

國立交通大學

資訊工程學系

博士論文

中文寫作自動評閱之概念化方法

Conceptualization Methodology for Chinese Automatic Essay Scoring



研究生：張道行

指導教授：李嘉晃 教授

中華民國九十六年六月

中文寫作自動評閱之概念化方法  
Conceptualization Methodology for Chinese Automatic Essay Scoring

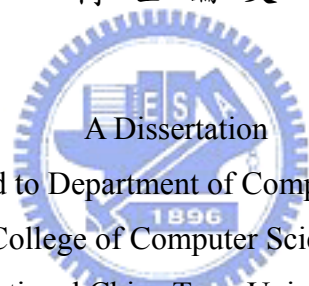
研究生：張道行

Student : Tao-Hsing Chang

指導教授：李嘉晃

Advisor : Chia-Hoang Lee

國立交通大學  
資訊工程學系  
博士論文



A Dissertation  
Submitted to Department of Computer Science  
College of Computer Science  
National Chiao Tung University  
in partial Fulfillment of the Requirements  
for the Degree of  
Ph. D.  
in  
Computer Science

June 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年六月

# 中文寫作自動評閱之概念化方法

學生：張道行

指導教授：李嘉晃 博士

國立交通大學資訊學院 資訊工程學系



寫作自動評閱技術在教育研究領域是相當重要的研究工具。雖然已經有許多英文的寫作自動評閱工具問世，然而由於其主要的架構設計仍依靠文法分析，因此後續的發展及應用均受到很大的限制。另外，由於寫作架構本質上的差異，使得中文寫作也無法使用現有工具進行自動評閱。本論文以寫作者的概念發展歷程為基礎，提出以概念-連結架構進行評閱的新方法，目的在測量概念發展過程中的表現。因為這樣的架構設計著重於作品的語意分析，使得中文寫作自動評閱技術發展的困難得以突破，其評閱結果也較具有教育上的應用價值。實驗結果顯示這個以寫作本質特徵為基礎的方法具有相當良好的效能。

# Conceptualization Methodology for Chinese Automatic Essay Scoring

Student : Tao-Hsing Chang

Advisor : Prof. Chia-Hoang Lee

Department of Computer Science  
College of Computer Science  
National Chiao Tung University



## Abstract

Automatic essay scoring system is a very important research tool for such areas as educational testing and psychometrics. Although some English AES systems have been proposed and developed successfully, it is still a difficult and interesting issue for Chinese AES. The thesis proposed a novel methodology for scoring Chinese essays based on the extraction and analysis of conceptual frameworks in essays. There are three characteristics in the methodology. First, it performs well based on the analysis of semantics in essays even if it does not employ surface features and syntax features. Second, the result of evaluation can be used for instructional feedback to the authors because it refers the conceptualization progress of authors to evaluate the quality of essays. Third, it overcomes the difficulties of applying current English AES systems to Chinese. Experimental results show that the performance of the methodology is quite close to that of current English AES systems.

## 誌 謝

本文提出了一個假設：一個次概念只有在其他次概念存在時存在。從開始這個論文題目起的幾千個日子，經歷的人生波折實在刻骨銘心。但是有許多關心我、照顧我、陪伴我、幫助我的人，讓我可以走到現在。沒有這些人，我和這篇論文的存​​在也就沒有意義。我不想像頒獎典禮致謝詞般的點名致謝，因為這些人就是我的存在、我的過去、我的未來。

謹以此文獻給影響我一生的人們



# CONTENTS

Abstract (in Chinese).....	i
Abstract.....	ii
Acknowledgements.....	iii
Contents.....	iv
List of Tables.....	vi
List of Figures.....	vii
Chapter 1 Introduction.....	1
Chapter 2 Previous Studies.....	4
2.1 Current AES.....	4
2.2 Contrastive Rhetoric.....	7
2.2.1 Sentence structure.....	8
2.2.2 Topics.....	8
2.2.3 Organization.....	9
2.2.4 Figure-of-Speech.....	10
2.3 Extraction and Classification of Chinese Words.....	10
Chapter 3 Conceptualization of Essays.....	12
3.1 Extraction of Words in Subconcepts.....	14
3.2 Unknown Word Extraction.....	16
3.2.1 Retrieving Meaningful Strings.....	17
3.2.2 Filtering Familiar Strings.....	18
3.2.3 Recovering Prefixed/Suffixed Words.....	19
3.2.4 Comparison with Previous Method.....	20
3.3 Unknown Word Classification.....	21
3.3.1 Inferring Category Based on Morphological Rules.....	22
3.3.2 Inferring Category Based on Contextual Rules.....	23
3.4 Thematic Subconcept Hierarchy.....	24
3.4.1 Asymmetrical Semantic Relation Matrix.....	25
3.4.2 Constructing Thematic Subconcept Hierarchy.....	26
Chapter 4 Selection of Concepts.....	29
4.1 Set of Literary Sememes.....	30
4.2 Extraction of Literary Sememes.....	31
4.2.1 The Correlation between Candidate and Essay.....	31
4.2.2 Using Candidate Sets to Score Essays.....	31
4.2.3 Estimating the Performance of Candidates Quantitatively.....	32
4.3 Usefulness of Literary Sememes for Scoring Essays.....	33
4.3.1 Experimental Corpus and Definition of Performance.....	33

4.3.2	Performance of Literary Sememes for Scoring Essays.....	35
Chapter 5	Connection of Concepts.....	36
5.1	Extraction and Transformation of the Concepts in C-L Structure .....	36
5.2	Inter-paragraph Connections.....	37
5.2.1	Similarity Measure between Inter-paragraph Connections.....	37
5.3	Intra-paragraph Connections.....	39
5.3.1	Similarity Measure between Intra-paragraph Connections.....	41
5.3.2	Scoring Essays by Intra-paragraph Connections .....	42
5.4	Usefulness of C-L Structures for Scoring Essays .....	42
Chapter 6	Decoration of Concepts .....	44
6.1	Building Sets of Connectives and Literary Connectives .....	45
6.2	Extracting FOS “Pi-yu” .....	46
6.3	Extracting FOS “Pai-bi” .....	47
6.4	Usefulness for Scoring Essays .....	48
Chapter 7	Performance of Conceptualization for Scoring Essays.....	49
7.1	Predicting Model.....	49
7.2	Performance of Predicting Model.....	51
Chapter 8	Conclusions.....	53
Reference	.....	54



## LIST OF TABLES

Table 3.1	Part of the Morphological Rule Base.....	20
Table 3.2	An Example for Morphological Rules Containing Two Categories .....	22
Table 4.1	A Performance Table.....	32
Table 4.2	A Weighted Table .....	33
Table 4.3	Performance of Using Literary Sememes to Scoring Essays.....	35
Table 5.1	Performance of Inter-paragraph Connection.....	43
Table 5.2	Performance of Intra-paragraph Connection.....	43
Table 6.1	The Distributions of the Ratios of Essays to All Essays .....	48
Table 7.1	Performance of the Improved MBM for Scoring Essays.....	52
Table 7.2	Performances of the Proposed Method using Different Feature Sets .....	52





## LIST OF FIGURES

Fig. 3.1	Two Viewpoints for Semantic Structure of Chinese Essays.....	12
Fig. 3.3	Design for Extracting Concepts in C-L Structure.....	14
Fig. 3.4	Methods for Achieving Set of Words in Subconcepts.....	15
Fig. 3.5	Architecture of Unknown Word Extraction and Classification.....	16
Fig. 3.6	Comparison the Proposed Method with [28].....	21
Fig. 3.7	An Example for Estimating Association among Words .....	26
Fig. 3.8	Part of a Thematic Subtopic Hierarchy .....	27
Fig. 3.9	Algorithm for Determining the Levels of Subconcepts.....	28
Fig. 5.1	Methods for Predicting Scores Using the Connection of Concepts .....	36
Fig. 5.2	Illustration for C-chains.....	37
Fig. 5.3	Illustration for P-chains and S-chains.....	38
Fig. 5.4	An Example of Principles for Scoring S-chains.....	39
Fig. 5.5	Forms of R-chains .....	40
Fig. 5.6	An Example for Comparison between R-chains .....	41



## Chapter 1 Introduction

Automatic essay scoring (AES) system is a very important research tool for such areas as educational testing and psychometrics because studies in these domains often rely on a large number of writings to conduct various analyses. It is, however, often very difficult to obtain a large number of graded writings due to expensive cost and time consuming process of human grading. In English, successful development of automatic essay scoring system in the past has overcome these limitations and largely facilitated the progress of the stated research area. By contrast, the lack of Chinese automatic essay scoring system (CAES) has limited the scale, validity and progress of these research areas.

AES is still a difficult, intricate and interesting issue for researchers in artificial intelligences and natural language processing though some English AES systems have been proposed and developed successfully. In general, current syntactical and lexical analysis tools and techniques serve the basis for AES systems. However, AES systems should analyze the semantic characteristics of an essay, since the essence of an essay is the semantic representation of an author on given theme. Presently, the issue of AES systems lack the methodology for analyzing contextual semantics is still quite difficult in the domains of artificial intelligences and natural language processing. The development of such methodology will play an important role in overcoming the bottleneck of the researches in AES.

In addition, many studies also indicate semantic analysis is crucial and critical factor for Chinese AES. Some studies demonstrate that Chinese essays tend to be organized with parataxis structure connecting concepts with their semantics. By contrast, English essays tend to be organized with hypotaxis connecting concepts with grammatical framework and connectives. The studies for extracting hypotaxis structure in English essays have been used to develop English AES successfully. Based on the observation, developing methods for extracting parataxis structure become a very important task for constructing Chinese AES.

Furthermore, the linguistic difference of the languages between other languages and English also suggest the need to reconsider such factors as syntactic structure in designing AES for other languages. For example, current English AES systems successfully analyze the syntax and sentence structure of an essay by developing effective parser and grade the essays accordingly. By contrast, to develop of a

effective parser in Chinese is an extremely difficult task because the grammar specification is quite loose and fuzzy. Accordingly, identifying and extracting valid sentences is quite difficult task for machine. Beside, the importance of grammar and sentence structure may not play the same role as that in English.

We propose to employ the conceptualization progress of writing to the semantic analysis of essays for next generation AES systems, especially Chinese AES. Essays can be regarded as a set consisting of concepts on a given theme and the semantic association among the concepts, while a concept represents a subset consisting of the subconcepts of the concept and the semantic association. Based on the observation, writing progress is composed of three conceptualization phases. First, an author selects major concepts from his background knowledge based on a given theme. Then, the author organizes and arranges the occurrence order and space of these concepts. Finally, the author develops and constructs the major concepts with many subconcepts, and uses many writing skills to enrich and enhance the subconcepts. In this thesis, the three phases are respectively denoted as the selection of concepts, the connections among concepts and the decorations for concepts.

The conceptual framework of an essay represents the outcome of conceptualization progress of an author on a given theme. In this thesis, we will explore the various issues and effectiveness of using conceptual framework to score an essay. There are three tasks we have to solve before scoring. First is to develop the extraction, classification and organization of concepts on given theme. These concepts will be transformed into set of subconcepts and these subconcepts will be into a subconcept hierarchy. Using the concepts and subconcept hierarchy, essays can then be decomposed into conceptual frameworks. Second task is to construct the extractions of concepts, connections and decorations in the frameworks. Third is to design an integrated model for predicting scores from various components in the frameworks

This thesis will be organized as follows. In Section Chapter 2, some current English AES systems will be discussed and some studies about contrastive rhetoric which indicate the difference between essays in English and Chinese are reviewed. In addition, some approaches for Chinese text processing are also studied. Section Chapter 3 defines and analyzes conceptual structure of essays and discusses some methods for constructing subconcept hierarchy. The subconcept hierarchy is used to various conceptual methods for Chinese AES in Section Chapter 4, Chapter 5 and

Chapter 6. Section Chapter 4 develops methods for collecting particular subconcepts and then using them to score essay. Section Chapter 5 proposes methods for constructing the structures of inter-paragraph and intra-paragraph in essays and then measuring similarities between the structures to score essay. In Section Chapter 6, some identification methods for figures-of-speech are discussed. These figures-of-speech will be input features of scoring system. Section Chapter 7 discusses an improved multi-variate Bernoulli model for CAES which integrates various conceptual features. Section 8 gives a conclusion.



## Chapter 2 Previous Studies

This section reviews previous studies related to the various design issues of Chinese AES. Subsection 2.1 surveys the methodologies and performances of current AES systems. Subsection 2.2 explores intrinsic differences between Chinese essays and English essays based on many studies of contrastive rhetoric. Since there are no blank separating Chinese words, Chinese essays need to be first segmented into a set of words. This task includes such related procedures as word segmentation, unknown word extraction and classification. Subsection 2.3 will review previous methods and techniques for above procedures.

### 2.1 Current AES

All current AES systems are designed and implemented for English essays. A few of these systems claim that they are also suited for other languages. In the earlier 1990s, various AES methods were proposed and further commercialized. Although the technical details of the AES systems are rarely published in academic literatures due to business secret, some studies [16][48][54] for evaluating current AES systems are available. Below, five AES systems will be reviewed based on these studies.

#### PEG

PEG [44] is an AES system developed early in 1960s and being still improved even in recent years. The design of PEG bases on an assumption denoted as trins-proxes: the intrinsic variables of an essay could be measured using the approximations of the variables. For instance, the length of an essay can approximate intrinsic variable fluency; the number of preposition and relative pronouns represents the degree of complexity of sentence structure; the variation of word length displays author's writing skill on diction. In PEG, the intrinsic variables and corresponding approximations are denoted as trins and proxes respectively.

PEG predicts the score of a test essay using the following three steps. Firstly, PEG calculates the values of all proxes of each essay in training corpus. Secondly, the values of proxes and the scores of training essays are used as input to a statistical model called multiple linear regression. The model yields the coefficients of regression function which represent the relationship between the values and the scores

given by human. Finally, PEG calculates the values of proxies in a test essay, and then regression function generates a predictive score for the test essay.

The advantages of PEG are easy implementation and fast operation. Experimental results show the multiple regression correlation between predictive score and graded score as 0.87. However, PEG is criticized for lack of evaluating the semantics of an essay. The drawback results in that PEG is easily tricked by prankish essays. Furthermore, since PEG only relied on proxies, it cannot provide instructional feedback to students.

