which contains word elements (including phrase, POS, and context), syntactic elements (including dependency relations and governing words), and sentiment element (opinion words). Each element is also explained in detail in **Table 2**.

Table 2. Core elements of rule template and their explains

Elements	Contents	Explains
Word form elements	Phrase	Element denotes a phrase
	POS	Element denotes the POS of the current phrase
	Context (front or back)	Element denotes the phrases that locate at the front of the current phrase or at the back of the current phrase
Syntax elements	Dependency relation	Element denotes the dependency relation between the current phrase and its governing word
	Governing word	Element denotes the governing word that belong to the dependency relation between them
Opinion elements	Opinion words	Governing word is an opinion word or not

These elements describe the current phrase and the concerned information around it that are very useful to identify product feature. The utilization unit of these elements can be described as a three tuple $< p, \Omega, "T" >$ which is explained in **Figure 7**.

Where p denotes the position information of the elements. Ω denotes the content information of the element such as phrase (0), POS (1), dependency relation (2), governing word (3), and opinion word (4). And T denotes the value corresponding to the element that is determined based on p and Ω. For example, the unit [1, 1,"n"] means that the POS of the phrase that is next to the current phrase is a noun. Using this mode, we can design the contents at a given position to deal with the various expression forms of Chinese language. In practice, the elements in **Table 2** are always combined when establishing the rule template to increase the accuracy and efficiency of extracting product features. The combination forms of elements and its implications are presented in **Figure 8**.

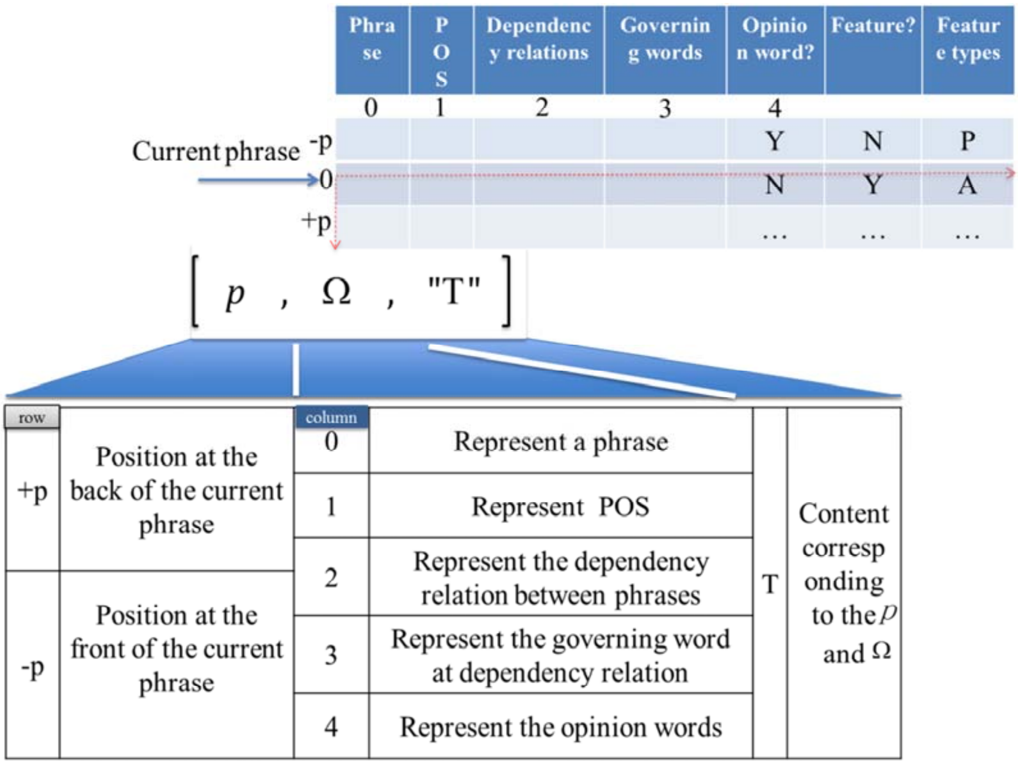


Figure 7. Unit structure and its explains

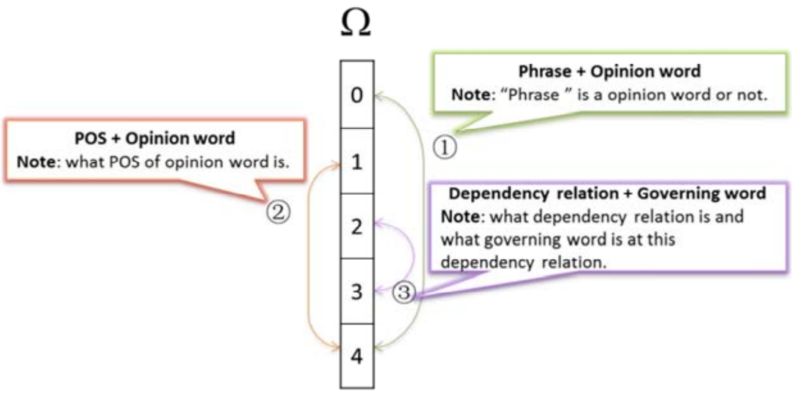


Figure 8. Combination forms of the elements and its implication

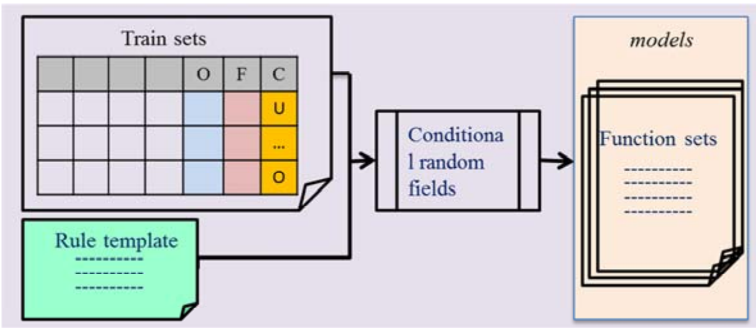
There are mainly three kinds of combination forms in this work namely ***phrase*** + ***opinion word***, ***POS*** + ***opinion word***, and ***dependency relation*** + ***governing word***. The combination "***phrase*** + ***opinion word***" describes whether the current phrase is an opinion word or not. The combination "***POS*** +***opinion word***" describes what is the POS of the opinion word. And the combination "***dependency relation*** + ***governing word***" describes the dependency relation between two phrases and what is the governing word of this dependency relation. These combination utilizations of the elements, together with their sole utilizations, form a complex architecture of rule template which is illustrated in **Figure 9**. Based on it, product feature extraction for specific task can be achieved well.

$$
\begin{array}{l}
\cdots \\
[\ p_k\ ,\ \Omega_i\ ,\ \text{"T"}\ ] \\
\cdots \\
[\ p_k\ ,\ \Omega_i\ ,\ \text{"T"}\ ]\cdots[\ p_{k+h}\ ,\ \Omega_i\ ,\ \text{"T"}\ ]\cdots \\
\cdots \\
[\ p_k\ ,\ \Omega_i\ ,\ \text{"T"}\ ]\cdots[\ p_{k+h'}\ ,\ \Omega_{i'}\ ,\ \text{"T"}\ ]\cdots \\
\cdots
\end{array}
$$

$$
\begin{array}{l}
s.t. \\
i,i' \in \{0,1,2,3,4\} \quad and \quad i \neq i' \\
k,h,h' = const.
\end{array}
$$

Figure 9. General organization form of rule template

Based on train set and rule template, the models of CRF can be established through in-depth learning process which is presented in **Figure 10**. This learning process constructs a large amount of function sets which will be used in models to calculate the conditional probability of elements co-occurring with the form of rule unit description at consumer reviews. Then these results are used to calculate the probabilities at the transfer character and those of the state character in **Equation (7)**, respectively.



O: Opinion **F**: Feature **C**: Classification

Figure 10. Training the models of CRF

5. Quantitative Characters of Product Features

Quantitative description is the foundation of analyzing product features precisely. In this work, the quantitative characters of product features are investigated from two aspects: the frequency of product feature and the sentiment score of product feature which reflect the extent of consumer paying attention to them, and the positive or negative feeling of consumer for them, respectively.

5.1 Frequency of Product Feature Occurring at Consumer Reviews

The frequency of product feature occurring at consumer reviews reflects the extent of customer paying attention to it. For example, consumer maybe like a product feature very much or disappoint very much when the frequency of it is very high. The frequency of a product feature occurring at consumer reviews is generalized as follows:

$$Num\_F_i = \sum_s^{n_s} k_{is} \tag{8}$$

where n_s denotes the number of all consumer reviews. k_{is} denotes the number of the ith product feature appearing at the sth consumer review. Num_F_i denotes the frequency of the ith product feature occurring at consumer reviews.

5.2 Sentiment Score of Product Feature

Generally, the evaluation of consumers to a product feature is either positive or negative, and its strength is different as well. How to describe this kind of distinguishes and how to measure its strength are very important to insight into the preference of consumers precisely.

After analyzing 3,000 consumer reviews, we find that the language pattern of consumer evaluating a product feature is mainly manifested as follows:

$$(adv., adj., pf) \tag{9}$$

where pf denotes a product feature. $adj.$ denotes the adjective that modifies product feature pf. And $adv.$ denotes the adverb that modifies the adjective $adj.$. The adverb $adv.$ and the adjective $adj.$ modify the product feature pf together.

The adverb $adv.$ and the adjective $adj.$ that modify the product features are always qualitative descriptions at consumer reviews. In order to describe the strengths of these adjectives $adj.$ and their polarity as well as those of adverb $adv.$ for the goal of calculation and comparison, the adjective $adj.$ and the adverb $adv.$ should be transformed to numerical value according to their strength and polarity. In this work, the adjective $adj.$ is defined as the range [-9, +9], and the adverb $adv.$ is also defined as the range [-9, +9]. From 1 to 9, strength is increasing gradually. And the minus sign denotes opposite polarity (namely negative). Then, the sentiment score of the ith product feature pf is generalized as follows:

$$Sco_{F_i} = \frac{1}{T}\left\{Sco_{F_{iP}} + Sco_{F_{iN}} + Sco_{F_{iM}}\right\}$$
$$= \frac{1}{T}\{\sum_{x=1}^{a}(Strong_{Px} + Strong_{PxA}) - \sum_{y=1}^{b}(Strong_{Ny} + Strong_{NyA}) +$$
$$\sum_{z=1}^{c}[\sum_{z1=1}^{pz}(Strong_{Mz_{Pz1}} + Strong_{Mz_{Pz1A}}) -$$
$$\sum_{z2=1}^{nz}(Strong_{Mz_{Nz2}} + Strong\_Mz\_Nz2A)]\} \quad (10)$$

$$T = a + b + \sum_{z=1}^{c}(pz + nz) \quad (11)$$

where Sco_F_i denotes the sentiment score of the ith product feature F_i. a, b, and c denote the number of the positive consumer reviews concerned with the ith product feature F_i, the number of the negative consumer reviews concerned with the ith product feature F_i, and the number of the neutral consumer reviews (the consumer review that has multiple different polarity opinion words is defined as neutral consumer review in this work because it is difficult to identify its exact polarity) concerned with the ith product feature F_i, respectively. $Strong\_Px$ denotes the score of the adjective nearby the ith product feature F_i at the xth positive consumer review. And $Strong\_PxA$ denotes the strength of the adverb that modifies the nearest adjective at the xth positive consumer review. $Strong\_Ny$ denotes the sentiment score of the adjective nearby the ith product feature F_i at the yth negative consumer review. $Strong\_NyA$ denotes the strength of the adverb that modifies the nearest adjective at the yth negative consumer review. pz is the number of the positive adjective that correspond to product feature F_i at the zth neutral consumer review, and nz is the number of the negative adjective that correspond to product feature F_i at the zth neutral consumer review. $Strong\_Mz\_Pz1$ denotes the sentiment score of the $z1$th positive adjective of the zth neutral consumer review, and $Strong\_Mz\_Pz1A$ denotes the strength of the adverb that modifies the $z1$th positive adjective at the zth neutral consumer review. Likeness, $Strong\_Mz\_Nz2$ denotes the sentiment score of the $z2$th negative adjective of the zth neutral consumer reviews, and $Strong\_Mz\_Nz2A$ denotes the strength of the adverb that modifies the $z2$th negative adjective at the zth neutral consumer review.