## IEA

IEA [30] is an AES system based on latent semantics analysis (LSA) technique. LSA uses a two dimensional matrix to express the semantic space of documents and words. For example, a set of 200 essays containing 3000 words could construct a  $200 \times 3000$  matrix to represent the semantic space of the set. Since the space is often sparse, LSA uses singular value decomposition (SVD) to transform the semantic space to three smaller-dimension matrices. The algebraic operation could also discover latent semantic relations between documents and words. For instance, although an essay containing term “president of United State” does not mention term “white house”, the semantic relation between the essay and other essays mentioning term “white house” could be constructed against SVD.

The vectors in new semantic space can represent the semantics of essays more exactly and precisely. IEA uses the vectors to estimate the semantic similarity between a test essay and training essays scored by human. Then, IEA gives the predictive score of the test essay with the score of the training essay which is most similar to the test essay. The accuracy of similarity measurement of essays on semantics also increases. Experimental results show that the accuracy of IEA is between 0.85 and 0.91. In addition, IEA can effectively detect plagiarism while most AES systems cannot. The characteristic is very necessary to high-stake assessments because plagiarism affects the validation of assessments.

Although IEA has the advantages mentioned above, its application is still limited due to lack of other aspects of human graders, e.g. sentence structure and organization. The drawback results in that IEA cannot satisfy instructional requirements and provide meaningful feedback to students.

e-rater

e-rater [1][4] employs corpus-based methods to extract the features of scoring rubrics by analyzing essays scored by human graders. The basic tasks are to identify such features as discourse, syntax and topic-domain.

Discourse organization is identified by using the concept framework of conjunctive relations. The concept framework can be extracted by cue words, terms or sentence structures. For instance, cue word “perhaps” may express a belief; term “in summary” may represent a conclusion; complement clauses may identify the beginning of a new argument.

The versatility of syntactic structures can be evaluated by identifying the subjects, verbs and various clausal structures in sentences. These components in sentences can be derived by Microsoft natural language processing tool (MsNLP).

Features of topic-domain are used to evaluate the vocabulary usage of essays. It uses cosine as the metric to evaluate the similarity on the variety and type of vocabulary between a test essay and training essays. The basic assumption is that good essays usually resemble that of other good essays and differ from that of poor essays.

e-rater employs two modules to score test essays based on the above features. One module assigns different weights to features while the other module measures the similarity of features between a test essay and training essays. It then predicts the score of the test essay according to the result of the comparison.

By contrast to IEA, e-rater considers more aspects of scoring rubrics. e-rater has been applied to score over 75000 essays of famous assessment GMAT and the accuracy of scoring essays is between 0.87 and 0.94. Since e-rater heavily relied on grammar structure features, it cannot be applied to many languages in which grammar parsers are difficult to design.

## BETSY

BETSY [46] uses feature-based Bayesian models to predict the score of an essay. Different Bayesian models have been widely applied to text classification. BETSY uses multi-variate Bernoulli model (MBM) as one of its predictive modules. MBM can effectively incorporate with various features to predict score. BETSY claims that it uses the best features from PEG, IEA and e-rater as the features of MBM. It also

employs three procedures, namely stemming, stop words and feature selection, to normalize text in order to extract features more precisely and grades essays more accurately. Experimental results show that the accuracy of scoring essays by BETSY is over 0.8.

BETSY has two major advantages: (i) it employs the features from different aspects of scoring rubrics. (ii) it uses a high-performance classifier to increase the accuracy of predicting scores. However, there are still two constraints in BETSY. First, BETSY requires a large number of training essays. Secondly, BETSY also needs an effective grammar parser to maintain the accuracy.

### IntelliMetric

IntelliMetric [17] is a business AES system developed by Vantage Learning Company. The basic architecture of IntelliMetric is also feature-based methods. IntelliMetric claims that it contains over 300 semantic-, syntactic- and discourse-related features. These features can be classified into five categories of scoring rubrics: focus and unity, organization, development and elaboration, sentence structure, and mechanics and conventions. IntelliMetric contains two powerful tools CogniSearch and QuantumReasoning. CogniSearch includes a grammar parser to analyze sentence structure. It also uses training essays to construct a non-linear scoring model. However, the detail of features, tools and scoring model in IntelliMetric is not available. The company claims that the accuracy of IntelliMetric is high as 0.96 and can be applied to other languages.

## 2.2 Contrastive Rhetoric

A large number of researches [5][6][32], called contrast rhetoric and contrast linguistics, standard the difference of writing features between Chinese and English. These studies focus on four directions covering sentence structure, organization, topics as well as figures-of-speech. Sentence structure compares the difference of syntax and its structure. Topics and organization studies the difference in thinking and plot skill. Although organization can be analyzed at the levels of both paragraph and essay, it will be treated as the same in this study. Figures-of-speech compare the difference for the usage and categories of figures-of-speech between Chinese and English writings. Below, we will discuss the issues separately.



### **2.2.1 Sentence structure**

Many studies pointed out that the syntax in Chinese is more flexible and loose than that in English. Jaio [24] noted three fundamental differences. First, the sentence structure in English can be perceived as a tree structure while the sentence structure in Chinese can be perceived as a linear structure. Second, sentences in English emphasize hypotaxis which refers to grammatical subordination while sentences in Chinese focus on parataxis which refers to grammatical coordination. Parataxis often places sentences side by side without any connectives while hypotaxis uses word to express the temporal, logical and syntactic relations between sentences. Third, the sentence boundary in English is very crisp. As long as the subject and predicate appears, the sentence is basically complete. By contrast, the amount of information in Chinese sentence is not strictly constrained. A sentence is complete only when a sequence of actions is over. Each of the long sequence of actions can appear as a complete sentence or short sentence or a single action or modifier. Hence it is much harder to identify the boundary of a sentence.

Lee and Zeng [33] pointed out that the ordering of three basic elements (subject, verb and object) and the type of tags show no fundamental difference in both Chinese and English discourse. However, the constituents may be very different. For instance, the verb and the adjective can be used as subject in Chinese while the verb and adjective have to be changed into infinitive or participle before it can be used as a subject in English. The same restriction also appears in the predicates. In the case of structure changes, Chinese can move the predicate to the beginning of a sentence to form a derived sentence. By contrast, in English, there is no such pattern.

The above discussion illustrates that the Chinese syntax is quite flexible and more diversity in patterns than that in English. Hence, developing a powerful parser for Chinese is a quite difficult task.

### **2.2.2 Topics**

One writing requirement is to select materials and the illustration has to be consistent with the topic. However the material selection is quite different due to the difference of East West cultures. Chinese students often quote authoritative argument and classical article, and seldom express personal opinion and feelings. The usage of rhetoric is quite indirect and moderate. By contrast, western students like to present many evidences to support their arguments or viewpoints. Critical and logical

discussions are quite important for English writings. The observation indicates that the material selections of a Chinese essay should not be judged by the English automatic scoring system because of the different perspectives.

### **2.2.3 Organization**

Many studies noted a fundamental difference of discourse organization in Chinese and English writing. In particular, Kaplan [25] pointed out that the organization of discourse in English writing often appears as a linear sequence. By contrast, the discourse organization in Chinese writing often appears like a spiral pattern. A good organization in English writing starts with a topical sentence and develops various arguments in succeeding sentences to illustrate the topic. On the other hand, a good organization in Chinese will discuss the topic from various perspectives and repeatedly illustrate the topic with various semantic expressions without word connection. Hu [21] further observes either the last paragraph or the end of the middle paragraph often contains highlight in Chinese.

Some studies made a further analysis on the difference of the discourse organizations. Zeng [58] pointed out that grammatical framework connecting subject and verb does not exist in Chinese writing. Instead of using connectives for combining sentences, it uses the sequence and logical order to connect sentences. Many times, topic idea is not evident in a paragraph. Instead it likely contains an invisible or a subconscious topic idea. On the other hand, the principles of organization for paragraph and discourse are consistent. They all submit to coherence, completeness. Hence, topic sentence, concluding sentence and supporting sentence can be seen quite often. The tree structure of discourse [36] also shows that the coherence of discourse in English tends towards the relationship between main clause and subjective clause. By contrast, the structure of Chinese discourse tends towards parallel sequence.

Although, effective paragraph all stresses unity, coherence, and completeness in both English and Chinese writing, the different thinking process behavior from the east and west culture would naturally lead to distinction in their own discourse. For instance, Scollon et al. [47] remarks that people in western cultures use a deductive method of reasoning or argument, while people in eastern cultures use an inductive method of reasoning. This indicates that human grader from different cultures may give different judgments for the topic organization.

### 2.2.4 *Figure-of-Speech*

The usage of figure-of-speech is an important factor for high quality of writing. Although there are many common usages of figures-of-speech in Chinese and English writing, each language has its own unique figures-of-speech. Bai and Shi [3] studied the differences of figures-of-speech in Chinese and English. Of 16 figures-of-speech used in Chinese and 26 figures-of-speech in English, they each have 6 figures-of-speech not observed in another language. Even, for the similar figures-of-speech, the occasions and skill to use it are different. Hence, the grading factors may be different in using the figures-of-speech to judge essays.

### 2.3 **Extraction and Classification of Chinese Words**

Most studies have focused either on word extraction [14][19][28][35][37][49][57] or word classification [2][11][41][43][50][53]. This subsection briefly reviews extraction and classification approaches for the purpose of comparison and discussion.

Extraction approaches can be categorized into rule-based [14][19][35][49] and statistics-based [28][37][57] methods. Rule-based extraction approaches generally analyze the characteristics of unknown words in terms of linguistic factors such as morphemes, roots and semantic networks, to identify whether unknown words occur and to locate their boundaries. Although the precision rates of these approaches are usually higher than those of statistics-based approaches, they suffer three major drawbacks. First, a rule base cannot easily include all rules for identifying unknown words. Second, both semantic and syntactic ambiguities reduce the reliability of the rules. Third, the approaches need to rely on a large training corpus, in which the sentences are segmented and the terms are tagged, to generate rules. Large, prepared and domain-specific corpora are impractical in the domain of this work.

Statistics-based extraction approaches considered characters that occur frequently together as possible evidence of a word. Although these statistical approaches overcome some of the drawbacks discussed above in the rule-based approaches, they have two shortcomings. First, the statistical approaches often extract meaningless strings as unknown words. Second, they perform badly for unknown words that appear rarely even though the training corpus is very large.

Most classification approaches [2][11][41][43][50] are rule-based since the

possible categories of unknown words, critical parameters in statistical models, cannot be obtained from current lexica. The rule-based approaches often first generate morphological, syntactical and contextual rules from universal lexica or corpora, and then apply these rules to generate a set of possible categories of unknown words. They also suffer such drawbacks as reliability of the rules and incompleteness of the rule base discussed before. Additionally, properties such as affixes in alphabetic languages cannot be used in approaches for non-alphabet languages.

Although these extraction and classification approaches are inadequate for domain-specific small corpora, several studies contain interesting observations. Chen and Ma [14] proposed a refined rule-based extraction method, which first applies a rule-based approach to detect the occurrence of Chinese unknown words, and then applies morphological rules and statistical methods to confirm the boundaries of the unknown word. Finally, a verification procedure, which depends on the validity of the structure and syntax, and on local consistency, is utilized to refine and improve the result of previous steps. Lai and Wu [28] proposed an effective statistics-based method to extract unknown words, in which the PLU-based likelihood ratios (PLR) of all Chinese character sequences are computed from the co-occurrence in the sequences. The sequences with a high PLR value or frequency are then classified into several groups, in which quality sequences are regarded as unknown words.

Tzeng et al. [53] proposed a hybrid method to classify Chinese unknown verbs into subcategories by utilizing a set of morphological rules to determine the categories of the unknown verbs. An instance-based categorization algorithm is then employed to classify words that cannot be classified using morphological rules. Manning and Schutze [40] observed that all successful classification approaches are based on lexical information from words. The lexical information, gathered from a huge corpus in which the words are segmented and tagged, is essential because of the extremely uneven distribution of a word's usage across different parts of speech. Since the lexicon information of unknown words cannot be obtained directly from the corpus, Bai et al. [2] scored possible categories of the unknown word using contextual rules to simulate its lexicon information, and then classifies it with the highest-scoring category.

### Chapter 3 Conceptualization of Essays

Many studies mentioned in Subsection 2.2.3 point out the organization of essays in Chinese is parataxis structure. Concepts are crucial elements in parataxis structure because concepts are semantic units for expressing paragraphic topic. Furthermore, the connection among concepts, including the order and position of the appearances of concepts, also perform latent semantics. Both selection and arrangement of concepts, hence, affect the quality of essays.

A general viewpoint for semantic structure of Chinese essays is shown in Fig. 3.1(a). The essay in Fig. 3.1 (a) consists of three paragraphs and seven concepts for a theme. Conventional viewpoint considers that a paragraph interprets a subtopic of the theme and all subtopics should be semantically related to each other. It is abstract and incomplete because the model does not express the importance of selection and arrangement of concepts in paragraphs.

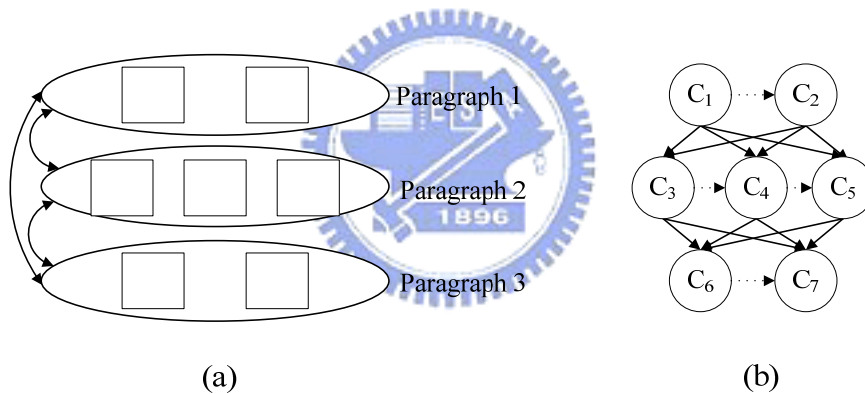


Fig. 3.1 Two Viewpoints for Semantic Structure of Chinese Essays

Fig. 3.1(b) shows an extension and modification of the essay structure in Fig. 3.1(a). In Fig. 3.1(b), the paragraphs are transformed to three sets of concepts and the relations among adjacent paragraphs are translated into unidirectional relations between concepts. In addition, the concepts in a paragraph are related to each other. The essay structure in Fig. 3.1 (b), denoted concept-link structure, C-L structure for short, has two characteristics. Firstly, the total sum of the semantics for all concepts in a set of concept can fully represent the semantics of subtopic in the corresponding paragraph. Secondly, different appearance orders of concepts represent different semantics. C-L structure will serve as the basis for developing methods for automatic scoring essays in Section Chapter 4, Chapter 5 and Chapter 6.