The sentiment score reflects the preference of consumers to a product feature and its extent comprehensively. It can provide the evidences for retailer, designer, or manufacturer to precisely implement product improvement, and market strategy *et al*.

6. Product Feature Structure Tree Constructing

Product features that correspond to the attributes of the product, components, or parts should be connected with relevant objects (namely product, components, or parts) in order to implement in-depth analysis and comprehensive applications at product features level. According to the classifications of product features, product features form a tree structure in general which is presented in **Figure 6**, namely product feature structure tree. In order to construct this product feature structure tree, a basic product structure is employed which is an existing product structure and used as frame, and the nodes of it are also the potential parent nodes for attributed product features. It needs to be noted that the nodes of the basic product structure should also the product features extracted from consumer reviews. Therefore, the key effort of constructing product feature structure tree is to find corresponding parent nodes for each attributed product feature from the results of word segmentation, and compare the corresponding parent nodes with the nodes of basic product structure.

6.1 Finding Potential Parent Node for Current Product Feature

The potential parent nodes of current product features are always the parts, components, or even product. They are noun phrases. And they always co-exist with these attributed product features. Besides, considering the expression habits of Chinese consumer reviews e.g. some consumers may mention the parts or product first when they comment an object, and then evaluate its attributes for example "照相机的像素太低(The pixels of the camera is too poor)" while some other consumers may evaluate the attributes of the parts or product first, and then mention the parts or product for example "续航时间长(long battery life)，电池杠杠的(the battery very good)". Thus, keeping the current product feature pf_i as a central point, the process of finding potential parent node for current product feature pf_i is to search the phrase that satisfies with specific POS and type requirements, or the dependency relation with current product feature pf_i based on given step-length from left and right direction illustrated as **Figure 11**. The pseudo-code description for the algorithm of finding potential parent node for current product feature is presented in **Figure 12**.

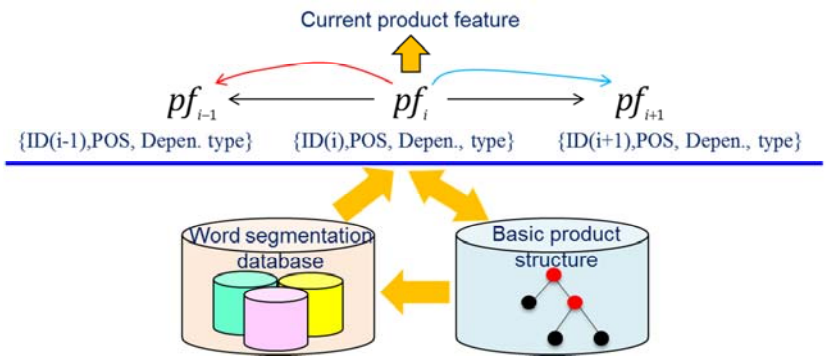


Figure 11. Principle of finding the parent nodes for current product features

Algorithm 1. Pseudo-Code for finding potential parent node for the current product feature

```
//Input:     R – Results of word segmentation including phrase, POS
                 dependency relation, governing word, opinion, feature, type
             P  –  Basic product structure
//Output: PCP – Parent –children pairs
          PCP=Ø
          For each tagged review rn ∈ R
              PCP=Ø
                  For i=1 to end of review rn
                     If i<Length(rn) – 2 then x=3
                     Else If i=length(rn) – 2 then x=2
                          Else If i=length(rn) – 1 then x=1
                               Else x=0
                               End
                          End
                     End
                     For j = 1 to x
                        GW = phrase(i+j)     /* Potential parent node namely phrase(i+j) of rn */
                        GT  = T(i+j)         /* POS Tag of phrase(i+j) of rn */
                        If GT is a nous and the dependency relation between phrasei
                          and phrase(i+j) is a ATT
                          and the type of phrase(i+j) is Product or Parts

                          then
                          i=i+j
                          PCP=PCP ∨ PCPi
                          Break
                        End
                     End
                  End
              Number (PCP)++
          End
```

Figure 12. Pseudo-code of finding potential parent node

6.2 Similarity between Potential Parent Nodes and the Nodes of Basic Product Structure

It is necessary to confirm whether a potential parent node of the current attributed product feature exists at basic product structure or not before adding the attributed product features into basic product structure. Comparing the similarity between the potential parent nodes and the nodes of basic product structure is a valid measure. Considering the characters of Chinese language, the similarity between the potential parent nodes and the nodes of basic product structure is calculated from two aspects: literal similarity and context similarity.

6.2.1 Literal Similarity

Word is the basic unit of constructing a phrase. For Chinese language, many phrases whose meanings are similar always contain the same words (Xia, 2007). Based on these facts, the similarity between potential parent nodes and the nodes of basic product structure can be calculated through the status of words appearing at these nodes (product features) namely literal similarity which is influenced by two factors: quantitative and position (Wang, Zhou & Sun, 2012).

Let pf_A and pf_B are two product features that the similarity between them need to be calculated. The literal similarity $LitSim(pf_A, pf_B)$ between pf_A and pf_B is generalized as follows (Xia, 2007; Wang *et al*., 2012):

$$LitSim(pf_A, pf_B) = \alpha \times (\frac{|SameHZ(pf_A, pf_B)|}{|pf_A|} + \frac{|SameHZ(pf_A, pf_B)|}{|pf_B|})/2$$

$$+\beta \times d_p \times (\frac{\sum_{i=1}^{|pf_A|} Weight(pf_A, i)}{\sum_{i=1}^{|pf_A|} i} + \frac{\sum_{j=1}^{|pf_B|} Weight(pf_B, j)}{\sum_{j=1}^{|pf_B|} j})/2 \quad (12)$$

and $0 \leq LitSim(pf_A, pf_B) \leq 1$.

where α and β are the weights that describe the importance of quantitative factor at the literal similarity calculation and the importance of position factor at the literal similarity calculation respectively, and $\alpha + \beta = 1$. In addition, d_p defines the ratio of the number of words at these two product features.

$$d_p = \min\{\frac{|pf_A|}{|pf_B|}, \frac{|pf_B|}{|pf_A|}\}$$

$Weight(pf_A, i)$ denotes the weight of the ith word of the product feature pf_A.

$$Weight(pf_A, i) = \begin{cases} i, & \text{if } pf_A(i) \text{ at } SameHZ(pf_A, pf_B) \\ 0, & others \end{cases}$$

where $|pf_A|$ and $|pf_B|$ denote the number of words at product feature pf_A and product feature pf_B, respectively. $pf_A(i)$ denotes the i th word of product feature pf_A. $SameHZ(pf_A, pf_B)$ denotes the set of the words that are contained in both product feature pf_A and product feature pf_B at the same time. $|SameHZ(pf_A, pf_B)|$ is the number of the set $SameHZ(pf_A, pf_B)$.

6.2.2 Context Similarity

In addition, some Chinese phrases are similar at sematic but they don’t contain any the same words such as “外观(appearance)” and “样子(shape)”. In order to calculate the similarity of these kinds of product features, it should make full use of the context information around these product features because the phrases which modify the same sematic phrases are always similar (Tu, Zhang, Zhou & He, 2012). Thus, the similarity calculation between product features based on context can be generalized as follows:

Each product feature pf_i is described as a n dimensional vector.

$$pf_i = (S_{i1}, S_{i2}, \cdots, S_{ij}, \cdots, S_{in}) \tag{13}$$

where S_{ij} is the co-occurrence frequency between product feature pf_i and the jth modified phrase.

Thereupon, the similarity calculation among product features is transformed into the similarity between two vectors. It is generalized as follows (Tu *et al.*, 2012):

$$Sim(pf_a, pf_b) = \frac{\sum_{k=1}^{n} S_{ak} \times S_{bk}}{\sqrt{\sum_{k=1}^{n} S_{ak}^2 \sum_{k=1}^{2} S_{bk}^2}} \tag{14}$$

where S_{ak} denotes the co-occurrence frequency between product feature pf_a and the kth modification phrase. S_{bk} denotes the occurrence frequency of product feature pf_b and the kth modification phrase. n is the total number of the modification phrases in an existing group. And $Sim(pf_a, pf_b)$ is the similarity between product feature pf_a and product feature pf_b.