Two basic issues for extracting the C-L structures of essays should be discussed

first: (i) how to define concepts of C-L structure in essays? (ii) how to extract the concepts of essays? This thesis employs subconcepts and thematic subtopic hierarchy to define and extract the concepts respectively. In general, the semantics of an essay entitled a theme can be distinguished hierarchically as three layers of subtopic, concept and vocabulary. Subtopic contains several concepts while vocabularies are the components of concepts. Instead of defining and extracting concepts and subtopics, most of the studies use the associations of vocabularies to analyze the semantics of documents, for it is a very difficult task to precisely define the semantics of the concept and subtopics. However, the usage of vocabulary often performs poorly because the associations are uncertain and imprecise.

Using a set of subconcepts to represent concept can overcome above dilemma. The relation between concept and its subconcepts is shown as Fig. 3.2. Concepts are defined herein as a set of subconcepts. A subconcept consists of a word and the relations between subconcepts on given theme. The semantics of subconcepts is defined by the theme and other related subconcepts, not the semantics of the word of the subconcept in lexicon. Using subconcepts and the relations between them, the semantics of a concept can be presented clearly and concretely. Furthermore, a concept often contains one or more major subconcepts while other subconcepts are the assistants of major subconcepts. For example, concept “events in classroom” includes subconcepts “classroom”, “textbook”, “classmates”, “conversation” and so on. In the concept, subconcept “classroom” is a major subconcept and other subconcepts are its assistant. Hence, major subconcepts of a concept can be regarded as the representatives of the concept.

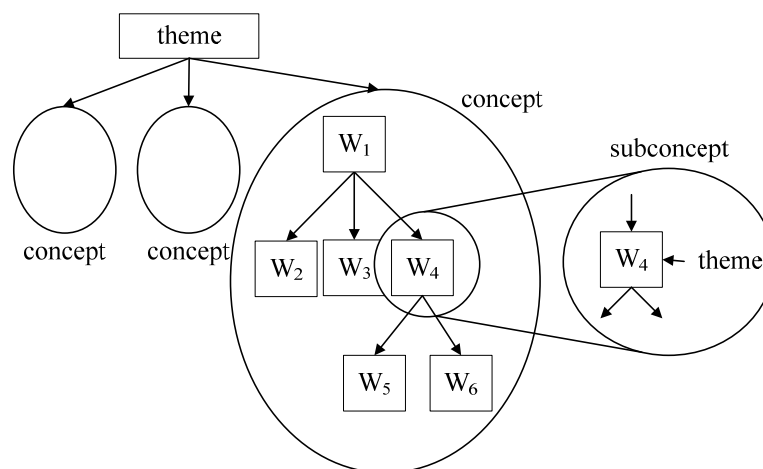


Fig. 3.2 Definition of Concept and Subconcept in C-L Structure

Subconcepts can be also used to extract concepts in C-L structure. Fig. 3.3 shows our design containing of three procedures for extracting the concepts. First procedure extracts words in subconcepts in training corpus. Second procedure employs an algorithm and the co-occurrence of the words to construct a thematic subconcept hierarchy. Since the hierarchy contains the classes of all subconcepts on a theme, third procedure can transform an essay into C-L structure using the hierarchy.

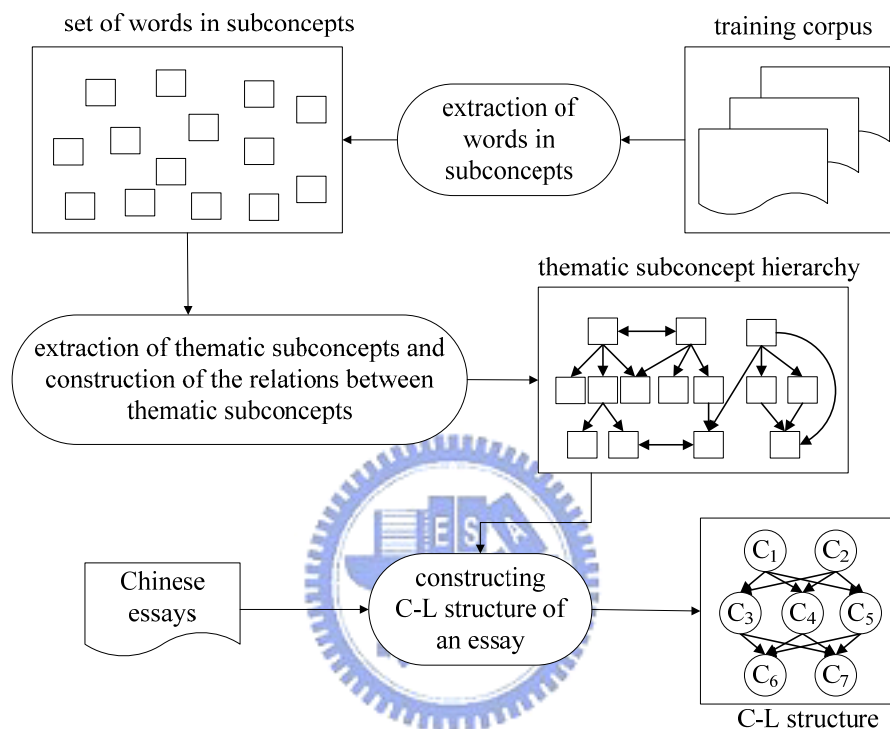


Fig. 3.3 Design for Extracting Concepts in C-L Structure

Below subsections will further discuss three important and fundamental issues. Subsection 3.1 introduces our methods for extracting the words in subconcepts. Since the words of some subconcepts are not appear in lexicon, Subsection 3.2 and 3.3 proposed our methods for extracting and classifying unknown words of subconcepts. Subsection 3.4 discusses the extraction of thematic subconcepts and the construction of thematic subtopic hierarchy.

### 3.1 Extraction of Words in Subconcepts

The words in subconcepts are often used to describe such important elements as person, place, action and events, these words can be collected from nouns, verbs, adverb and adjective in essays. This task is trivial for English essays, but not for Chinese because there is no blank between Chinese words. Chinese sentences must

first be segmented into words and phrases. Our strategy for this issue, which is shown in Fig. 3.4, consists of two steps: (i) extracting and classifying unknown words to be known words first. (ii) segmenting text into words of subconcepts. Sinica Autotag [38] is one of the tools developed to segment text into known words and tag these words precisely.

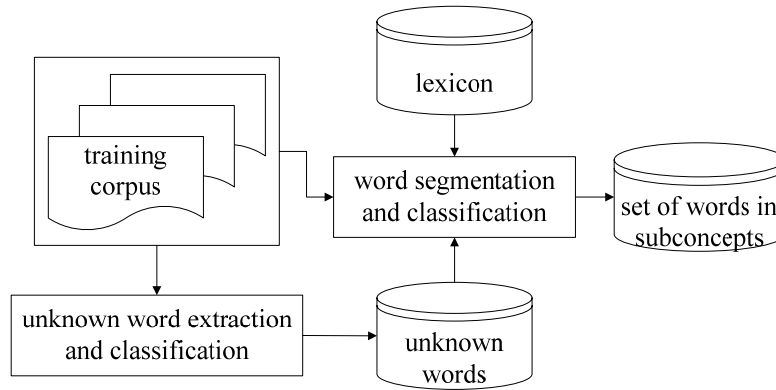


Fig. 3.4 Methods for Achieving Set of Words in Subconcepts

However, none of current approaches work well for unknown words in a small corpus. The limitation greatly decreases the performance of Chinese AES because most of training corpora is small in real AES applications. Hence, based on two inferences, this thesis proposes new methods for unknown word extraction and classification. First, extraction should remove meaningless strings based on strict conditions, and recall meaningful terms from the strings based on high-accuracy rules. The hybrid design can maintain both precision and recall rates of extraction. Second, conversely, classification procedure employs high-accuracy rules as far as possible to tag unknown words, and classify the remaining unknown words with the contextual rules in both universal and domain-specific corpora.

Fig. 3.5 shows the architecture of the proposed method, which consists of training, extracting, and classifying phases. The training phase generates morphological and contextual rule bases using a current lexicon and corpus. Both rule bases are referenced by the procedures in the extracting and classifying phases. The detail of extracting and classifying phases will be discussed in Subsection 3.2 and 3.3.



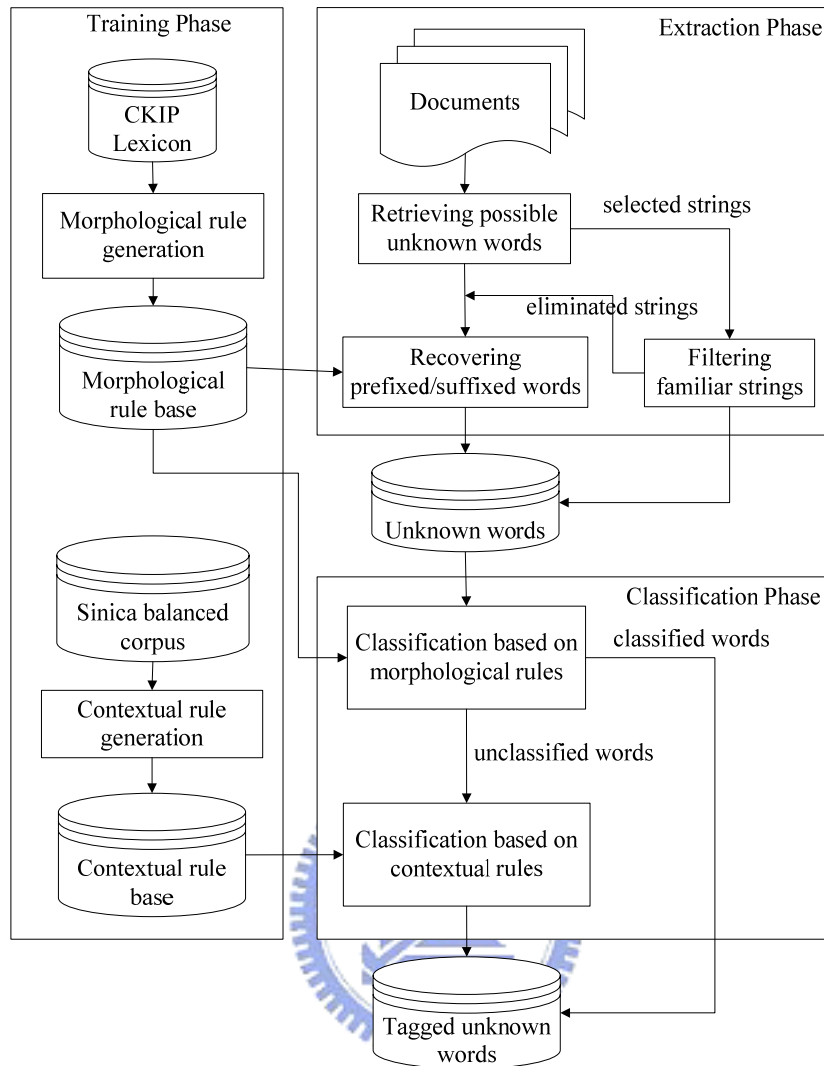


Fig. 3.5 Architecture of Unknown Word Extraction and Classification

### 3.2 Unknown Word Extraction

The extracting phase of the proposed method shown in Fig. 3.5 contains three procedures, namely retrieving possible unknown words, filtering familiar strings and recovering prefix/suffix words. The retrieval procedure estimates the probability that a series of characters is an unknown word. The filtering procedure uses a syntactic approach to eliminate familiar patterns that appear frequently in the corpus but do not form specific words. The recovery procedure reexamines all removed strings and recall qualified ones to the set of unknown words based on the morphological rule base. The three procedures will be discussed in below subsections.

### 3.2.1 Retrieving Meaningful Strings

The retrieval procedure in the proposed method estimates the probability that a series of characters is an unknown word based on a strict concept of PLR called SPLR. In contrast to conventional PLR approaches, the proposed SPLR method uses a strict condition to exclude a large number of meaningless terms, at the risk of sacrificing a few meaningful terms. The proposed method first retrieves all strings, and computes their frequencies of occurrence in a corpus where the maximum length of strings is previously defined and set. Additionally, stop words are removed from the set of these strings.

The following assumption is used throughout this work. For a frequently occurring string  $t$  of  $n$  characters  $c_1, c_2, c_3, \dots, c_n$ , the substring  $t_L$  of  $t$  is defined as the first  $n-1$  characters  $c_1, c_2, c_3, \dots, c_{n-1}$ , the substring  $t_R$  of  $t$  is defined to be the last  $n-1$  characters  $c_2, c_3, \dots, c_{n-1}, c_n$ . Apart from the monosyllabic strings, each string  $t$  always has  $t_L$  and  $t_R$ , of which the length of each is one less than that of the string  $t$ . Next the parent strings  $tp$  of  $t$  are defined to be  $c_0, c_1, c_2, c_3, \dots, c_n$  or  $c_1, c_2, c_3, \dots, c_n, c_{n+1}$  for some  $c_0$  or  $c_{n+1}$ . The parent string is clearly not unique. The example  $t = \text{“學習步道”}$ ,  $t_L = \text{“學習步”}$ ,  $t_R = \text{“習步道”}$  and  $tp = \text{“在學習步道”}$  illustrates these notations. The strict phrase likelihood ratio (SPLR) of  $t$  is defined as

$$SPLR(t) = \frac{tf(t)}{\max(tf(t_L), tf(t_R))}, \quad (3.1)$$

where  $tf$  represents the frequency of occurrence of a string in documents. Since the number of occurrences of a string  $t$  must be less than that of both  $t_L$  and  $t_R$ , the SPLR of any string is less than 1.

A string  $t$  with high-SPLR value signifies that neither substrings  $t_L$  and  $t_R$  appear in other contexts themselves but appear only when string  $t$  appears. Conversely, a string with low-SPLR value indicates that at least one of its substrings frequently appears in another context. Hence, a string  $t$  with high-SPLR can reasonably be considered as a likely unknown word, because it cannot be reduced to a smaller substring. This observation is reflected in the first part of Condition (3) in (3.2).