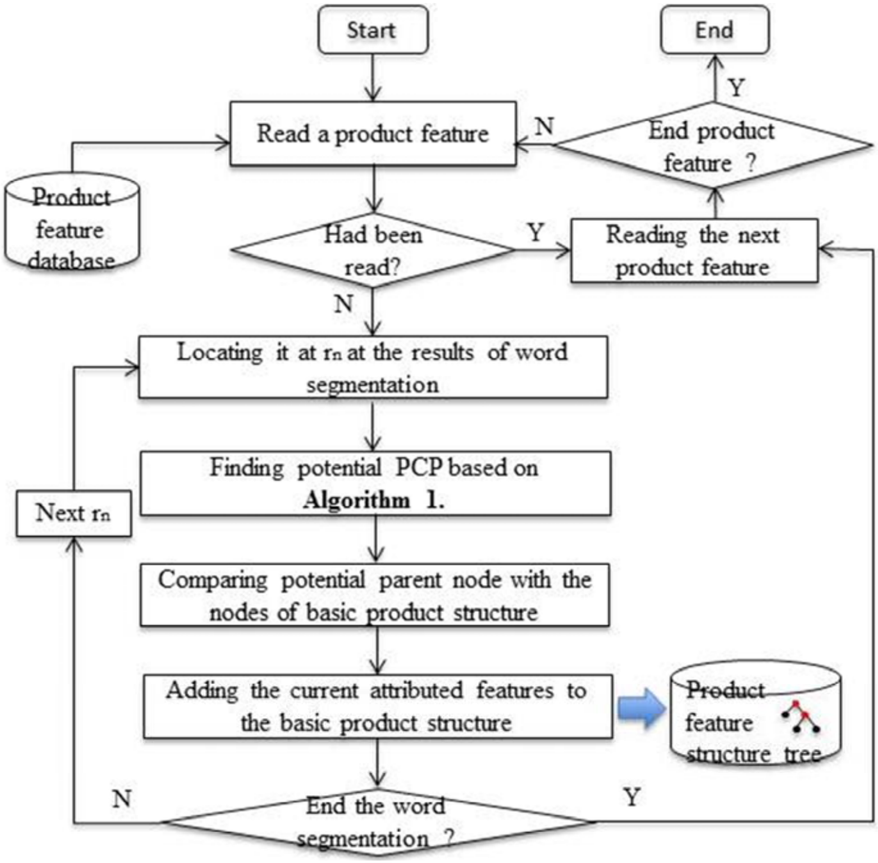


Figure 13. Process of constructing product feature structure tree

Based on potential parent node searching and similarity calculating, the process of constructing product feature structure tree is presented in **Figure 13**. First, picking out a product feature from product feature database, and locating it at the results of word segmentation e.g. the r_nth consumer reviews. Second, searching the potential parent-child pairs (PCP) by calling **Algorithm 1**, and then comparing the parent nodes of potential PCP with the nodes of basic product structure based on similarity analysis. If exists, adding the product features (namely attributed product feature) into the corresponding nodes of basic product structure as its children. Repeating this process, until all the attributed product features are added into basic product structure. This process connects not only the attributed product features but also their quantitative descriptions such as frequency and sentiment score with their parent nodes.

7. Experimental Analysis

Product feature extraction from Chinese consumer reviews is a complicated task and is also a crucial task because its results influence the efficiency of similarity analysis and comprehensive applications directly.

Many factors influence the results of product feature extraction. In order to insight into these factors and provide evidences to control the process of product feature extraction effectively, we design extended experiments from different perspectives based on 5,806 Chinese consumer reviews retrieved from e-commerce platforms ***Taobao.com***, ***Suning.com***, and ***Zhongguancun.com***.

7.1 Results of Word Segmentation

The results of word segmentation provide the data resources for product feature extraction and product feature structure tree constructing. Therefore, a valid word segmentation should keep enough correctness. In this work, a two-stage optimizing word segmentation process is proposed which is presented in **Figure 3**. In order to show the effectiveness and necessity of two-stage optimizing word segmentation process, we designed two experiments: the word segmentation based on tool *ictcals* only and the word segmentation based on our proposed two-stage optimizing word segmentation method. And then the correct rate, which is defined as the ratio between the number of correct word segmentation and the number of total word segmentation result, is used as index to evaluate the effectiveness of different word segmentation methods and different data sources such as ***taobao.com***, ***suning.com***, and ***zhongguancun.com***, respectively. These results are presented in **Figure 14**. Black rectangles describe the correct rates of product features that are extracted based on *ictcals* system only from ***taobao.com***, ***suning.com***, and ***zhongguancun.com*** respectively (*taobao*:90.16%, *suning*:90.5%, and *zhongguancun*:95.29%.). Red rectangles describe that correct rates of product features that are extracted based on our two-stage optimizing word segmentation from ***taobao.com***, ***suning.com***, and ***zhongguancun.com*** respectively (*taobao*:98.39%, *suning*:95.97%, and *zhongguancun*:97.65%.). Obviously, the correct ratios of red rectangle are all higher than those of black rectangle.

Furthermore, we also calculate the average correct rate of word segmentation based on the total data from ***taobao.com***, ***suning.com***, and ***zhongguancun.com*** which is illustrated in **Figure 15**. The correct rate is also increased by 6.16%. Therefore, it is very necessary to implement two-stages optimizing word segmentation in order to increase the correctness of Chinese consumer reviews and provide valid data sources for product feature extraction.

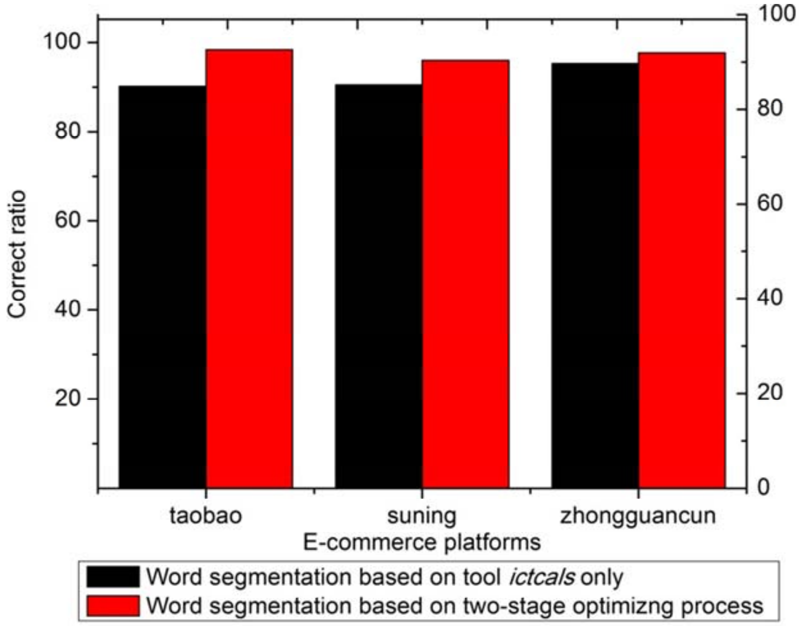


Figure 14. Correct rates of two word segmentation methods for three data sources

7.2 Contents of Rule Template

The elements of rule template and its organization form determine the solution of extracting product features. Different rule templates will lead to different effectiveness of product feature extraction. In order to explore a valid rule template including elements and its organization form for our work, 10 rule templates that are developed based on different elements which are presented at **Table 2** and **Figure 7** and organization forms are designed. The efficiency of product feature extraction based on these 10 rule templates are evaluated respectively based on existing popular index such as precision, recall, and *F*-score which is illustrated in **Figure 16.**

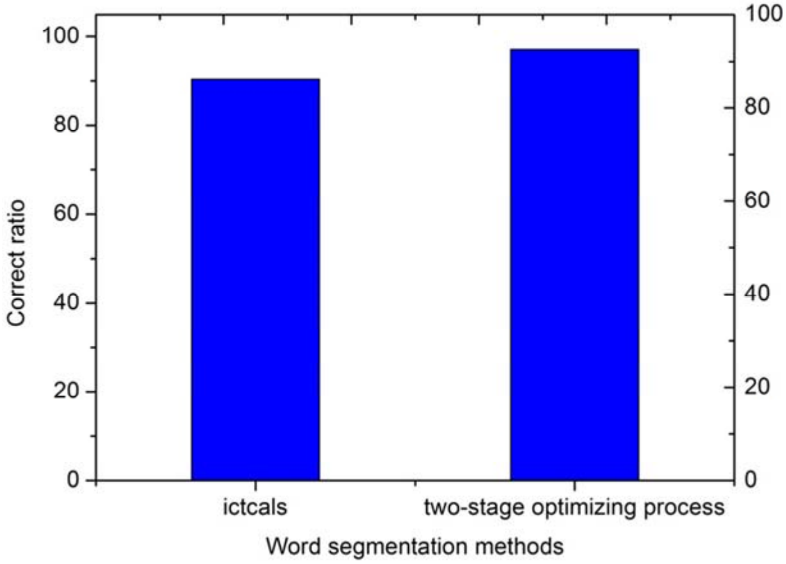


Figure 15. Correct rates of two word segmentation methods for total data

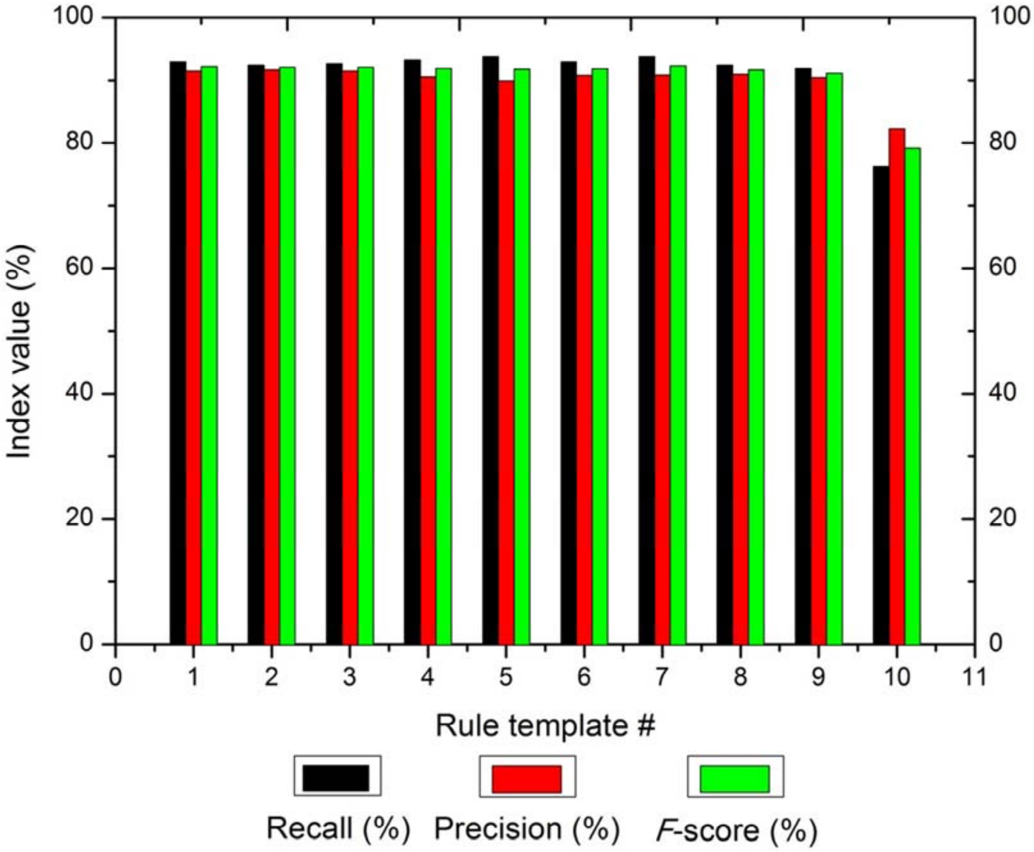


Figure 16. Precisions, recall, and F-score of product feature extraction based on different rule templates

We found that the precision, recall, and *F*-score corresponding to the 7th rule template are 90.86%, 93.8%, and 92.31%, respectively. These are comprehensive optimal comparing with those of product feature extraction processes based on the other 9 rule templates. Thus, the rule template for CRF in this work will be established according to the 7th rule template.