However, this approach still admits many meaningless terms into the set of unknown words. The next observation, reflected in the first part of Condition (4) in (3.2), is that if there is a parent string  $tp$  of  $t$  such that both of  $tp$  and  $t$  are frequently

co-occur and cannot be separated, then the string  $t$  cannot be considered as an unknown word, even if it satisfies Condition (3). To summarize the observations discussed above, a string  $t$  that meets the following four conditions is considered as a possible unknown word:

$$\left\{ \begin{array}{l} (1) n > 1 \\ (2) tf(t) \geq c \\ (3) SPLR(t) \geq 1 - \varepsilon \text{ or } \frac{SPLR(t) \cdot tf(t)}{(length(t))^2} \geq d \\ (4) tf(t) \geq v \cdot tf(t_p) \text{ or } SPLR(t) \geq \mu \cdot SPLR(t_p) \end{array} \right. \quad (3.2)$$

where  $length(t)$  represents the number of Chinese characters comprising  $t$ ;  $t_p$  represents the parent strings of  $t$ ;  $c$  and  $d$  represent thresholds from experiments;  $\varepsilon$ ,  $\mu$   $v$  represent the coefficients from experiments which are used to deal with real data, and  $n$  represents the length of  $t$ .

The first condition states that monosyllabic strings are removed from the set of unknown words. The second condition states that the number of occurrences of a string must exceed a threshold. The second part of the third condition handles disyllabic strings, which often have SPLR values below the threshold. The second part of the fourth condition handles very frequent strings.

### 3.2.2 Filtering Familiar Strings

The set of possible unknown words identified by SPLR approaches includes familiar strings. For instance, the string “從教室裡衝出來” (rushing out of the classroom) can be seen quite often in the essays entitled “Recess at School”. Obviously, it is a familiar string but not an unknown word. Because unknown compounds are difficult to distinguish from familiar strings, any statistics-based procedures would admit both into the set of unknown words irrespective of how the threshold is adjusted.

Notably, familiar strings often comprise common words from domain-specific corpora, while unknown compounds often comprise particular words. For instance, from the essays entitled “Recess at School”, the familiar string “從教室裡衝出來” (rushing out of the classroom) contains the common words “從”, “教室” and “衝出

來”。 By contrast, the unknown compound “文具用品”(stationery) comprises words “文具”(stationery) and “用品”(appliance), both of which rarely occur in domain-specific corpus. Thus, a familiar string  $t$  can be easily distinguished from the unknown compounds by the following condition:

$$\frac{tf(t)}{\min tf(t_c)} \leq \tau, \quad (3.3)$$

where  $t_c$  represents a set of known words included in  $t$ , and  $\tau$  represents a threshold from experiments.

### 3.2.3 Recovering Prefixed/Suffixed Words

SPLR can successfully eliminate meaningless strings, but it also sacrifices some meaningful terms comprising a familiar stem and a monosyllabic prefix or suffix. For instance, the string “總領隊”(總(*general*)領隊(*leader*), general leader) would not be identified as an unknown word because “leader” often appears as a familiar stem, resulting in a low SPLR value for “general leader”. Instead of lowering the threshold used in Formula (3.2), which would reduce the accuracy of extraction, morphological procedure is employed to identify these unknown words.

Many approaches based on morphological rules [13][14][41][50][53] to extract and classify unknown words have been explored recently. In our case, a morphological rule represents a pattern comprising a monosyllabic prefix or suffix and the category of a meaningful stem. The pattern can be formed from various known words. For instance, “總經理”(總(*general*)經理(*manager*), general manager), “總教練”(總(*general*)教練(*coach*), head coach), and “總司令部”(總(*general*)司令部(*headquarters*), headquarters) are known words listed in the CKIP lexicon [15]. These known words generate a pattern, called a morphological rule, which comprises the monosyllabic prefix “general” and category “Na” of a meaningful stem. Since the string “總領隊” can match the morphological rule, it can then be called an unknown word.

Table 3.1 Part of the Morphological Rule Base

Type	Affix	The category of stem	Frequency
Prefix	總(General)	Na	15
Prefix	副(Assistant)	Na	22
Suffix	局(Department)	Na	47
Suffix	室(Room)	VC	18

Table 3.1 shows some of the morphological rules generated by the proposed algorithm. To accelerate the matching procedure, all morphological rules can be initially generated from lexica.

### 3.2.4 Comparison with Previous Method

The difference between the performances of previous PLR-based method and our proposed method derives from their assumptions. Previous method [28] bases on an assumption: if two substrings of a string always occur together, the string is identified probably as an unknown word or phrase. The assumption could function well in universal and large corpora, but the familiar strings in a domain-specific corpus will be treated as unknown word based on the assumption. For example, the string “多同學” is meaningless but quite familiar in the corpus on theme “class recess”. Both frequency and PLR of string “多同學” satisfy the conditions of the previous method and is therefore classified as an unknown word. Although the method uses a purification procedure to further sieve out some of the meaningless strings, it is still not sufficient. For example, the string “多同學” could not be filtered out because it is the substring of various strings such as “很多同學”(many classmates) and “許多同學”(numerous classmates) and does not meet the criterion of the purification procedure. Therefore, numerous familiar and meaningless strings would be extracted by the PLR-based methods.

Our proposed method bases on different assumption: if a string appears much less often than any of its substring, the string would not be classified as an unknown word. Our proposed method does not classified the string “多同學” as a unknown word because its substring “同學” (classmates) appears much often. The experiments [8] show that the SPLR method could deal with the difficult issue of familiar and meaningless strings.

Fig. 3.6 shows how the PLR and SPLR procedures affect the performance of the extraction methods. Although the PLR approach can efficiently extract unknown words from a small corpus, it also misclassifies many meaningless strings as unknown words. Furthermore the purification process can only discard a small number of meaningless strings; many other meaningless strings will survive the process and remain as unknown word classification. By contrast, the SPLR approach not only effectively extracts unknown words but also discards many meaningless strings. Filtering procedure will further exclude more meaningless strings. The fact that SPLR can efficiently eliminate many meaningless strings greatly helps to increase the performance of extracting unknown words in a domain-specific small corpus.

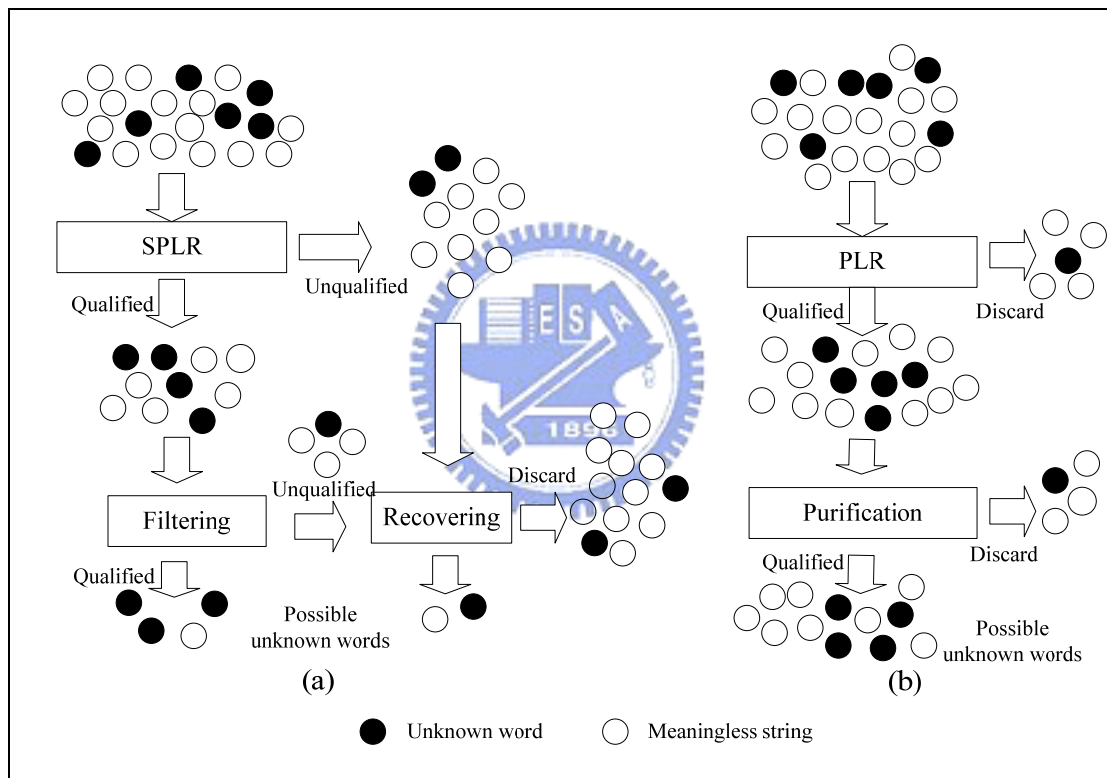


Fig. 3.6 Comparison the Proposed Method with [28]

### 3.3 Unknown Word Classification

The classifying phase of the proposed method shown in Fig. 3.5 comprises two procedures, namely morphological and contextual rule classification. The unknown words that can be identified by the recovery procedure in the extracting phase are classified through the procedure of morphological rules, while the remaining unknown words are classified through the procedure of contextual rules.

Unknown concepts are extracted by the recovering and filtering procedures

respectively. For unknown concepts extracted from recovering procedure, morphological rules are used to infer their categories. For unknown concepts extracted from filtering procedure, contextual rules from corpora are created to compute the degrees of the likelihood of an unknown concept conditioned on categories, denoted as LD. Based on the degrees, the category of the unknown concept can be determined by the current tagging method [40].

### 3.3.1 *Inferring Category Based on Morphological Rules*

The morphological rules comprise a monosyllabic prefix or suffix and a category of meaningful stem. Since morphological rules are generated from known words in lexica, the categories of the known words can be used to infer the category of an unknown word according to the rules. For instance, a morphological rule comprises a prefix “general” and a category of general noun, as revealed by such known phrases as “總經理” (general manager), “總教練”(head coach) and “總司令部” (headquarters). These known words all have a general noun as their category, hence the unknown word “總領隊”(general leader) can be inferred as a general noun. However, the known words that generate the same rule do not necessarily have the same category, as shown in Table 3.2 which includes two categories of known words associated with a morphological rule. Additionally, the stem of an unknown word can have multiple categories, complicating the inference process. In these cases, the Hidden Markov Model (HMM) can be used to infer the category of the unknown words.

Table 3.2 An Example for Morphological Rules Containing Two Categories

type	affix	POS of stem	POS of word	frequency
suffix	界 (the world of)	Na	Na	6
			Nc	40

An HMM-based method [40] uses the lexical information of a word, which includes the probabilities of the word conditioned on categories, to infer the category of the word. Because unknown words have no lexicon information, the probability of the unknown word on category  $c_i$  can be estimated from both the frequency of the stem on category  $c_i$ , and the frequency of the rule that infers category  $c_i$ . The degree of the likelihood of unknown word  $w$  conditioned on category  $c_i$ , denoted as LD, can

be computed as follows.

$$LD(w|c_i) = \frac{tf(b|c_i)}{tf(b)} \cdot \frac{\sum_{m \in S} tf(m|c_i)}{\sum_{m \in S} tf(m)}, \quad m \in S \quad (3.4)$$

where  $b$  represents the stem of unknown word  $w$ ;  $S$  represents the set of the rules matched by  $w$ , and function  $tf$  represents the parameter frequency. Once the LDs of unknown word  $w$  on different categories are available, the tagging method [40] can treat the LDs as the lexicon information of  $w$ , and thus infer its category.

### 3.3.2 Inferring Category Based on Contextual Rules

Unknown words that do not match any morphological rule can be classified based on their contextual information in both universal and tagged corpora and the corpora in our domain. Consider a sequence of words  $p, t, s$  in documents where  $t$  represents an unknown word,  $p$  represents a preceding known word, and  $s$  represents a succeeding known word. A form is defined as  $p*s$ , where  $*$  represents a wild card representing context cue associated with  $t$ . In the tagged corpora, the form may appear in many different locations, or may not appear at all, while the wild card may match various words or none at all. If these words are tagged with a common category, then the unknown word can be assigned to the same category. For instance, unknown word “空蕩蕩” appears on the following sentence in documents:

原本 空蕩蕩 的 球場 被 大家 的 歡笑聲 填滿  
(The originally empty field is filled with people's laughter.)

The form of “empty” includes preceding word “原本”(originally), wild card and succeeding word “的” (of). In corpus CKIP, the form matches the following fragments: ”原本 優異(excellent) 的”, “原本 光禿禿(bare) 的”, and “原本 追漲 (Overheated) 的”. Since the words, matched with the wild card, “excellent”, “bare”, and “overheated” are all tagged with category “VH”, the word “empty” is classified as category “VH”.

The inference is reliable and reasonable, but two exceptions should be considered. First, the middle words of a form may belong to different categories. Second, the form of an unknown word may not appear in tagged corpora. To integrate these observations, Formula (3.5) is developed to estimate the degree of the likelihood



of an unknown word conditioned on categories and associated with a form, denoted as FLD. Assuming the form  $m$  of unknown word  $w$ , the FLD of  $w$  conditioned on category  $c_i$  is as follows:

$$FLD(w_m | c_i) = \alpha \frac{ff(m | c_i)}{ff(m)} + (1 - \alpha) \frac{ef(m | c_i)}{ef(m)} \quad (3.5)$$

where  $ff(m)$  represents the number of form  $m$  in a universal corpus;  $ef(m)$  represents the number of the extension of  $m$  in the corpus;  $ff(m | c_i)$  represents the number of  $m$  conditioned on category  $c_i$  in the corpus. The extension of form  $p*s$  comprises three types,  $c(p)*s$ ,  $p*c(s)$  and  $c(p)*c(s)$ , where  $c(p)$  and  $c(s)$  represents the category of  $p$  and  $s$ , respectively. For instance, the format comprising preceding category “ADV” and succeeding word “的” (of) is one of the extensions of the form comprising preceding word “原本” (originally) and succeeding word “的”(of). The second term of Formula (3.5) is designed to be a major factor only when the form seldom appears in tagged corpora, so the value of coefficient  $\alpha$  in Formula (3.5) should be very close to 1.

The FLDs of an unknown word conditioned on categories can be obtained from a form by Formula (3.5). Furthermore, the unknown word may appear many times and in various ways in domain-specific corpora, so could have many forms. The LD of unknown word  $w$  conditioned on category  $c_i$  can be estimated from the FLDs of  $w$  associated with various forms as follows:

$$LD(w | c_i) = \sum f(m) \cdot FLD(w_m | c_i), m \in F_w \quad (3.6)$$

where  $F_w$  represents a set of the forms of  $w$ , and  $f(m)$  represents the number of occurrences of form  $m$  in the corpora.