7.3 Widths of Searching Window

Consumer reviews are always irregular expression because the purpose of consumer commenting on products at network platform is to exchange and share information. Especially for Chinese language, its complex syntax, grammar and diversified expressions make it more serious. Therefore, a proper search range is very important in order to find the valid phrases which are correlated with the current object.

With these considerations, three widths of searching window which had been described in **Figure 11** are designed such as 3, 5, and 7 respectively to extract the potential parent nodes for current product features. We also employ precision, recall and *F*-score to measure the effectiveness of finding potential parent node at different widths of searching window, and the results of them are presented in **Figure 17**. It can be seen that the comprehensive result is the optimal when the width of searching window is 5 although the recall of it increases continuously along with the increasing of width. The precision and *F*-score will be decreased once expanding the width of searching window when the potential parent node cannot be found at given range. The reason is that the phrases that are found at expanding range maybe satisfy with the constraint conditions defined at our searching algorithms such as POS or rules, it may not correlate with the current product feature at all. Thus, it decreases the precision and the *F*-score in the end.

Considering the expression habits of Chinese language and the irregularity of consumer review, the potential parent nodes of the current product features are always omitted or implicated. Therefore, they cannot be found directly under these conditions. In order to deal with this issue, a workflow of identifying the potential parent nodes for this kind of product features is presented in **Figure 18**. It is to infer the potential parent node for the current product feature according to the existing searching results namely the potential parent nodes for the same product feature at the front of consumer reviews. If the infer results are null, then the design manual for the target product which records the correlations between components/parts and their attributes is used as evidences to identify its parent node. It avoids to searching at wider range aimlessly and keep the effectiveness of searching process as well.

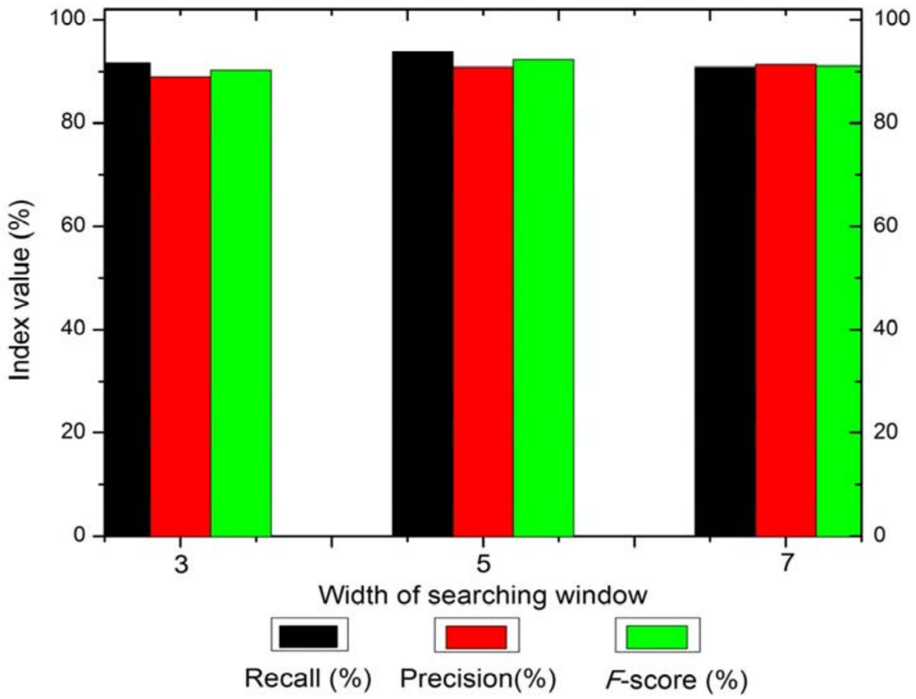


Figure 17. Precision, recall, and F-score under different widths of searching window

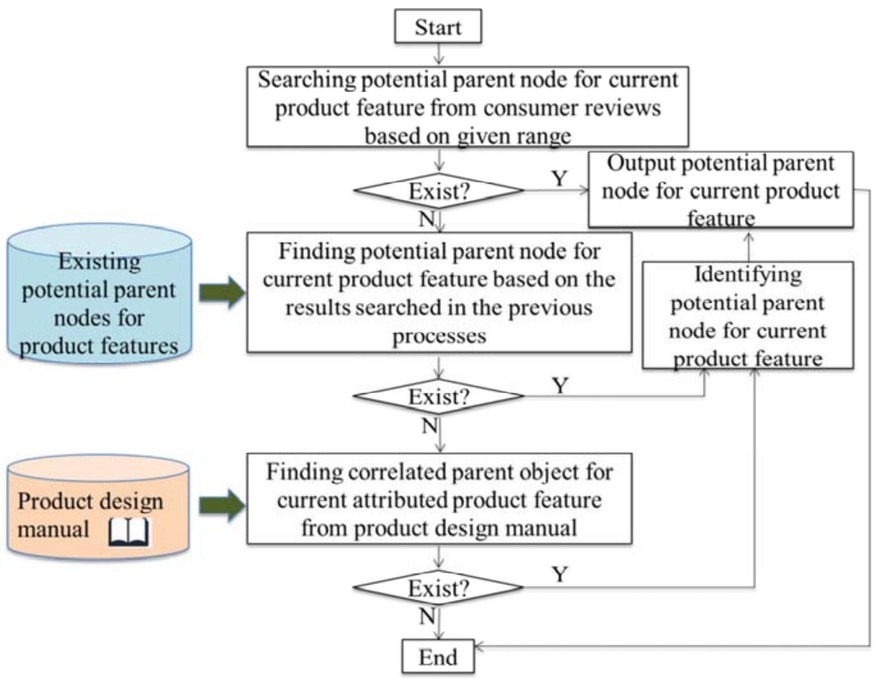
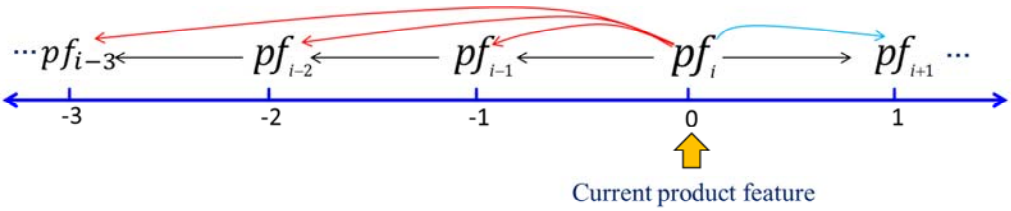
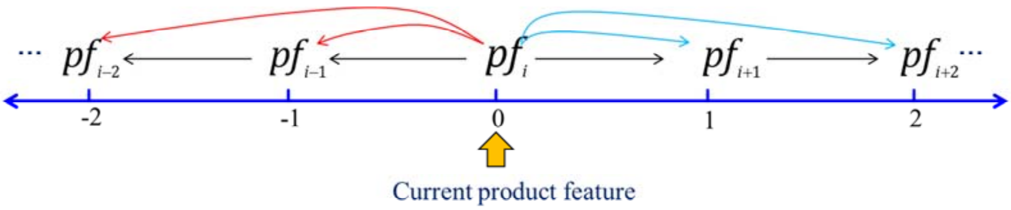


Figure 18. Workflow of identifying the implicit parent nodes for some product features

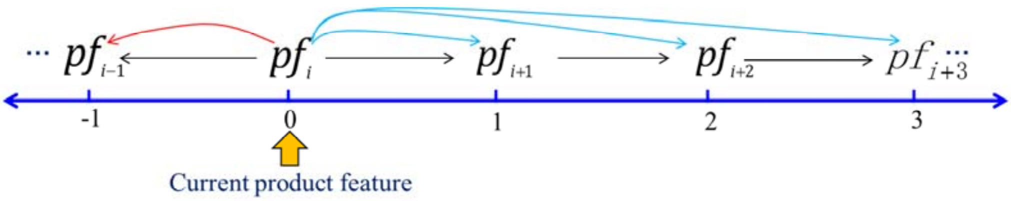
Moreover, the exact coverage regions of searching window may also be different even for the same width of searching window. Using the width 5 of searching window as example, three forms of coverage regions are presented in **Figure 19**. Accordingly, the efficiencies of searching potential parent nodes are evaluated through precision, recall, and *F*-score which are presented in **Figure 20**. The form of coverage region in **Figure (19-2)** corresponds to a better result. Therefore, the practical searching range and its coverage region are set based on this result in the case study.



(19-1) Searching range [-3, 1]



(19-2) Searching range [-2, 2]



(19-3) Searching range [-1, 3]

Figure 19. Different coverage forms of searching window

These experiments and their results provide the evidences for our research works at word segmentation, product feature extraction, and product feature structure tree constructing. They are very significant for keeping the validity of our proposed methods.

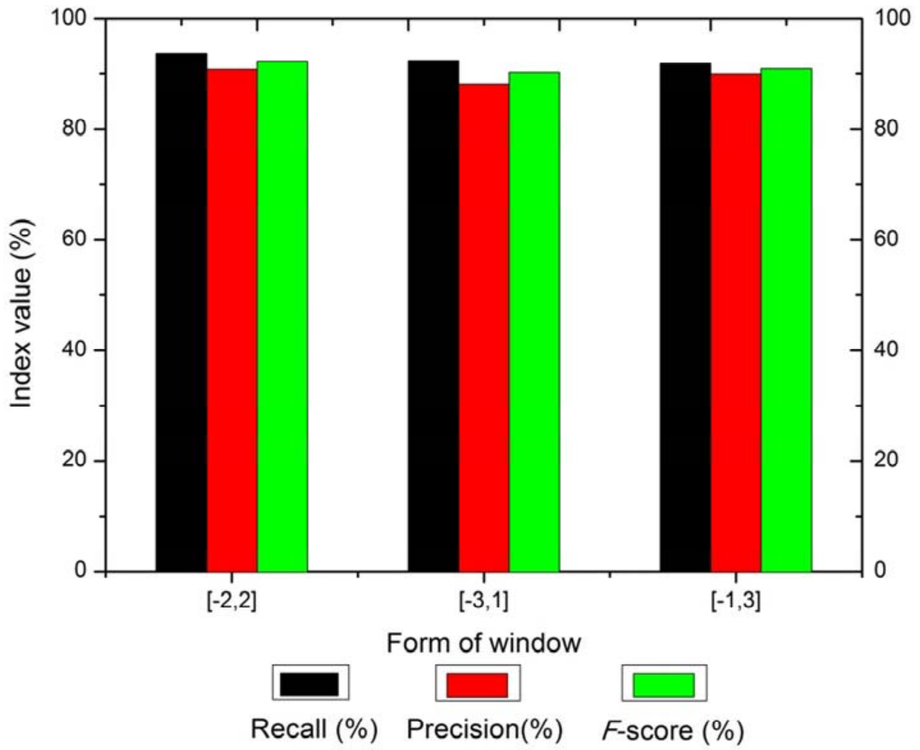


Figure 20. Efficiency of searching potential parent nodes under different forms of coverage regions

8. Case Study

Consumer reviews contain rich information regarding consumer requirements and preferences. Mining valuable information effectively from consumer reviews can provide evidences for designers, manufacturers, or retailers to implement product improvement or make market strategy. With the rapid expansion of e-commerce businesses based on network platforms and clients, more and more companies have realized the importance of this kinds of utilities. Aiming at Chinese consumer reviews, this section, using intelligent mobile phone xx-F2 as example, is to elaborate the implementations of the principles and methods mentioned above, and the applications based on product features.

By using web crawler tools *Goseeker* and *Train collector*, we retrieved 5,806 Chinese consumer reviews from e-commerce platform ***taobao.com*** (2,591), ***suning.com*** (1,243), and ***zhongguancun.com*** (1,972) which are used as analysis corpuses. According to the technique framework presented in **Figure 1**, we employ software *ictclas*, which is developed by Chinese Science Academic, as word segmentation tool to divide consumer reviews into discrete phrases and label their POS. At the same time, we employ software *ltp*, which is developed by

Harbin Institute of Technology of China to achieve syntactic parsing. And 82,724 raw phrases are obtained. After preprocessing for these raw phrases such as stop words, typos, and meaningless phrases, a two-stages optimization word segmentation process is performed presented at **Section 3.2** to make the results of word segmentation more suitable for our research tasks, and the key parameters of these optimization phases are set presented in **Table 3**. Finally, 50,785 valid phrases are obtained. These phrases are used as the data resources (corpus) for product feature extraction.

Table 3. Parameter setting of two-stages optimizing word segmentation

No	Parameters	Explains	Setting
1	n	The length of reconstructed string	n=5
2	f	The parameter of frequency filtering	f>2
3	c	The parameter of cohesive filtering	c>0.2
4	r	The parameter of left and right entropy filtering	r>0.8
5	q	The number of relation Fi-sematic-Fj occurring at the reviews	q>3

In order to extract product features from these results of word segmentation effectively based on CRF, 9,081 phrases obtained from 1,000 consumer reviews are used as train set. We invited 2 engineers from mobile phone development department and 1 linguist from the literature of our school to annotate these phrases manually including feature, type, and opinion. It took two days, 8 hours per day to implement this task. At the same time, rule template is developed according to the analysis results of experiment at **Section 7.2**.

Based on train set and rule template, the *model* of CRF is trained through a machine learning process. And then, product features for xx-F2 product are extracted from 50,785 valid phrases of 5,806 consumer reviews based on CRF. Finally, 80 product features are obtained after merging synonym, homoionym, and alternative names.

Product feature extraction is a very crucial step for ensuring the effectiveness of the next comprehensive analysis and application based on product features. In order to verify the validity of our proposed methods, we design a five-fold intersection experiments by using 5,000 phrases from the results of word segmentation. These 5,000 phrases are divided into 5 subsets which are labeled 1, 2, 3, 4, and 5 respectively, and each subset contains 1,000 phrases. The effectiveness of product feature extraction based on five-fold intersection experiments are measured through indexes precision, recall, F-score. At the same time, we calculated the precision, recall, and F-score of product feature extraction based on the methods proposed by Jakob's work, which is the closest with our works at the aspect of product feature extraction, by using the same phrase set. Finally, we compare the results obtained based on our methods with those obtained based on the methods of Jakob's work (Jakob & Gurevych, 2010) which are presented in **Table 4**. Obviously, the precision, recall, and F-score of our methods are all better than those of Jakon's work. It denotes that our methods of extracting product features from consumer reviews are valid, especially for Chinese consumer reviews.

Table 4. Experiment result comparision between our methods and Jakob's work

Precision(%)		Recall(%)		*F*-score(%)	
Our methods	Jakob's work	Our methods	Jakob's work	Our methods	Jakob's work
93.80	86.47	90.86	78.70	92.31	79.63

Based on the product features extracted above and the results of word segmentation, the frequency of each product feature is calculated, so does the sentiment score of it. And the potential parent nodes of the attributed product features are identified based on the **Algorithm 1** presented in **Figure 12** and the workflow presented in **Figure 18**. As a result, the attributed product features are added into the basic product structure of product xx-F2. Thus, product feature structure tree for xx-F2 is established which is illustrated in **Figure 21**. The unit of product feature structure tree is a four tuple: < F_i, frequency, score, F_j > where F_i is parent node and F_j is child node. Frequency denotes the times of product feature F_j appearing at consumer reviews. Score denotes the sentiment evaluation of consumer to product feature F_j. Based on the product feature structure tree and the data on it including frequency and sentiment score, the influence or interaction relations between the parent nodes of product feature structure tree and its child nodes can be inferred conveniently.

In this work, a Bayes theory based application is investigated based on product feature structure tree that is to infer the factors (namely child nodes) of leading to the negative valuations or low sentiment scores of their parent nodes. The mathematic description of this inferring process is as following:

For a Bayes network which is concerned on a set of variables $\boldsymbol{X} = \{\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_n}\}$, it contains two aspects: ① network structure **S** in which variables **X** are conditional independency, and ② local probability distribution $\boldsymbol{P}$ which connects with each variable. Let variable $\boldsymbol{X_i}$ corresponds to a node of Bayes network, and $\boldsymbol{X_j}$ is the parent node of variable $\boldsymbol{X_i}$, then the probability of child node leading to the low sentiment score of its parent node can be generalized as following.

$$\mathbf{P}(\boldsymbol{X_i} = \boldsymbol{L} | \boldsymbol{X_j} = \boldsymbol{N}) = \frac{\boldsymbol{P}(\boldsymbol{X_j} = \boldsymbol{N} | \boldsymbol{X_i} = \boldsymbol{L}) \boldsymbol{P}(\boldsymbol{X_i} = \boldsymbol{L})}{\boldsymbol{P}(\boldsymbol{X_j} = \boldsymbol{N})} \tag{15}$$

Where $\boldsymbol{P}(\boldsymbol{X_i} = \boldsymbol{L})$ is the ratio of unsatisfied consumer reviews (L) for product feature $\boldsymbol{X_i}$ relative to all consumer reviews. It is calculated as following.

$$\boldsymbol{P(X_i = L) = \frac{\sum_{s=1}^{S} m(s,X_i,L)}{\sum_{s=1}^{S} n(s,X_i)}} \tag{16}$$

Where $\boldsymbol{n(s, X_i)}$ denotes the times of product feature $\boldsymbol{X_i}$ appearing on the sth consumer review. $\boldsymbol{m(s,\ X_i, L)}$ denotes the times of product feature $\boldsymbol{X_i}$ appearing on the sth consumer review that has negative evaluating on the product feature $\boldsymbol{X_i}$.

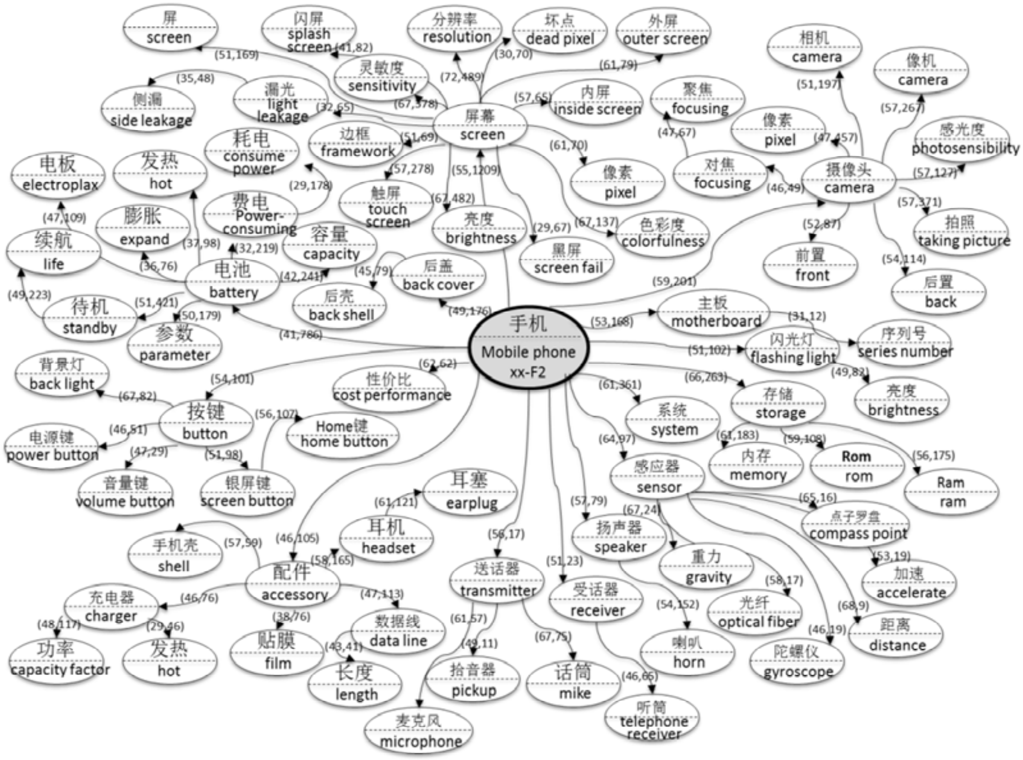


Figure 21. Product feature structure tree for product xx-F2

And, $\boldsymbol{P(X_j = N|X_i = L)}$ denotes the probability that a child node (product feature) $\boldsymbol{X_i}$ being evaluated as negative (described as L) leads to its parent node (product feature) $\boldsymbol{X_j}$ being evaluated as a poor feature (described as N) by consumers. Thus, $\boldsymbol{P(X_j = N|X_i = L)}$ can be generalized as following.

$$\boldsymbol{P(X_j = N|X_i = L) = \sum P\left(X_j = N\middle|X_i, \cdots, X_k\right)P(X_i = \Gamma)_{\Gamma=L,M,H}, \cdots, P(X_k = \Gamma)_{\Gamma=L,M,H}} \tag{17}$$

Where $\boldsymbol{P\left(X_j = N\middle|X_i, \cdots, X_k\right)P(X_i = \Gamma)}$ denotes the probability that the child nodes (product features) $\boldsymbol{\{X_i, \cdots, X_k\}}$ being evaluated as positive (described as H), negative (described as L), and neutral (described as M) lead to its parent node (product feature) $\boldsymbol{X_j}$ being evaluated as a poor feature (described as N) by consumers. The probability sum of these child nodes being evaluated as positive (H), negative (L), and neutral (M) respectively denotes the probability of parent node (product feature) $\boldsymbol{X_j}$ being evaluated as a poor feature (described as N) when the child node (product feature) $\boldsymbol{X_i}$ is evaluated as negative (described as L) namely $\boldsymbol{P(X_j = N | X_i = L)}$.