Like the LDs yielded by Formula (3.4), the LDs computed by Formula (3.6) can be used as the lexical information of an unknown word, which then enables an HMM-based method [40] to classify the unknown word into a unique category.

### 3.4 Thematic Subconcept Hierarchy

The thematic subconcepts in general represent familiar subconcepts in essays on a theme while the hierarchy formed by thematic subconcepts is used to extract concepts in C-L structures. A subconcept may play different roles in different themes. For instance, subconcept “concert” is difficult to associate with the theme “recess at

school” while the “concert” is easy to associate with the theme “activity on holiday”. For purpose of describing the hierarchy, we define the subordinate subconcept as a subconcept which is not easy to associate with a theme. The subordinate subconcept is often used to specialize other subconcepts which are called superordinate subconcepts.

A superordinate subconcept may be a subordinate subconcept of another subconcept. For example, on theme “recess at school”, subconcept “conversation” is the superordinate subconcept of “concert” and the subordinate subconcept of “classroom”. Hence, the category of a subconcept is determined by theme as well as other subconcepts.

Based on the superordinate-subordinate relations between subconcepts, all subconcepts can form a hierarchy relation. Our proposed methods will generate an asymmetrical semantic relation matrix to represent the superordinate-subordinate relations, and then use the matrix to construct a thematic subconcept hierarchy.

#### ***3.4.1 Asymmetrical Semantic Relation Matrix***

The association, describing the relationship between words in two subconcepts, is computed in a passage consisting of a fixed number of sequential sentences. The co-occurrence of two words in a passage is a good indicator for association since both words can be considered to address the same semantic or topic. Furthermore, the degree of association between the two words [42][54][55] is often measured and determined by both frequency and distance of the co-occurrence of two words.

In this thesis, the distance is defined as the number of the sentences between the two words. Fig. 3.7 illustrates the concepts of the co-occurrence and the distance of two keywords where the distance between word  $k_2$  and  $k_3$  is 1, and the distance between word  $k_3$  and  $k_4$  is 3. Since the length of the passage in Fig. 3.7 is set to be 3, word  $k_2$ ,  $k_3$ , and  $k_4$  are co-occurrence words in the passage, but word  $k_1$  and  $k_4$  are not.

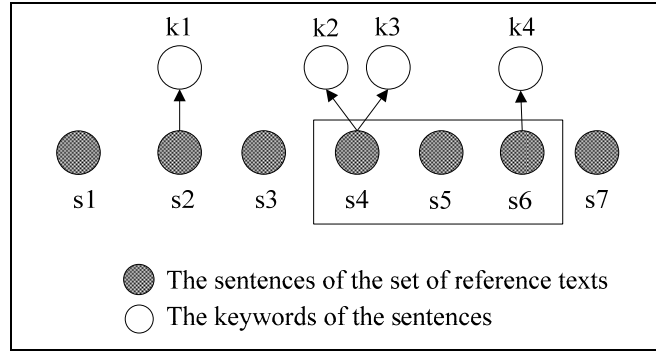


Fig. 3.7 An Example for Estimating Association among Words

Based on the concepts of both co-occurrence and distance, our method will generate a matrix, denoted as asymmetrical semantic relation matrix, to record the degrees of the association among cue words. First, all of cue words are retrieved from training corpus. Assume there are  $n$  reference words in a corpus and relative semantics matrix  $R$  is a  $n \times n$  matrix. The element  $r_{i,j}$  of matrix  $R$  represents the association degree of  $i^{th}$  word  $w_i$  to  $j^{th}$  word  $w_j$ , and is computed as following:

$$r_{i,j} = \frac{\sum_{t \in T} \sum_{p \in t} occ(w_i, w_j)}{freq(w_i)} \quad (3.7)$$

where  $freq(w_i)$  is the number of occurrences of word  $w_i$  in the training corpus;  $t$  is the text in the corpus  $T$ ;  $p$  is a text segment in the text  $t$ , and  $occ(w_i, w_j)$  is

$$occ(w_i, w_j) = \begin{cases} \frac{1}{dist(w_i, w_j)}, & \text{where the words } w_i \text{ and } w_j \text{ both exist in } p. \\ 0, & \text{otherwise} \end{cases} \quad (3.8)$$

where  $dist(w_i, w_j)$  is the distance between words  $w_i$  and  $w_j$  in  $p$ .

### 3.4.2 Constructing Thematic Subconcept Hierarchy

Because of the asymmetric of matrix  $R$ ,  $r_{i,j}$  and  $r_{j,i}$  in  $R$  may be different. If  $r_{i,j}$  is high for the two words  $w_i$  and  $w_j$ , it means that  $j$  always occurs whenever  $i$  occurs, which allows us to use  $w_j$  to represent  $w_i$ . On the other hand, if  $r_{i,j}$  is low, then there is no reason for  $w_j$  to represent  $w_i$ . Based on above discussion, there are three relations between words  $w_i$  and  $w_j$ . First, the relation between  $w_i$  and  $w_j$  is coordinate each other if both  $r_{i,j}$  and  $r_{j,i}$  are high. Second,  $w_i$  is the subordinate word of  $w_j$  if  $r_{i,j}$  is

much higher than  $r_{j,i}$ . Third, there is no relation between  $w_i$  and  $w_j$  if both  $r_{i,j}$  and  $r_{j,i}$  are low. The three relations among words can be used to construct a thematic subconcept hierarchy.

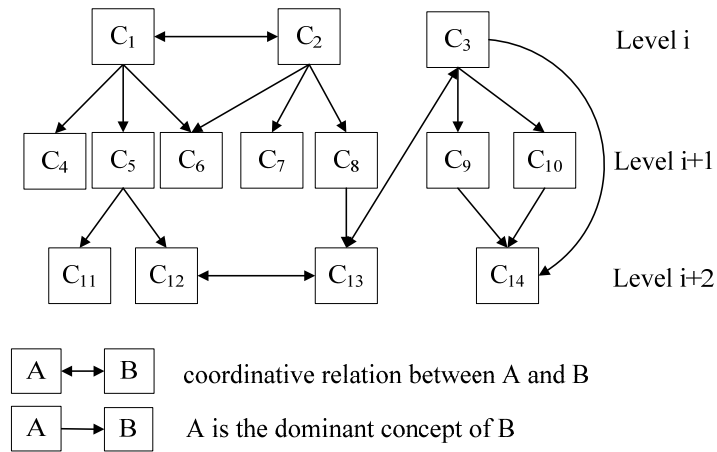


Fig. 3.8 Part of a Thematic Subtopic Hierarchy

A thematic subconcept hierarchy shown in Fig. 3.8 can be constructed by the algorithm shown in Fig. 3.9. There are two characteristics in thematic subconcept hierarchy. First, each subconcept serves as the subordinate subconcept of at least one superordinate subconcept. Secondly, the levels of two subconcepts on a path represent the relative relation between them, i.e.  $c_1$  is the superordinate subconcept of  $c_{11}$ . The thematic subconcept hierarchy constructed from all training essays can be regarded as the collective opinion of writers about the knowledge structure of a theme.

**Determining Level and Related Subconcepts**(relative semantics matrix R)**variable:**

$r_{i,j}$ : the entry of matrix R which represents the relation between word i and j

$M_i$ : a set of word, degree and attribute of relation words for word i

$a_{i,j}$ : the attribute of word j relative to word i which is one of superordinate, subordinate correlation and processed.

THr: threshold of relations which is effective

THk: maximum of the number of relation words in set  $M_i$

THd: maximum of the deepness of concept hierarchy

**main** {

/\* step1: get k relative words of each word in which relationship is the highest top k

**for** i:=1 **to** n

**for** j:=1 **to** n

**if** ( $r_{i,j} + r_{j,i}$ ) > THr **then** {

**if** ( $r_{i,j} > THk$ ) **or** ( $k >$  the number of elements in  $M_i$ ) **then** {

**if** ( $r_{i,j} < r_{j,i} + e$ )     **then**  $a_{i,j} :=$  subordinate

**else if** ( $r_{i,j} > r_{j,i} + e$ ) **then**  $a_{i,j} :=$  superordinate

**else**  $a_{i,j} :=$  correlation

        remove word d from  $M_i$  if  $r_{i,d}$  is smallest relation degree in  $M_i$

        add word j,  $a_{i,j}$  and  $r_{i,j}$  to  $M_i$

        THr :=  $r_{i,j}$

      }

    }

/\* step2: given the level of  $M_i$

processing\_level := 0

**repeat**

**for** i:=1 **to** n {

**if** the attributes of all words in  $M_i$  are subordinate, correlation or processed

**then** the level of  $M_i :=$  processing\_level

  }

**for** j:=1 **to** n {

**if** (the level of  $M_j$  is unknown) **then**

**if** (word  $p \in M_j$ ) **and** (the level of  $M_p$  has been given)

**then**  $a_{j,p} :=$  processed

    }

  processing\_level := processing\_level+1

**until** (processing\_level > THd) or (there is no subconcept whose level is assigned in this cycle.)

the level of sets M will be assigned with processing level if its level is unknown

**return**

}

Fig. 3.9 Algorithm for Determining the Levels of Subconcepts

## Chapter 4 Selection of Concepts

Given a theme, a writer must use concepts and subconcepts to present his/her observations, opinions and illustrations for the theme. In general, writers must first select a set of concepts and organize them around the theme. Furthermore, writers often tend to adopt some concepts with which he/she is familiar because there are differences between the writing abilities of writers. For example, on theme “recess at school”, most of writers can describe subconcept “classroom” sufficiently and effortlessly. Conversely, the description of subconcept “campus scenery” is laborious for some writers for lack of observation, literary device or the abilities for organizing concepts and subconcepts on a theme. These writers may either neglect or choose not to use such concepts and subconcepts.

Based on the discussions mentioned above, this thesis makes a simple assumption: some concepts denoted as literary concepts only performed by skillful writers. Given the assumption, it becomes important to extract the literary concepts in essays. Since the definition and extraction of concepts are very difficult, concepts including literary concepts are first transformed into set of subconcepts in this thesis.

However, it is not an easy task to extract literal subconcepts either. Although it is obvious that literary subconcepts usually appear more often in high-scored essays than the low-scored essays, extracting literary subconcepts directly from high-scored essay based on frequency does not work well. This is due partly to small size of training corpora. In small training corpus, some literary subconcepts do not necessarily appear and can not be collected into a set of literary subconcepts. Besides, above method may extract some subconcepts appeared in both low-scored and high-scored essays and treat them as literary subconcepts.

The usage of HowNet, a universal semantic network in Chinese, can overcome above difficulties derived from small training corpora. HowNet constructs its semantic network with two major elements: term and sememe. A term corresponding to subconcept uses one or more sememes to represent its semantics while various terms may share some sememes. For instance, term “school” consists of sememes “place”, “education”, “learning” and “teaching” while term “teacher” consists of “education” and “person”. The task of extracting sememes will be easier than that of extracting literary subconcepts. For example, subconcept “school” appears in small training corpus while subconcept “teacher” does not. But, sememe “education” of

subconcept “teacher” will be identified because of subconcept “school”. It indicates that a set of sememes can be obtained from small training corpus.

Based on the discussion mentioned above, our proposed method will employ literary sememes to score essays. Below, Subsections 4.1 and 4.2 will introduce our methods for extracting literary sememes. Subsection 4.3 shows the performance of using literary sememes to score essay.

#### 4.1 Set of Literary Sememes

Literary sememes are defined as the sememes which frequently occur in higher-score essays but do not occur in lower-score essays. The degree of the literature of sememe  $s$  and the reliability of its degree, denoted as  $d(s)$  and  $r(s)$  respectively, are shown as follows.

$$d(s) = \frac{\|H\|}{\|L\|}, \quad r(s) = \frac{\|A\|}{\|T\|} \quad (4.1)$$

where  $\|T\|$  represents the number of the essays in training corpus;  $\|A\|$  represents the number of the essays in which sememe  $s$  occurs;  $\|H\|$  represents the number of the high-scored essays in which sememe  $s$  occurs;  $\|L\|$  represents the number of the low-scored essays in which sememe  $s$  occurs. Obviously, higher  $d(s)$  represents sememe  $s$  seldom occurs in low-scored essays and higher  $r(s)$  represents the value of  $d(s)$  is reliable.

Based on Formula (4.1), various sets of sememes can be created. A sememe set  $CS_i$  can be defined as follows

$$CS_i = \{s \mid d(s) \geq D_i \text{ and } r(s) \geq R_i, s \in S\} \quad (4.2)$$

where  $D_i$  and  $R_i$  respectively represent the thresholds of being literary sememes.

Different values of  $D_i$  and  $R_i$  will generate different sets  $CS_i$ . The set of literary sememes on a theme is defined as the  $CS_i$  which yields highest accuracy for scoring essays when the scoring is based on the number of the occurrences of literary sememes in the essay. Next subsection will discuss an automatic procedure to determine the set of literary sememes from the sets of sememes denoted as candidates.

## 4.2 Extraction of Literary Sememes

An evaluation procedure consisting of three steps is designed for determining set of literary sememes. First step is to use sememes in a candidate to calculate the numbers of the sememes in a scored essay. Second step is to employ the numbers of sememes in an essay to predict the score of the essay. Third is to use the predictive scores and an evaluation function to calculate the accuracy of the candidate for scoring essays. The three steps in above procedure will be discussed in following subsections.