Using substructure 送话器(transmitter), which has a relative low sentiment score according to our statistical results and consists of three child nodes such as 麦克风(microphone), 拾音器(pickup) and 话筒(mike), as example, the correlation matrix between the child node evaluations and the parent node evaluations from consumers is established by experts based on their observations on 3,000 consumer reviews which is presented **Table 5**. On the basis of this, the influences between parent nodes and their child nodes can be calculated based on formulas (15)-(17). The results shown that the probabilities of a relative low sentiment scores of 送话器(transmitter) causing by its child nodes such as 麦克风(microphone), 拾音器(pickup) and 话筒(mike) are 0.415, 0.327, and 0.258, respectively. It can be seen that this relative low sentiment score of 送话器(transmitter) is the most likely caused by 麦克风(microphone). Thus, designers or manufacturers should improve the 麦克风(microphone) for the future in order to increase the satisfactions of consumers for their products, and gain profit margins under fierce market competition in the end.

Table 5. Observation results of influence among product features

Child nodes (product features)			Parent nodes (product features)	
麦克风 (microphone)	拾音器 (pickup)	话筒 (mike)	送话器（Transmitter）	
			Y	N
L	L	L	0.083	0.917
L	L	M	0.143	0.857
L	L	H	0.417	0.583
L	M	L	0.354	0.646
L	M	M	0.703	0.297

Similarly, the influences among other nodes on the product feature structure tree can also be analyzed in this way. It will provide valuable evidences for the designers, manufacturers, or retailers.

9. Discussions

The main goal of Online reviews from consumers is to exchange or share information among them. The languages from consumers are characterized as oral, haphazard, and irregular syntax. And some new words or terms are also created or introduced continuously, specifically for young people. Therefore, it is necessary to adopt two-stages optimization method for word segmentation. This process can deal with the error results of direct word segmentation first, and find some new words or terms. For example, "分辨率(pixel)", in fact, is a kind of attribute descriptions of intelligent electronic products. So "分", "分辨", and "辨率" are all error results of word segmentation but these results exactly exist in practice. Obviously, it is necessary to delete these error phrases from the results of original word segmentation process in order to keep the accuracy of our research and analysis works. Based on the results of original word segmentation, the correct form namely "分辨率(pixel)" can only be generated through word reorganization. However, new error forms can also be generated such as "*分", "分辨", and "率*", etc. Through three filter algorithms such as frequency filtering, cohesive filtering, and left & right entropy filtering, most of these error results can be deleted from the results of original word segmentation. In addition, some new terms or phrases can also be found such as "云存储(cloud storage)" and "语音识别(speech recognition)", etc. All these new words and terms, along with the correct results of word segmentation, are input into user dictionary again which is used to guide word segmentation at practice. And then, the process of word segmentation will be restarted based on this extended user dictionary. As a result, the correct rate of word segmentation is increased remarkably. For example, we used 1,000 consumer reviews as experiment corpus, and invited two development engineers of intelligent mobile phone and one linguist to divide reviews into phrases and annotate their POSs manually. The results are used as reference to evaluate the efficiency of word segmentation methods. And then, two kinds of word segmentation processes such as word segmentation based on *ictclas* tool directly and our proposed two-stages optimizing methods. Comparing with the reference results obtained from experts, the results generated from our two-stages optimizing method are more accuracy than those of *ictclas* tool directly which had been explained in **Figure 14** and **Figure 15**. Therefore, two-stages optimizing word segmentation method for Chinese consumer reviews is valid and necessary. It ensures to provide high quality data for the next product feature extraction analysis and application.

Product feature extraction is a complex task especially for Chinese consumer reviews, and also a crucial stage that will influence the effectiveness of applications based on product features directly. Therefore, product feature extraction in this work adopted supervised product feature extraction strategy due to its high precision. Thus, the core work is to design a reasonable rule template. Besides the elements of existing traditional rule templates, the rule template developed in this work added two kinds of elements such as governing word and opinion word to support product feature extraction and sentiment identification. By doing these, some implicit product features or sentiment expresses can be detected by combining these new adding elements with the existing elements of existing rule templates which were presented in **Figure 8**. For example, "杠杠的(ganggangde means very good)" is a recent popular express which describes a kind of positive evaluation. It is an opinion word but it isn't contained at user dictionary exactly. Thereupon, we added it into extended user dictionary, and annotated it as opinion word manually at train set. And then, the implicit product features concerned with it can be extracted conveniently, and their sentiment score can be calculated accurately. In addition, “战斗机(fighter)” is another popular express recently. In essence, it is a noun phrase. But it is always used as an adjective phrase to modify a product feature around it and express a positive sentiment. Likeness, this phrase is also not contained at user dictionary. Therefore, it is high significant for product feature extraction from Chinese consumer reviews to find new words especially for opinion words to extend existing user dictionary through two-stages optimizing word segmentation process, and annotate the opinion attributes of phrases at train set and rule template. After doing this, the implicit product features and their sentiment evaluation can be processed accurately. These were verified in **Figure 16** which presents the efficiency of product feature extraction based on 10 different rule templates, and the 7^{th} rule template which was proposed in this work has better results than those of rule templates.

In addition, product features are always internal correlated with each other. For example, “摄像头(camera)” and “像素(pixel)” are two product features, and may appear at different consumer reviews discretely. However, product feature “像素(pixel)” is one of the attributes of product feature “摄像头(camera)” in essence. Therefore, the internal correlation among them is an inevitable existence. Unfortunately, the existing researches don’t explore this fact. This paper discussed this issue. Product feature structure tree is the representation form of the internal correlations among product features. It integrates product features which distributes at consumer reviews concretely into a whole object, and makes the comprehensive applications based on product features feasible. However, the numbers between parent nodes (product features) and its child nodes (product features), according to our observations, don’t satisfy with cumulative calculation law both frequency and sentiment score e.g. between “送话器(transmitter)” and {“麦克风(microphone)”, “拾音器(pickup)”, and “话筒(mike)”}. The reason is that many consumers provide a snippet text description for products only for the goal of completing evaluation task required by platform or system. As a result, many product features at consumer reviews are not evaluated by consumers at all. Therefore, the influences among product features cannot be reflected by the numbers on product feature structure tree directly. For this reason, a method of inferring the influences among product features based on product feature structure tree is proposed by using Bayes theory. This method uses the sentiment scores of product features as evidences to identify the product features that need to be analyzed in depth because of its low or negative evaluations from consumers. At the same time, it makes full use of the practical evaluation results of each review from consumers. Therefore, the inferring results are more convince. For example, product feature “送话器(transmitter)” is determined as the object that need to be inferred the elements leading to its low or negative sentiment score. According to the data on product feature structure tree, child node (product feature) ”拾音器(pickup)” may be the potential element because of its lowest sentiment score. However, the inferring result from our proposed method based on Bayes theory is that child node “麦克风(microphone)” has the maximal possible of leading to low sentiment score of its parent node or negative evaluation. It is in accordance with fact. Even if the sentiment scores of “麦克风(microphone)” are not the lowest while the frequency of product feature received negative evaluation are very high which means a large amount of consumers pay attention on this product feature and give negative evaluation on this product feature. Therefore, this leads to a lower sentiment score of its parent nodes. From the perspective of probability theory and mathematical statistics, a minority events always have no statistic means in general. Therefore, product feature structure tree makes the research and analysis on the internal relations among product features feasible, and the inferring method based on Bayes theory is a valid method to keep the applications more reasonable.

10. Conclusions

A large amount of product reviews provide valuable consumer feedback. In the past decade, many researchers in computer science and information management have paid much attention to extract product features from consumer reviews, and analyze the opinion direction of consumer for product features. This paper, aiming at Chinese consumer reviews, investigates the issues of product feature extraction and the applications at product feature level. It is high significant because of emerging a huge e-commerce market in China.

In this work, a technique overview of extracting product features from Chinese consumer reviews is proposed in which two-stages optimizing word segmentation, product feature extraction based on CRF, and product feature structure tree establishing are investigated. Two-stages optimizing word segmentation process mainly consists of phrase reconstructing, frequency filtering, cohesiveness filtering, and left & right entropy filtering. It increases the correct rate of word segmentation through new phrase finding to expand user dictionary and the second word segmentation process. Likeness, an expanded rule template is proposed in which governing word and opinion word annotating are added to detect the implicit product features and infrequent opinion words. It increases both the efficiency of product feature extraction from Chinese consumer reviews and the accurateness of sentiment evaluation for product features. At the same time, two quantitative characters are defined to describe the preference extent of consumers for a product feature. Furthermore, product feature structure tree is established based on the inevitable internal correlations among product features. An algorithm is proposed to find the potential parent nodes for current product features from the results of word segmentation and different similarity functions are employed to evaluate the similarity between the potential parent nodes and the nodes of basic product structure in order to add the attribute product features into basic product structure. On the basis of these, an inferring application based on product feature structure tree is explored to identify the potential factors that lead to the low sentiment score of its parent node by using Bayes theory. This is high significant for designers, manufacturers, or retailers to implement product update, quality improvement, and market strategy, etc. Moreover, categories of comparative experiments and profound analysis are conducted on 5,806 real consumer reviews. The results generated from them provide the evidences for our research works. Finally, the case study verified the effectiveness of our proposed methods and applications.

Potential research work can be extended in many directions such as product quality and risk management and the dynamic evolution characteristics of the influences among product features, etc. These are also our future research directions.

Acknowledgements

This project was supported by the Natural Science Foundation of China (NSFC) under contract 51405462 and 51175486. Zhejiang Province Fund of Natural Science, China under contract LY16G010006 and LQ15G010005.

Reference

Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of 20th international conference on very large data bases(VLDB '94)*, 487-499.

Bahu, S. M. & Das, S. N. (2015). An Unsupervised Approach for Feature Based Sentiment Analysis of Product Reviews. *International Journal of Scientific Research Engineering & Technology*, *4*(5), 484-489.

Chang, Y. C., Chu, C. H., Chen, C. C. & Hsu, W. L. (2016). Linguistic Template Extraction for Recognizing Reader-Emotion. *International Journal of Computational Linguistics and Chinese Language Processing*, *21*(1), 29-50.

Chen, L., Qi, L. L. & Wang, F. (2012). Comparison of feature-level learning methods for mining online consumer reviews. *Expert System with Applications*, *39*(10), 9588-9601. doi: 10.1016/j.eswa.2012.02.158

Choi, Y. & Cardie, C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing(EMNLP '09)*, *2*, 590-598.

Dai, H. J., Tsai, R. T. H. & Hsu, W. L. (2014). Joint Learning of Entity Linking Constraints Using a Markov-Logic Network. *International Journal of Computational Linguistics and Chinese Language Processing*, *19*(1), 11-32.

Dave, K., Lawrence, S. & Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product review. In *Proceedings of the 12th* *international conference on World Wide Web (WWW 2003)*, 519-528. doi: 10.1145/775152.775226

Dellarocas, C. (2003). The digitization of word of mouth: promise and challenges of online feedback mechanisms. *Management Science*, *49*(10), 1407-1424. doi: 10.1287/mnsc.49.10.1407.17308

Duan, W., Gu, B. & Whinston, A. B. (2008). Do online reviews matter? An empirical investigation of panel data. *Decision Support System*, *45*(4), 1007-1016. doi: 10.1016/j.dss.2008.04.001

Etzioni, O., Cafarella, M., Doweny, D., Popescu, A.-M., Shaked, T., Soderland, S.…Yates, A. (2005). Unsupervised name-entity extraction from the Web: an experimental study. *Artificial Intelligence*, *165*(1), 91-134. doi: 10.1016/j.artint.2005.03.001

Forman, C., Ghose, A. & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: the role of reviewer identity discloser in electronic markets. *Information System Research*, *19*(3), 291-313. doi: 10.1287/isre.1080.0193

Godes, D. & Mayzlin, D. (2004). Using online conversations to study word of mouth communication. *Marketing Science, 23*(4), 545-560. doi:10.1287/mksc.1040.0071

Htay, S. S. & Lynn, K. T. (2013). Extracting Product Features and Opinion Words Using Pattern Knowledge in Customer Reviews. *The Scientific World Journal*, Article ID 394758. doi: 10.1155/2013/394758

Hu, M. & Liu, B. (2004a). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining*, 168-177. doi: 10.1145/1014052.1014073

Hu, M. & Liu, B. (2004b). Mining opinion features in customer reviews. In *Proceedings of the 19th national conference on Artifical intelligence*, 755-760.

Hu, Z. K., Zheng, X. L., Wu, Y. F. & Chen, D.-r. (2013). Product recommendation algorithm based on users' reviews mining. *Journal of Zhejiang University (Engineering Science)*, *47*(8), 1475-1485. [In Chinese]

Jakob, N. & Gurevych, I., (2010). Extracting opinion targets in a single- and cross- domain setting with conditional random fields. In *Proceedings of the the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP' 10)*, 1035-1045.

Jiang, M. T.-J., Shih, C.-W., Yang, T.-H., Kuo, C.-H., Tsai, R. T.-H. & Hsu, W.-L. (2012). Enhancement of Feature Engineering for Conditional Random Field Learning in Chinese Word Segmentation Using Unlabeled Data. *International Journal of Computational Linguistics & Chinese Language Processing*, *17*(3), 45-86.

Jindal, N. & Liu, B. (2006). Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, 244-251. doi: 10.1145/1148170.1148215

Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K. & Fukushima, T. (2004). Collecting evaluative expressions for opinion extraction. In *Proceedings of the first international joint conference on natural language processing (IJCNLP-04)*, 596-605.

Kobayashi, N., Iida, R., Inui, K. & Matsumotto, Y. (2005). Opinion extraction using a learning-based anaphora resolution technique. In *Proceedings of the second international joint conference on natural language processing (IJCNLP-04)*, 173-178.

Lafferty, J. D., McCallum, A. & Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning(ICML 01)*, 282-289.

Li, F. T., Han, C., Huang, M. L., Zhu, X., Xia, Y.-J., Zhang, S. & Yu, H. (2010). Structure aware review mining and summarization. In *Proceedings of the 23rd International Conference on computational Linguistics*, 653-661.

Li, S., Ye, Q., Li, Y. J. & Law, R. (2009). Mining features of products from Chinese customer online reviews. *Journal of Management Sciences in China*, *12*(2), 142-152. [In Chinese]

Li, Z. H. (2013). *Research on Key Technologies of Chinese Dependency Parsing* (Doctoral dissertation, Harbin Institute of Technology). Retrieved from http://hlt.suda.edu.cn/~zhli/papers/zhenghua-2013-phd-thesis.pdf. [In Chinese]

Liu, B., Hu, M. & Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the Web. In *Proceedings of 2005 World Wide Web conference(WWW 05)*, 342-351. doi: 10.1145/1060745.1060797

Liu, D. Y. & Wang, L. F. (2013). Keywords extraction algorithm based on semantic dictionary and lexical chain. *Journal of Zhejiang University of Technology*, *41*(5), 545-551. [In Chinese]

Liu, L. Z., Song, W., Wang, H. S., Li, C. C. & Lu, J. L. (2014). A Novel Feature-based Method for Sentiment Analysis of Chinese Product Reviews. *China Communications*, *11*(3), 154-164. doi: 10.1109/CC.2014.6825268

Liu, T., Wu, G. & Yao, T. (2006). Opinion Searching in Multi-Product Reviews. In *Proceedings of the Sixth IEEE International Conference on Computer and Information Technology (CIT'06)*. doi: 10.1109/CIT.2006.132

Liu, T. & Ma, J. H. (2009). Theories and Methods of Chinese Automatic Syntactic Parsing. *Contemporary linguistics*, *11*(2), 100-112. [In Chinese]

Lv, P., Zhong, L., Cai, D. B. & Wu, Y. T. (2014). Effective mining product features from Chinese review based on CRF. *Computer Engineering & Science*, *36*(2), 359-366. [In Chinese]

Ma, B. Z. & Yan, Z. J. (2014). Product features extraction of online reviews based on LDA model. *Computer Integrated Manufacturing Systems*, *20*(1), 96-103. [In Chinese]

Miao, Q., Li, Q. & Zeng, D. (2010). Mining fine grained opinions by using probabilistic models and domain knowledge. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 358-363. doi: 10.1109/WI-IAT.2010.193

Ouyang, C. P., Liu, Y. B., Zhang, S. Q. & Yang, X. H. (2015). Features-level Sentiment Analysis of Movie Reviews. *Adcanced Science and Technology Letters*, *81*, 110-113. doi: 10.14257/astl.2015.81.23

Popescu, A. & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the conference on human language technology and empirical methods in natural language processing(HLT '05)*, 339-346. doi: 10.3115/1220575.1220618

Song, H., Yan, Y. & Liu, X. Q. (2012). A grammatical dependency improved CRF learning approach for integrated product extraction. In *Proceedings of 2nd International Conference on Computer Science and Network Technology(ICCSNT)*, 1787-1794. doi: 10.1109/ICCSNT.2012.6526267

Srikant, R. & Agrawal, R. (1995). Mining generalized association rules. In *Proceedings of the 21th international conference on very large data bases(VLDB '95)*, 407-419.

Tu, X. H., Zhang, H. C., Zhou, K. F. & He, T. T. (2012). Extracting Structured Information from Chinese Wipipedia and Measuring Relatedness between Words. *Journal of Chinese Information Processing*, *26*(3), 109-114. [In Chinese]

Turney, P. D. (2002). Thumbs, up or thumbs down: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*, 417-424. doi: 10.3115/1073083.1073153

Wang, H. S., Liu, L. Z., Song, W. & Lu, J. (2014). Feature-based Sentiment Analysis Approach for Product Reviews. *Journal of Software*, *9*(2), 274-279. doi:10.4304/jsw.9.2.274-279

Wang, W. & Wang, H. W. (2016). Comparative network for product competition in feature-levels through sentiment analysis. *Journal of Management Sciences in China*, *19*(9), 109-126. [In Chinese]

Wang, W. P. & Meng, C. C. (2011). Opinion Object Extraction Based on the Syntax Analysis and Dependency Analysis. *Computer System Applications*, *20*(8), 52-57. [In Chinese]

Wang, Y., Zhou, X. G. & Sun, Y. (2012). Research on Automatic Building of Word Correlation Net Based on Statistic. *Computer & Digital Engineering*, *40*(2), 15-18. [In Chinese]

Wei, C. P., Chen, Y. M., Yang, C. S. & Yang, C. C. (2010). Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. *Information Systems and e-Business Management*, *8*(2),149-167.

Wong, T.-L. & Lam, W. (2005). Hot item mining and summarization from multiple auction Web sites. In *Proceedings of the fifth IEEE international conference on data mining (ICDM'05)*, 797-800. doi: 10.1109/ICDM.2005.78

Wong, T.-L. & Lam, W. (2008). Learning to extract and summarize hot item features from multiple auction Web sites. *Knowledge and Information and System*, *14*(2), 143-160. doi: 10.1109/ICDM.2005.78

Xia, T. (2007). Study on Chinese Words Semantic Similarity Computation. *Computer Engineering*, *33*(6), 191-194. [In Chinese]

Yi, J. & Niblack, W. (2005). Sentiment Mining in WebFountain. In *Proceedings of the 21st International Conference on Data Engineering (ICDE 2005)*, 1073-1083. doi: 10.1109/ICDE.2005.132

Yin, C. X. & Peng, Q. K. (2009). Sentiment Analysis for Product Features in Chinese Reviews Based on Semantic Association. In *Proceedings of International Conference on Artificial Intelligence and Computational Intelligence 2009(AICI 09)*, 81-85. doi: 10.1109/AICI.2009.326

Zhang, H. P., Yu, Z. G., Xu, M. & Shi, Y. L. (2011). Feature-level sentiment analysis for Chinese product reviews. In *Proceedings of 3rd International Conference on Computer Research and Development(ICCRD 2011)*, 135-140. doi: 10.1109/ICCRD.2011.5764099

Zhang, S. & Li, F. (2015). Opinion Target and Polarity Extraction Based on Iterative Two-Stage CRF Model. *Journal of Chinese Information Processing*, *29*(1), 163-169. [In Chinese]

Zheng, M. J., Lei, Z. C., Liao, X. W. & Chen, G. L. (2013). Identify Sentiment-Objects from Chinese Sentences based on Cascaded Conditional Random Fields. *Journal of Chinese Information Processing*, *27*(3), 69-77. [In Chinese]

Zhou, X. J., Wan, X. J. & Xiao, J. G. (2013). Collective opinion target extraction in Chinese microblogs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1840-1850.