### 4.2.1 The Correlation between Candidate and Essay

Given a candidate for the set of literary sememes, a numerical set is first generated from the candidate and the training essays. Assume candidate set  $C$  consists of sememes  $c_1, c_2, \dots, c_n$  and the training corpus  $T$  contains essays  $e_1, e_2, \dots, e_k$ . The degree of correlation between  $C$  and essay  $e_j$  can be computed using the following evaluation function.

$$h(e_j) = \sum_{i=1}^n \text{freq}(c_i | e_j) \quad (4.3)$$

where  $c_i \in C$  and  $\text{freq}(c_i | e_j)$  represents the number of the occurrence of sememe  $c_i$  in essay  $e_j$ . Furthermore, let  $R$  be the set consisting of the degree of correlation between  $C$  and each essay in  $T$ :

$$R = \{a_j \mid a_j = h(e_j), e_j \in T, 1 \leq j \leq k\} \quad (4.4)$$

Next, the following set  $M$  is obtained by sorting the elements in set  $R$ .

$$M = \{m_p \mid m_p \in R, m_{p-1} \leq m_p \leq m_{p+1}, 1 \leq p \leq k\} \quad (4.5)$$

### 4.2.2 Using Candidate Sets to Score Essays

The candidate set  $M$ , which is an ordered set, will be divided into several ordered subsets  $M_i$  which satisfy the following three conditions:



$$\begin{cases} (1) 1 \leq i \leq g \\ (2) \forall m_a \in M_i \text{ and } m_b \in M_{i+1}, m_a < m_b \\ (3) \|M_i\| = \|\{e_j \mid e_j \in T \text{ and the essay } e_j \text{ scored } i \text{ by human}\}\| \end{cases} \quad (4.6)$$

where  $g$  represents the maximum of score points;  $T$  represent the training corpus.

Now, the predictive score of an essay  $e_k$  based on literary sememes can then be assigned by the following formula:

$$gd(e_k) = i, \text{ if } h(e_k) \in M_i \quad (4.7)$$

In other words, the system will assign the score  $i$  to the essay if the number of sememes of the essay is within the range of  $M_i$ .

### 4.2.3 Estimating the Performance of Candidates Quantitatively

Using Formula (4.7), every essay in training corpus can obtain a predictive score from a candidate. The difference between the predictive and human scores of essays represents the performance of using the candidate set to score essays. Table 4.1 shows the performance table of a candidate set in which score points is between 1 and 6. The entry  $n_{i,j}$  in Table 4.1 represents the number of essays of which  $i$  and  $j$  respectively represent the human score and predictive score.

Table 4.1 A Performance Table

Human score	Predictive score					
	1	2	3	4	5	6
1	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	$n_{1,4}$	$n_{1,5}$	$n_{1,6}$
2	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	$n_{2,4}$	$n_{2,5}$	$n_{2,6}$
3	$n_{3,1}$	$n_{3,2}$	$n_{3,3}$	$n_{3,4}$	$n_{3,5}$	$n_{3,6}$
4	$n_{4,1}$	$n_{4,2}$	$n_{4,3}$	$n_{4,4}$	$n_{4,5}$	$n_{4,6}$
5	$n_{5,1}$	$n_{5,2}$	$n_{5,3}$	$n_{5,4}$	$n_{5,5}$	$n_{5,6}$
6	$n_{6,1}$	$n_{6,2}$	$n_{6,3}$	$n_{6,4}$	$n_{6,5}$	$n_{6,6}$

Third step of the evaluation is to estimate the performance of a candidate using the above performance table and a weighted table shown in Table 4.2. The size of weighted table is the same as that of performance table and the values in weighted table are related to the difference between human score and predictive score. In general, a higher weight corresponds to a smaller difference and weight one

corresponds to the lowest acceptable difference.

Table 4.2 A Weighted Table

Human score	Predictive score					
	1	2	3	4	5	6
1	2	1	0	-1	-2	-3
2	1	2	1	0	-1	-2
3	0	1	2	1	0	-1
4	-1	0	1	2	1	0
5	-2	-1	0	1	2	1
6	-3	-2	-1	0	1	2

Performance table and weighted table can be used to estimate the performance of a candidate. Assuming candidate set  $C$  generates a performance table. The performance of  $C$  can be estimated by Formula (4.8).

$$perf(C) = \sum_i \sum_j w_{i,j} \times n_{i,j} \quad (4.8)$$

where  $w_{i,j}$  represents a weight in a weighted table and  $n_{i,j}$  represents the value of a entry in the performance table.

Formula (4.8) can calculate the performance of every candidate. The candidate set which obtains highest value of the performance of all candidate sets will be regarded as the set of literary sememes for the theme. Using the set of literary sememes and its subsets derived from Formula (4.6), a test essay can be scored by Formula (4.7).

### 4.3 Usefulness of Literary Sememes for Scoring Essays

A set of essays is employed to evaluate the performance of using literary sememes to score essays. Subsection 4.3.1 describes the corpus and the evaluation which are also used in Subsection 5.4 and 6.4. Subsection 4.3.2 discusses experimental results.

#### 4.3.1 Experimental Corpus and Definition of Performance

The experimental corpus consists of 689 essays written by students from the eighth grade. The theme of the essays was “Recess at School”. The score of each essay ranged from one point to six points, where a higher point represented a higher quality. The score of an essay was obtained by averaging the scores from two or three

teachers. The numbers of essays corresponding to different scores in the range 1–6 were 45, 128, 210, 208, 91 and 7, respectively.

The performance of a method for scoring essays can be evaluated by four indicators, namely accuracy rate, exact rate, average accuracy rate and average exact rate. Assuming the number of test essays graded 1 to  $n$  by experts is  $k_1, k_2, \dots, k_n$  respectively. In the essays scored  $i$  by expert, the number of essays graded 1 to  $n$  by the method is  $c_{i,1}, c_{i,2}, \dots, c_{i,n}$  respectively. The four indicators are defined as follows.

$$\text{Accuracy rate} = \frac{\sum_{i=1}^n \sum_{j=1}^n a_{i,j}}{\sum_{i=1}^n \sum_{j=1}^n c_{i,j}}, \text{ where } a_{i,j} = \begin{cases} c_{i,j}, & \text{if } |i-j| \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (4.9)$$

$$\text{Exact rate} = \frac{\sum_{i=1}^n \sum_{j=1}^n e_{i,j}}{\sum_{i=1}^n \sum_{j=1}^n c_{i,j}}, \text{ where } e_{i,j} = \begin{cases} c_{i,j}, & \text{if } |i-j| = 0 \\ 0, & \text{otherwise} \end{cases} \quad (4.10)$$

$$\text{Average accuracy rate} = \frac{1}{n} \sum_{i=1}^n \frac{\sum_{j=1}^n a_{i,j}}{\sum_{j=1}^n c_{i,j}}, \text{ where } a_{i,j} = \begin{cases} c_{i,j}, & \text{if } |i-j| \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (4.11)$$

$$\text{Average exact rate} = \frac{1}{n} \sum_{i=1}^n \frac{\sum_{j=1}^n e_{i,j}}{\sum_{j=1}^n c_{i,j}}, \text{ where } e_{i,j} = \begin{cases} c_{i,j}, & \text{if } |i-j| = 0 \\ 0, & \text{otherwise} \end{cases} \quad (4.12)$$

High accuracy rate and exact rate represent the method performs well in whole test corpus. Of test corpus, middle-score essays are often the majority and high-score and low-score essays are the minority. Average accuracy rate and average exact rate can be used to further analyze the difference between the performances of different methods.

### 4.3.2 Performance of Literary Sememes for Scoring Essays

In the experiment, 343 training essays and 346 test essays are randomly chosen from the corpus. The proposed method finally selected 383 literary sememes from 1016 sememes in training essays. The test essays, in which the number of literary sememes is less than 2, are scored 1 point by the proposed method. The test essays, in which the number of literary sememes is between 3 and 6, are scored 2. The test essays, in which the number of literary sememes is between 7 and 13, are scored 3. The test essays, in which the number of literary sememes is between 14 and 23, are scored 4. The test essays, in which the number of literary sememes is between 24 and 43, are scored 5. The test essays, in which the number of literary sememes is higher than 43, are scored 6.

Table 4.3 shows the relationship between human scores and the scores graded by the proposed method. The entries in Table 4.3 represent the number of essays which are graded the human score by experts and graded the predictive score by machine, e.g. there are 35 essays which are graded 2 by experts and graded 3 by machine. In Table 4.3, the accuracy rate and exact rate are 0.916 and 0.442, respectively. The result shows that the occurrence of literary sememes in an essay is highly related to the score of the essay.

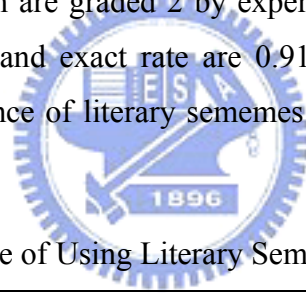


Table 4.3 Performance of Using Literary Sememes to Scoring Essays

human score	predictive score					
	1	2	3	4	5	6
1	7	10	6	0	0	0
2	4	20	35	5	0	0
3	1	15	62	26	1	0
4	0	4	45	51	4	0
5	0	0	9	24	13	0
6	0	0	0	3	1	0

## Chapter 5 Connection of Concepts

This section will propose a method which uses the similarity between the connections of concepts in different essays to predict the scores of essays. The method consists of three phases shown in Fig. 5.1. First phase is to transform essays in training corpus and test essay into C-L structures relied on thematic subconcept hierarchy. Second phase is to measure the similarities between the C-L structures of test essay and that of every training essay. Third phase is to score test essay with the result of measuring similarity by a scoring method.

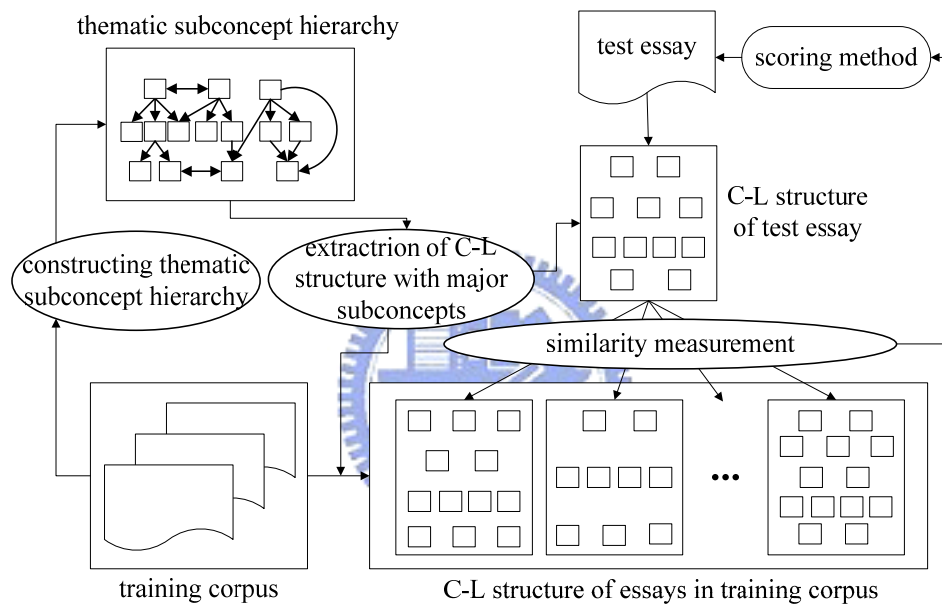


Fig. 5.1 Methods for Predicting Scores Using the Connection of Concepts

### 5.1 Extraction and Transformation of the Concepts in C-L Structure

A paragraph often includes one or more concepts which compose several subconcepts respectively. Some of the subconcepts are the major subconcepts of concepts and others are the assistants of the major subconcepts. Because assistant subconcepts influence the accuracy of measuring similarity between C-L structures of essays, it is necessary that extracting major subconcepts from subconcepts to represent concepts in C-L structures.

Basically, major subconcepts place in higher part of thematic subconcept hierarchy while the assistants place in lower part of the hierarchy. Based on the observation, major subconcepts of a paragraph can be extracted via three steps. First step is to extract all subconcepts of the paragraph. Second is to assign the level

number of most abstract subconcept in the paragraph to minimum level of major subconcepts. The reference level of major subconcepts equals the minimum level plus a fault tolerate coefficient. Third is to extract the subconcepts to be major subconcepts whose level numbers are less than the reference level. Below, major subconcepts will be treated as the representatives of concepts in C-L structure.

In addition, because C-L structures consist of inter-paragraph connections and intra-paragraph connections, Subsections 5.2 and 5.3 will discussed the similarity measurements and scoring methods of the two connections respectively.

## 5.2 Inter-paragraph Connections

Fig. 5.2 shows the inter-paragraph connections of an essay in which major subconcepts are employed to represent concepts. The inter-paragraph connections can be treated as the conjunction of various concept sequences. For instance, the inter-paragraph connections in Fig. 5.2 can be transformed to the conjunction of twelve sequences (c1,c3,c6), (c1,c3,c7), (c1,c4,c6), (c1,c4,c7), ..., (c2,c5,c6) and (c2,c5,c7). The concept sequences are denoted as “C-chains”. Since different appearance orders of concepts in C-L structure represent different semantics, a C-chain is treated as part of thematic semantics. In addition, because the concepts in C-chains only contain a one-way connection, similarity measure between C-chains is relatively easier than that between inter-paragraph connections. Hence, the issue about similarity measure between two inter-paragraph connections will be transformed to similarity measure between the C-chains in two structures.

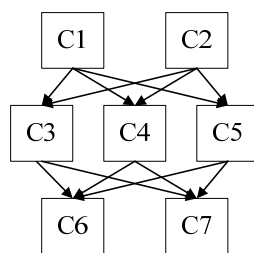


Fig. 5.2 Illustration for C-chains

### 5.2.1 Similarity Measure between Inter-paragraph Connections

Fig. 5.3(a) represents a test essay consisting of four paragraphs and twelve concepts. Paragraph 1 to 4 respectively contains three, two, four, three concepts. Fig. 5.3(b) shows the test essay which shares six of C-chains to a corresponding essay in

Fig. 5.3(b). The C-chains in a test essay are denoted as “T-chains”; The T-chains shared with a corresponding essay are denoted as “P-chains”.

In Fig. 5.3(b), the six P-chains are sound. However, most essays do not appear many complete P-chains because the number of paragraphs in the essays differs to that in test essay. Hence, three variations of P-chains, denoted as “S-chains”, should be discussed. First variation is the sub-chain of a T-chain which lacks several concepts in the head or tail of T-chains. For instance, S-chain (c5,c7,c12) in Fig. 5.3(c) is part of T-chains (c1,c5,c7,c12), (c2,c5,c7,c12) and (c3,c5,c7,c12).

Second variation is the S-chain which lacks middle concepts in T-chains. For example, S-chain (c1,c7,c12) in Fig. 5.3(c) lacks middle concepts c4 and c5 corresponding to T-chains (c1,c4,c7,c12) and (c1,c5,c7,c12). Third variation is the S-chain which occurs on serial paragraphs in a test essay but on alternate paragraphs in the corresponding essay. For example, the concepts of S-chain (c5,c7,c10) appear on paragraph 2, 3 and 4 in Fig. 5.3(a) while appear on paragraph 2, 3 and 5 in Fig. 5.3(c).