Zhuang, L., Feng, J. & Zhu, X. Y. (2006). Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management(CIKM '06)*, 43-50. doi: 10.1145/1183614.1183625

Zu, L. J. & Wang, W. P. (2014). Research of Extracting Product Features from Chinese Online Reviews. *Computer System Applications*, *23*(5), 196-201. [In Chinese]

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member： NT$ 1,000 (US$50.-)
group membership： NT$20,000 (US$1,000.-)
life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel.：886-2-2788-3799 ext. 1502 Fax：886-2-2788-1638

E-mail: aclclp@hp.iis.sinica.edu.tw Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#： ________________

Name： ___________________________________ Date of Birth： _______________

Country of Residence： ____________________ Province/State： ____________________

Passport No.： ____________________________ Sex: _________________________

Education(highest degree obtained)： ____

Work Experience： ____

____

Present Occupation： ____

Address： ____

____

Email Add： ____

Tel. No： ____________________________ Fax No： ____________________________

Membership Category：☐ Regular Member ☐ Life Member

Date： _____/_____/_____ （Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member ： US$ 50.- （NT$ 1,000）
Life Member ： US$500.- （NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

中華民國計算語言學學會

宗旨：

（一） 從事計算語言學之研究
（二） 推行計算語言學之應用與發展
（三） 促進國內外中文計算語言學之研究與發展
（四） 聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）

（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目

（三）收集國內外有關計算語言學知識之圖書及最新發展之資料

（四）發行有關之學術刊物，論文集及通訊

（五）研定有關計算語言學專用名稱術語及符號

（六）與國際計算語言學學術機構聯繫交流

（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會

2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員： 10,000.- （US$ 500.-）
個人會員： 1,000.- （US$ 50.-）
學生會員： 500.- （限國內學生）
團體會員： 20,000.- （US$ 1,000.-）

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

中華民國計算語言學學會
個人會員入會申請書

<table>
<tr><td>會員類別</td><td>☐終身 ☐個人 ☐學生</td><td colspan="2">會員編號</td><td colspan="2">（由本會填寫）</td></tr>
<tr><td rowspan="2">姓　　名</td><td rowspan="2"></td><td rowspan="2">性別</td><td rowspan="2"></td><td>出生日期</td><td>年　月　日</td></tr>
<tr><td>身分證號碼</td><td></td></tr>
<tr><td>現　　職</td><td></td><td colspan="2">學　　歷</td><td colspan="2"></td></tr>
<tr><td>通訊地址</td><td colspan="5">□□□</td></tr>
<tr><td>戶籍地址</td><td colspan="5">□□□</td></tr>
<tr><td>電　　話</td><td></td><td colspan="2">E-Mail</td><td colspan="2"></td></tr>
<tr><td colspan="6">申請人:　　　　　　（簽章）

中　華　民　國　　　年　　月　　日</td></tr>
</table>

審查結果:

1. 年費：

   終身會員：　10,000.-
   個人會員：　1,000.-
   學生會員：　500.-（限國內學生）
   團體會員：　20,000.-

2. 連絡處：

   地址：台北市南港區研究院路二段128號 中研院資訊所(轉)
   電話：(02) 2788-3799　ext.1502 傳真：(02) 2788-1638
   E-mail：aclclp@hp.iis.sinica.edu.tw　　網址: http://www.aclclp.org.tw
   連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)
PAYMENT FORM

Name: ________________________________(Please print)   Date: ______________

Please debit my credit card as follows: US$ ____________

❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD   Issue Bank:_________________

Card No.: ________ -_________-_________ -_________ Exp. Date:___________(M/Y)

3-digit code: _________ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE: ________________________________

Phone No.: ________________________E-mail: _________________________________

Address: ____

PAYMENT FOR

US$ _____ ❑ Computational Linguistics & Chinese Languages Processing (IJCLCLP)

Quantity Wanted: ______

US$ _____ ❑ Journal of Information Science and Engineering (JISE)

Quantity Wanted: ______

US$ _____ ❑ Publications:______________________________

US$ _____ ❑ Text Corpora: ____________________________

US$ _____ ❑ Speech Corpora:___________________________

US$ _____ ❑ Others: _________________________________

US$ _________ ❑ Membership Fees ❑ Life Membership ❑ New Membership ❑Renew

US$ _________ = Total

Fax 886-2-2788-1638 or Mail this form to:

ACLCLP

% IIS, Academia Sinica

Rm502, No.128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan

E-mail: aclclp@hp.iis.sinica.edu.tw

Website: http://www.aclclp.org.tw

中 華 民 國 計 算 語 言 學 學 會
信用卡付款單

姓名: ____________________________(請以正楷書寫)　　日期:：____________________

卡別：❑ VISA CARD　　❑ MASTER CARD ❑ JCB CARD　　發卡銀行：______________

信用卡號：__________-__________-__________-__________　　有效日期：__________(m/y)

卡片後三碼：______________（卡片背面簽名欄上數字後三碼）

持卡人簽名：_______________________________________(簽名方式請與信用卡背面相同)

通訊地址：____

聯絡電話：____________________________E-mail：_________________________________

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$__________ ❑ 中文計算語言學期刊(IJCLCLP) _______________

NT$__________ ❑ Journal of Information Science and Engineering (JISE)

NT$__________ ❑ 中研院詞庫小組技術報告_____________________

NT$__________ ❑ 文字語料庫 ____________________________________

NT$__________ ❑ 語音資料庫 ____________________________________

NT$__________ ❑ 光華雜誌語料庫1976~2010

NT$__________ ❑ 中文資訊檢索標竿測試集/文件集

NT$__________ ❑ 會員年費：❑續會　　　❑新會員　　　❑終身會員

NT$__________ ❑ 其他: _____________________________________

NT$______________ = 合計

填妥後請傳真至 02-27881638 或郵寄至:
11529台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

		Surface	AIR (US&EURP)	AIR (ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本) ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05 中文詞類分析	10	15	13	_____	_____
7.	no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02 古漢語字頻表	11	16	14	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03 訊息為本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01 古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02 論語詞頻表	9	14	12	_____	_____
16.	no.98-01 詞頻詞典	18	30	26	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: ___________Total Due:__________

- **OVERSEAS USE ONLY**
- PAYMENT： ☐ Credit Card (Preferred)
  ☐ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or "中華民國計算語言學學會"
- E-mail：aclclp@hp.iis.sinica.edu.tw

Name (please print): ____________________________ Signature: ____________________________

Fax: ____ E-mail: ______________________________

Address：____

中華民國計算語言學學會
相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息為本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01　古漢語詞頻表（甲）	400	450	_____	_____
15.	no.97-02　論語詞頻表	90	100	_____	_____
16	no.98-01　詞頻詞典	395	440	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義（中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊（一年四期） 年份：______（過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※ 此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人： 黃琪 小姐、何婉如 小姐　　E-mail:aclclp@hp.iis.sinica.edu.tw

訂購者：______________________ 收據抬頭：______________________

地　　址：____

電　　話：______________________ E-mail: ______________________

Information for Authors

nternational Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites ubmission of original research papers in the area of computational linguistics and speech/text processing of atural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously npublished and cannot be under consideration elsewhere. Submissions should report significant new research esults in computational linguistics, speech and language processing or new system implementation involving gnificant theoretical and/or technological innovation. The submitted papers are divided into the categories of egular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full etails. Short papers can focus on a smaller research issue. And survey papers should cover emerging research ends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no rict length limitation on the regular and survey papers. But it is suggested that the manuscript should not xceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced 4 pages. All contributions will be anonymously reviewed by at least two reviewers.

opyright：It is the author's responsibility to obtain written permission from both author and publisher to eproduce material which has appeared in another publication. Copies of this permission must also be enclosed ith the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to cilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright rm, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP ciety, will be required before the manuscript can be accepted for publication. The papers published by CLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database rvices.

yle for Manuscripts: The paper should conform to the following instructions.

Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size 11 points or larger.

Title and Author: The first page of the manuscript should consist of the title, the authors' names and stitutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and nail address. The title of the paper should use normal capitalization. Capitalize only the first words and such er words as the orthography of the language requires beginning with a capital letter. The author's name ould appear below the title.

Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords equired. The abstract should not only indicate the scope of the paper but should also summarize the author's nclusions.

Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-nd margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).

Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with perscript numbers. Footnotes may appear at the end of manuscript

Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or itten clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in entheses.

References: All the citations and references should follow the APA format. The basic form for a reference ks like

`Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article.` *`Title of Periodical, volume number`*`(issue number), pages.`

e shows an example.

`Scruton, R. (1996). The eclipse of listening.` *`The New Criterion, 15`*`(30), 5-13.`

e basic form for a citation looks like `(Authora, Authorb, and Authorc, Year)`. Here shows an example. ruton, 1996).

ase visit the following websites for details.

APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)

APA Stytle (http://www.apastyle.org/)

page charges are levied on authors or their institutions.

al Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply l manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

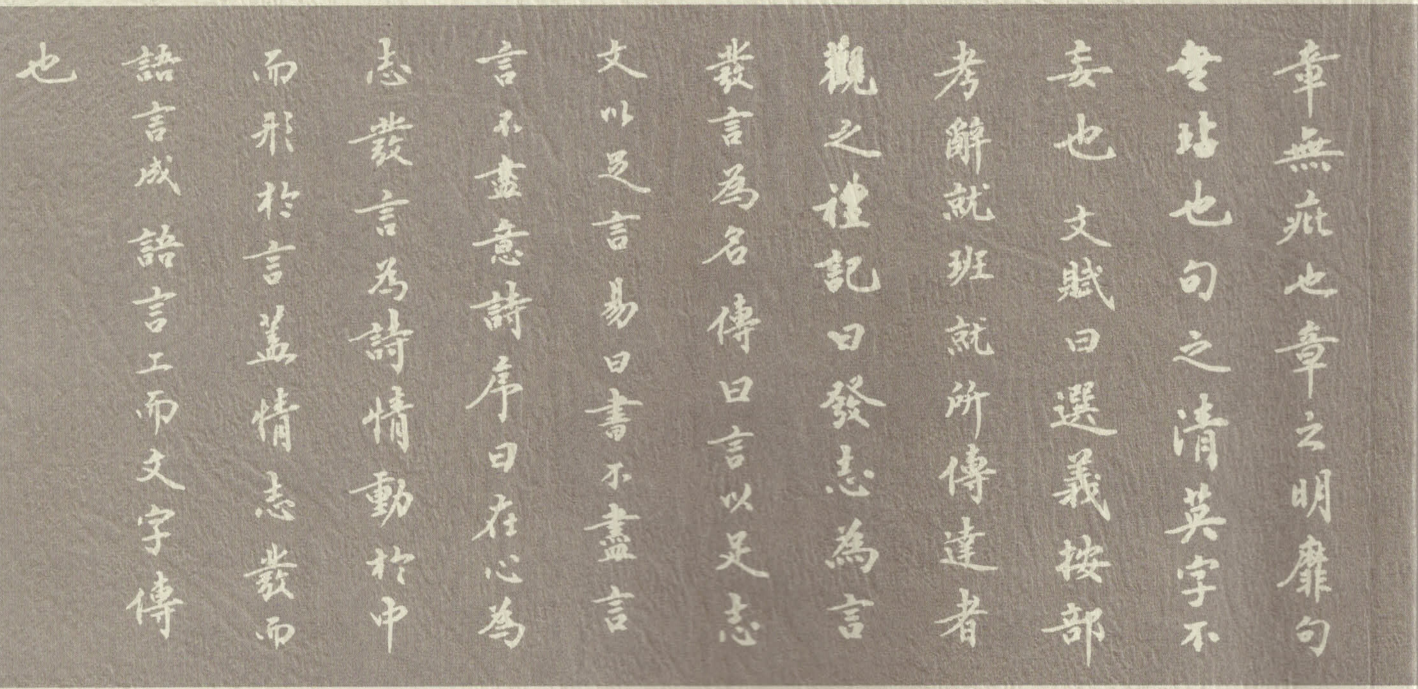
line Submission: http://www.aclclp.org.tw/journal/submit.php

ase visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.22 no.1 June 2017

Contents

Papers



ISSN: 1027-376X
The Association for Computational Linguistics and Chinese Language Processing