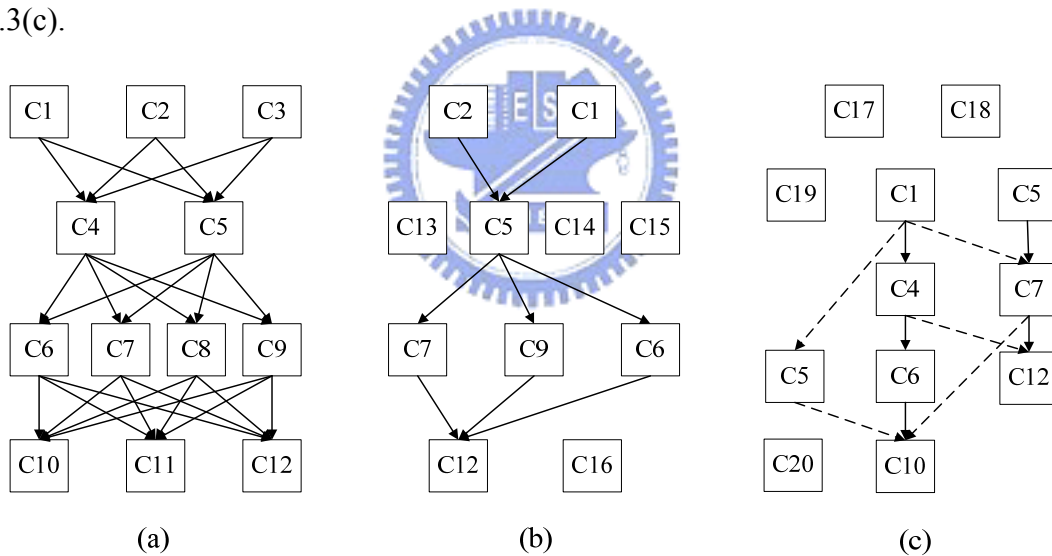


Fig. 5.3 Illustration for P-chains and S-chains

Using the subsets of P-chains can measure the similarity between test essay and corresponding essays. The subsets are classified into sets of consecutive and alternate sub-chains. For example, in Fig. 5.3(c), S-chain (c5,c7,c12) contains three consecutive sub-chains (c5,c7), (c7,c12) and (c5,c7,c12) while S-chain (c5,c7,c10) contain two alternate sub-chains (c5,c7,c10) and (c7,c10) and one consecutive sub-chains (c5,c7). Assume a set P of P-chains is derived from test essay  $t$  and corresponding essay  $c$ , the similarity  $sim(t,c)$  between essays  $t$  and  $c$  is as follows.

$$sim(t, c) = \|A\| + 2 \times \|S\| \quad (5.1)$$

where  $S$  represents the set of consecutive sub-chains of  $P$ ;  $A$  represents the set of alternate sub-chains of  $P$ ;  $\|A\|$  represents the number of the elements in  $A$ .

Equation (5.1) satisfies two principles: (i) the longer S-chains should be weighted first. (ii) if the length of a consecutive sub-chain is same as that of alternate sub-chains, the consecutive sub-chain shall be weighted. The two principles are necessary for the accuracy and precision of measuring similarity. For example, Fig. 5.4(a), (b) and (c) show three S-chains consisting of three concepts. Using Equation (5.1), the S-chain in Fig. 5.4(a) is scored with six to which consecutive sub-chains (c1,c2), (c2,c3) and (c1,c2,c3) contribute two respectively. In the other hand, the total of two S-chains in Fig. 5.4(b) is scored with four while the S-chain in Fig. 5.4(c) is also scored with four to which consecutive sub-chains (c9,c10) contributes two, alternate sub-chains (c8,c9) and (c8,c9,c10) only contributes one respectively.

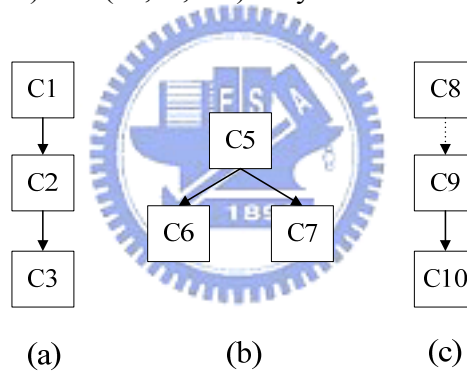


Fig. 5.4 An Example of Principles for Scoring S-chains

After measuring the similarities between test essay and essays in training corpus, a test essay will be scored by the score of the training essay to which test essay is most similar.

### 5.3 Intra-paragraph Connections

Based on the definition in Section Chapter 3, a paragraph of essays is used to describe a subtopic of a given theme and it can correspond to set of concepts in C-L structure. In addition, the occurrence order of these concepts also represents different semantics. The phenomenon that two writers use same concepts in a paragraph refers to two possibilities. First, two writers use different appearance order of the concepts to express different subtopics. Second, two writers try to express same subtopics but



the different writing skills result in different appearance order of same concepts. Both cases indicate that the occurrence order of concepts in a paragraph could be the evaluation of essay quality. Hence, the structure consists of concepts and their occurrence order is defined as conceptual structure of intra-paragraph.

Fig. 5.5(a) shows a conceptual structure of intra-paragraph. The structure composes of five concepts and a appearing sequence (c1,c2,c3,c4,c5). The quality of intra-paragraph conceptual structure of a test essay could be evaluated relied on searching the same structure in training corpus, but it is rare that a structure appears in two essays. Hence, the evaluation of the quality must rely on measuring similarity between the structures in test essay and training essays. Fig. 5.5(b) shows a conceptual structure of intra-paragraph which shares sub-chain (c1,c2,c3,c4) with Fig. 5.5(a). The sub-chain displays that the structure in Fig 4.4(a) is very similar to that in Fig. 5.5(b). For simplify, the sub-chain to which two structures shared is denoted as “R-chain”.

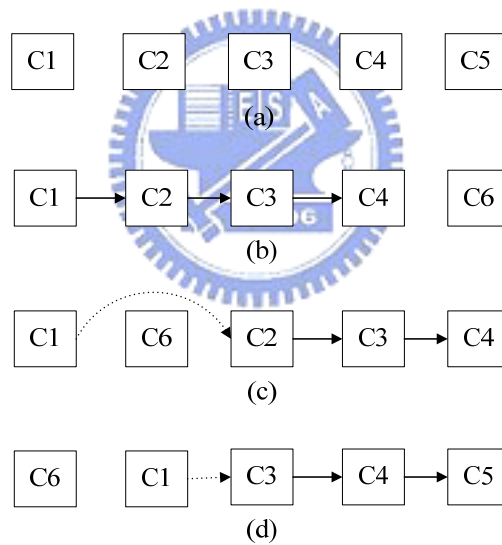


Fig. 5.5 Forms of R-chains

There are three forms of R-chain shown as Fig. 5.5(b) to (d). The structure in Fig. 5.5(c) shares concepts c1,c2,c3 and c4 with that in Fig. 5.5(a). These concepts occur continuously in Fig. 5.5(a) but concept c1 and c2 do not appear continuously in Fig. 5.5(c). Comparatively, the structure in Fig. 5.5(d) shares four concepts with that in Fig. 5.5(a). These concepts occur continuously in Fig. 5.5(d) but do not in Fig. 5.5(a). The connections between c1 and c2 in Fig. 5.5(b) and connections between c1 and c3 in Fig. 5.5(c) are denoted as “weak links”. In R-chains, the appearance of weak links implies two possibilities. First, weak links are still regular connections which are

merely influenced by insignificant concepts. Second, the structure containing weak links and the structure without weak links respectively express different subtopics. Due to the usage of R-chains should concern for the possible errors from weak links, R-chain could not contain weak links too many. In our experience, the number of weak links in R-chain should be smaller than two.

### 5.3.1 Similarity Measure between Intra-paragraph Connections

Three familiar R-chains are shown in Fig. 5.6(b) to (d). According to the discussion about weak links, the R-chain between the structures in Fig. 5.6(a) and (b) should be either (c1,c2,c3) or (c2,c3,c5) since (c1,c2,c3,c5) contains two weak links; That in Fig. 5.6(a) and (c) is either (c3,c4,c5) or (c2,c4,c5), not (c1,c3,c4,c5); and that in Fig. 5.6(a) and (d) should be either (c1,c4,c5) or (c1,c2,c3). Since there are two or more candidates of R-chain sometimes, genuine R-chain can be identified from candidates according to following three rules. First, the candidate which contains maximum concepts should be R-chain. Secondly, if two or more candidates satisfy first rule, candidate without weak links shall be R-chain. Thirdly, if two or more candidates still satisfy both rules 1 and 2, R-chain shall be randomly selected from the candidates.

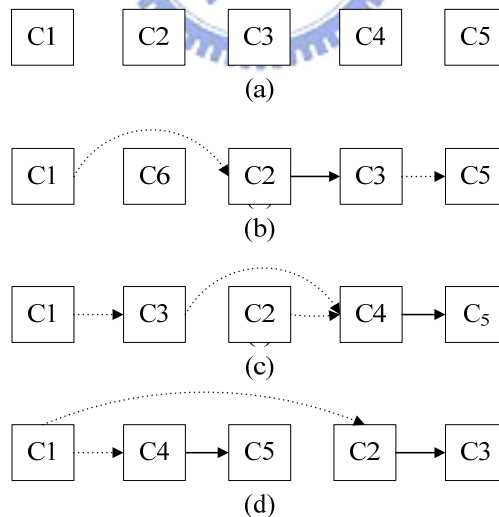


Fig. 5.6 An Example for Comparison between R-chains

R-chain can be used to estimate the similarity between two conceptual structures of intra-paragraph. Assume two structures  $s_1$  and  $s_2$  create a R-chain which consists of  $r$  concepts, the similarity between  $s_1$  and  $s_2$  is as follows.

$$\text{simp}(s_1, s_2) = \frac{C_2^{r-w}}{\|s_1\| \times \|s_2\|} \quad (5.2)$$

where  $\|s_1\|$  and  $\|s_2\|$  represent the number of the concepts in structures  $s_1$  and  $s_2$ , respectively, and

$$w = \begin{cases} 1, & \text{if } s \text{ contains a weak link} \\ 0, & \text{otherwise} \end{cases} \quad (5.3)$$

### 5.3.2 Scoring Essays by Intra-paragraph Connections

In Subsection 5.2, test essay is graded with the score of the training essay to which test essay is most similar. By contrast, since an essay often composes of several conceptual structures of intra-paragraph, the result of similarity measure of a structure in an essay cannot infer to that of every structure in the essay. In addition, an intra-paragraph connection in test essay may be similar to that in several essays which are graded with different scores.

Based on above observations, our proposed method integrates the results of similarity measure of all structures in test essay to score the essay. First, a threshold of similarity is used to examine whether a structure in corresponding essays is similar to that in test essay or not. Since qualified structures may be in exceed of quota, only the structures whose similarities are top  $n$  of all similarities are remained. The structures which pass above examination are collected into a set of similar structures denoted as  $S$ . Then, the score of test essay can be estimated by following equation.

$$\text{score} = \frac{\sum_{s \in S} gd(s)}{\|S\|} \quad (5.4)$$

where  $\|S\|$  represents the size of set  $S$  and  $gd(s)$  represents the score of the essay which contains the conceptual structure  $s$  of intra-paragraph.

## 5.4 Usefulness of C-L Structures for Scoring Essays

Experimental corpus and the definition of performance are the same as that in Subsection 4.3.1. In the corpus, the predictive scores of 445 and 563 essays respectively generated by inter-paragraph connections and intra-paragraph connections are available. Table 5.1 shows the result of using inter-paragraph

connections to score essays. An entry in Table 5.1 shows the number of essays which are graded the human score by experts and graded the predictive score by machine. In Table 5.1, the accuracy rate and exact rate are 0.82 and 0.39, respectively. Table 5.2 shows the result of using intra-paragraph connections to score essays. In Table 5.2, the accuracy rate and exact rate are 0.84 and 0.37, respectively. Experimental results show inter-paragraph connections and intra-paragraph connections are efficient conceptual features for scoring essays.

Table 5.1 Performance of Inter-paragraph Connection

human score	<u>predictive score</u>					
	1	2	3	4	5	6
1	0	2	3	1	0	1
2	0	8	11	18	1	2
3	0	17	55	58	20	2
4	0	8	36	84	34	1
5	0	0	20	28	28	0
6	0	0	1	2	4	0

Table 5.2 Performance of Intra-paragraph Connection

human score	<u>predictive score</u>					
	1	2	3	4	5	6
1	3	5	8	7	0	0
2	5	17	49	34	1	0
3	1	9	94	70	1	0
4	1	12	70	94	4	0
5	0	2	19	56	5	0
6	0	0	3	4	0	0

## Chapter 6 Decoration of Concepts

Sentence patterns and figures of speech are employed to refine and enhance concepts in essays. English AES systems often evaluate the quality of essays based on the variety and versatility of sentence patterns and the occurrence of such patterns as inverted sentence and relative clause. However, Chinese AES systems cannot use the evaluating methods because the definition of Chinese sentences is uncertain and loose. The issue also results in the lack of studies about the relativity between the quality and sentence patterns of essays.

Although methods for identifying sentence patterns in essays cannot be used to Chinese AES, figures of speech are useful for scoring essays. Some studies indicate that the writers who use figures of speech in essays possess better writing skills. Ko [26] notes that the usage of figures of speech in Chinese essays is an important factor in essay scoring. [10][20] states that students' writing skills can be enhanced when they practice or study the usage of figures of speech.

Many studies [22][51][52] have proposed various definitions and classification of figures of speech in Chinese articles. Although the definitions and classifications are varied, the manifestations of the figures of speech “pi-yu” and “pai-bi” are similar to each other. These observations indicate that it is feasible to extract figures-of-speech “pi-yu” and “pai-bi” from essays.

In the other hand, some figures of speech contain both basic and advanced representations. Huang [23] notes that ten syntactic rules for Chinese figures of speech are all included in the textbooks of elementary schools, but Chen's experiments [12] show that the literary “pi-yu” is not used as often as the basic “pi-yu” in sixth grade students' Chinese essays. It implies that a better writer will use an advanced representation of familiar figure of speech skills in essays among many alternatives.

This thesis proposes methods for extracting figures of speech “pi-yu”, “pai-bi” and literary “pi-yu”. Subsection 6.1 will discuss the representation of figure of speech “pi-yu”. Subsection 6.2 presents methods for identifying figure of speech “pi-yu” and literary “pi-yu”. Subsection 6.3 presents methods for extracting figure of speech “pai-bi” in essays.

In addition, although comma in Chinese functions as both comma and period in

English, the issue of ambiguity does not influence the performance of our method. In brief, this section treats a Chinese character sequence ended with comma, period, interrogation, exclamation and semicolon as a sentence.

### 6.1 Building Sets of Connectives and Literary Connectives

Figure-of-speech “pi-yu” makes a comparison between two unlike elements having at least one quality or characteristics in common. There are mainly four subcategories in “pi-yu”: “ming-yu” (明喻), “an-yu” (暗喻), “jie-yu” (借喻) and “lue-yu” (略喻). “Ming-yu” and “an-yu” comprises three elements: tenor, connective and vehicle. For example, in sentence “the campus is similar to a market on recess” (下課時校園就像菜市場), words “campus”, “similar” and “market” stand for respectively the tenor, connective and vehicle. “Ming-yu” and “an-yu” are both similar to simile in English, but “ming-yu” differs from “an-yu” in the degree of relationship between tenor and vehicle using different connectives. Because “ming-yu” and “an-yu” occur in essays with specific patterns, this paper only discusses the two subcategories of “pi-yu”.

Connectives are significant identifiers for retrieving the pattern of “pi-yu”. Based on our observations, the parts-of-speech of connectives could be classified into classificatory verbs and conjunctives, respectively denoted as VG and Caa in [15]. For example, words “變成” and “好像”, which are respectively synonymous to word “become” and “like”, are classificatory verbs. Words “跟” (as) and “和” (as) are conjunctives. Since the classificatory verbs and conjunctives contain very few words in Sinica CKIP lexicon, experts can manually select qualified connectives.

Some of the connectives, e.g. word “如” (similar), almost do not appear in low-score essays, but occur in high-score essays frequently. These connectives, denoted as literary connectives, are found to be seldom used in colloquialism. Based on our observations, literary connectives should be useful for essay scoring.

Formula (6.1) is used to retrieve literary connectives from training data. First essays in training data are divided into a subset of high-score essays and a subset of low-score essays. A literary connective  $w$  is defined to satisfy the following condition:

$$\frac{Hf(w)}{Hf(w) + Lf(w)} \geq \beta \quad (6.1)$$

where  $Hf(w)$  represents the numbers of the occurrence of  $w$  in the high-score subset,  $Lf(w)$  represents that in the low-score subset,  $\beta$  represents a threshold ranged from 0.5 to 1. The higher  $\beta$  value is used, the more discriminatory power the connective has. However, it will result in a small number of literary connectives. Based on our experience from experiments, the best choice of  $\beta$  is 0.6.

## 6.2 Extracting FOS “Pi-yu”

The appearance of connectives can identify two patterns of figure-of-speech “pi-yu”. The first pattern comprises “noun+connective+noun” in single sentence. For instance, the below sentence:

這時候 學校 變成了 一個 嘈雜的 菜市場  
 now Campus become a noisy market  
 (Campus becomes a noisy market now.)

contains the sequence “campus+become+market” which matches the pattern “noun+connective+noun”. Formula (6.2) describes the rule for the first pattern in detail:

$$> (Na | Nb | Nca | Ncb) > \text{Connective} > (Na | Nb | Nca | Ncb) > \quad (6.2)$$

where symbol “>” represents several words or no word, symbol “|” represents logical operator “OR”. Parts-of-speech Na, Nb, Nca, Ncb represent general noun, proper noun, proper place noun and general place noun, respectively.

The second pattern comprises either “connective+adjective+noun” or “connective+noun+adjective” in a single sentence. In addition, it should satisfy two conditions: (i) there is no noun before the connective, (ii) the preceding sentence ends with comma and contains a noun. For example, considering the two adjacent sentences:

校園 充滿 交談的 聲音，就 如 菜市場 般 熱鬧非凡，  
 Campus fill conversation voice as market boisterous  
 (Campus fills with conversation voice, just as a boisterous market.)

in which the preceding sentence end with comma and includes noun ”campus”, and the succeeding sentence includes pattern “connective+noun+adjective” corresponding to

the sequence “as+market+boisterous” and there is no noun before the connective. Formula (6.3) describes the rule of the pattern in detail:

$$> \text{Noun} > , > \text{Connective} > ((\text{Adjective} > \text{Noun}) | (\text{Noun} > \text{Adjective})) > \quad (6.3)$$

where the definitions of symbols “>” and “|” are the same as that in Formula (6.2), “Noun” represents the component (Na | Nb | Nca | Ncb) in Formula (6.2), “Adjective” represents a word whose part-of-speech is denoted as VH or A in [15].

Both rules for “pi-yu” in our proposed method effectively conform to the theoretical structure consisting of tenor, connective, and vehicle. Formula (6.2) often appears in English sentences and short Chinese sentences. Formula (6.3) is a mutation of Formula (6.2) where tenor and vehicle appears on different sentences. This is needed because of elaborated description for the tenor and vehicle.

### 6.3 Extracting FOS “Pai-bi”

Figure-of-speech “pai-bi” uses two sentences or sets of sentences, of which the syntactic structure is similar to each other, to express two concepts of the same property and domain. For example, both sentences “打球 的 打球、散步 的 散步” (Players are playing, walkers are walking.) describe actions in campus using three words and the same syntactic structure: verb following noun. Our proposed method identifies the two single sentences as using the writing skill “pai-bi”.

The following criterion is used to identify if “pai-bi” appears in the essay. If two sentences appearing in a small segment of content contain the same number of words and the same part-of-speech sequence, then the “pai-bi” is considered to occur. For example, in the four serial sentences “到操場走走，可以看到有人悠閒的慢跑；到合作社看看，可以看到有人瘋狂的搶購。” (Some guys are running leisurely on field; some guys are shopping irrationally on snack bar.), the word segmentation and part-of-speech tagging for the first and third sentence is as follows.

到(P) 操場(Ncb) 走走(VA)

到(P) 合作社(Ncb) 看看(VA)

Both sentences consist of three words and the same parts-of-speech of the words. In particular, a preposition, general place noun and verb are in the sequence. Our method hence identifies the occurrence of “pai-bi” in the four serial sentences.



The above example actually shows the delicate aspect of “pai-bi” where the first and the third constitute a usage while the second and fourth also constitute another usage of “pai-bi”. This is an advanced usage of “pai-bi” and is not considered in this study due to its rare occurrence.

#### 6.4 Usefulness for Scoring Essays

Table 6.1 shows how figures-of-speech affect the scores of essays. Row 1 in Table 6.1 shows the ratios of essays to all of the essays in the corpus under different scores. Row 2 shows the ratios of the essays to all of the essays containing the usage of “pai-bi” under different scores. Row 3 shows the ratios of the essays to all of the essays containing the usage of “pi-yu” under different scores. Row 4 shows the ratios of the essays to all of the essays containing the usage of literary “pi-yu” under different scores. The different distribution or spread of the ratios shows that the usage of figure-of-speech in fact affects the score of the essays.

The total ratios in the expanded column of higher score for row 2, 3 and 4 are 0.57, 0.55 and 0.80 respectively while the total ratios for row 1 are 0.37. It shows that the essays using figure-of-speech increase the odds to obtain higher scores. Further, the data from row 2, 3 and 4 shows that the odds are increased if the advanced skill of figure-of-speech is used. In other words, graders trend to grade essays containing advanced writing skills to higher score against common skills.

Table 6.1 The Distributions of the Ratios of Essays to All Essays

	<u>Lower score</u>			<u>Higher score</u>		
	1	2	3	4	5	6
All essays	0.07	0.19	0.31	0.30	0.13	0.01
FOS “pi-yu”	0.03	0.15	0.28	0.35	0.19	0.01
literary “pi-yu”	0.00	0.11	0.19	0.46	0.22	0.03
FOS “pai-bi”	0.03	0.10	0.22	0.47	0.16	0.03

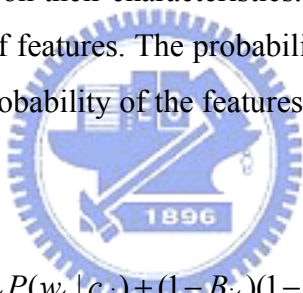
By contrast to selection and connection of concepts, the number and similarity of figures of speech in essays cannot be used alone to predict exact scores of the essays. However, the experimental result indicates that the occurrence of figures of speech will be a useful and important factor for Chinese AES.

## Chapter 7 Performance of Conceptualization for Scoring Essays

Sections Chapter 4, Chapter 5, and Chapter 6 have shown the effectiveness of scoring essays based on individual features of conceptualization of the essays. Although the features are useful for scoring essays, the scores respectively derived from the features may not be consistent. Furthermore, the occurrence of figures of speech in essays cannot be used alone to predict the scores of the essays. A model is needed to integrate the results of these features for scoring essays. Subsection 7.1 discusses a multi-variate Bernoulli model (MBM) for incorporating the features. Subsection 7.2 shows the performance of the model.

### 7.1 Predicting Model

In this thesis, we design an improved multi-variate Bernoulli model to incorporate the features based on their characteristics. Conventional MBM regards an essay as a special case of set of features. The probability of a test essay conditional on a score is the product of the probability of the features conditional on the score, which is shown in Formula (7.1).


$$P(d_i | c_j) = \prod_{t=1}^V \left[ B_{it} P(w_t | c_j) + (1 - B_{it})(1 - P(w_t | c_j)) \right] \quad (7.1)$$

where  $d_i$  represents essay  $i$ ;  $c_j$  represents score  $j$ ;  $B_{it} \in \{0,1\}$  indicates whether feature  $t$  appears in essay  $i$ ;  $V$  represents the number of features;  $P(w_t | c_j)$  represents the probability that feature  $w_t$  appears in an essay scored with  $c_j$ ;  $P(w_t | c_j)$  can be calculated by Formula (7.2).

$$P(w_t | c_j) = \frac{1 + \sum_{i=1}^{D_j} B_{it}}{J + D_j} \quad (7.2)$$

where  $D_j$  is the number of essays in the training corpus scored  $c_j$ ;  $J$  is a constant term which is assigned by 1 in this thesis. Based on Formulae (7.1) and (7.2), essay  $d_i$  can be graded with  $c_h$  which generates the maximum of the probabilities in Formula (7.1).

Conventional MBM assumes that the features of essays either appear or not appear. Of all the features used in the conceptualization, figures of speech satisfy such assumption. For example, among 91 training essays scored 5 points, 36 essays contain figure of speech “pi-yu”. The probability that feature “pi-yu” appears in an essay scored 5 points is therefore  $(1+36)/(1+91) = 0.40$ .

However, features of literary sememes and C-L structures do not satisfy such assumption since they can be derived from all of the training essays. It indicates that the features are useless in Formula (7.1). Given such observation, a new approach consisting of two steps is used to yield  $P(w_t | c_j)$  in Formula (7.1). First step is to use the score  $u$  of essay  $i$  graded by feature  $t$  to compute  $P(w_t | c_j)$  of essay  $i$  in Formula (7.1), which is shown in Formula (7.3).

$$P(w_t | c_j) = \frac{1 + K_{ju}}{J + \sum_{r=1}^n K_{jr}} \quad (7.3)$$

where  $K_{jx}$  represents the number of essays which are scored  $j$  by experts scored  $x$  by using feature  $t$ ;  $n$  represents the number of scores.

Formula (7.4) is used to compute  $P(w_t | c_j)$  of essays which cannot be scored by using feature  $t$ .

$$P(w_t | c_j) = \frac{1 + U_{jt}}{J + D_j} \quad (7.4)$$

where  $U_{jt}$  represents the number of essays which are scored  $j$  by experts and cannot be scored by using feature  $t$ .

For instance, in the training corpus, 91 training essays are scored with 5 by experts. 76 essays which can be scored by using feature inter-paragraph connections and 15 essays cannot be scored. For reader’s convenience, feature inter-paragraph connections is denoted as  $es$  herein. Of the 76 essays, 28 essays are graded 4 points by using feature  $es$ . Hence, in Formula (7.1), the  $P(w_{es} | c_5)$  of the test essay scored 4 points by using feature  $es$  is  $(1+28)/(1+76) = 0.38$ ; the  $P(w_{es} | c_5)$  of the test essay which cannot be scored by using feature  $es$  is  $(1+15)/(1+91) = 0.17$ .

Second step is to smooth the probabilities derived from Formulae (7.3) and (7.4). In general, the  $P(w_t | c_j)$  of a test essay which is scored with  $u$  by feature  $t$  should satisfy the three following constraints:

$$\left\{ \begin{array}{l} \max_j P(w_t | c_j) = P(w_t | c_u) \\ \forall j < u, P(w_t | c_j) \geq P(w_t | c_{j-1}) \\ \forall j > u, P(w_t | c_j) \geq P(w_t | c_{j+1}) \end{array} \right. \quad (7.5)$$

However, the values of few probabilities cannot satisfy the constraints because of small size of training corpus. Hence, the inconsistent probabilities will be revised by using interpolation and extrapolation. For example, in the training corpus, the probabilities of feature  $es$  corresponding to scores 1 to 6 are 0.14, 0.45, 0.38, 0.52, 0.37 and 0.29 respectively. The probabilities satisfy the constraints in Formula (7.5) except second probability 0.45. Using interpolation, the probability can be corrected to 0.26.

Using the methods mentioned above, all features developed in Section Chapter 4, Chapter 5 and Chapter 6 can be incorporated into the improved MBM. Next, the performance of using the MBM for scoring essays will be shown.

## 7.2 Performance of Predicting Model

Experimental corpus and the definition of performance are the same as that in Subsection 4.3.1. In the experiment, 343 training essays and 346 test essays are randomly chosen from the corpus. Table 7.1 shows the performance of the improved MBM employs all features, consisting of literary sememes, figures of speech “pi-yu”, “pai-bi” and literary “pi-yu”, inter-paragraph connections and intra-paragraph connections, to score essays in the corpus. An entry  $(i,j)$  in Table 7.1 represents the number of essays which are graded  $i$  by experts and graded  $j$  by the proposed method. For instance, the entry (2,3) represents there are 11 essays which are graded 2 by experts and graded 3 by the proposed method in test essays. In Table 7.1, the accurate rate and exact rate are 0.89 and 0.48. The results show the performance of the proposed method is close to that of current AES systems.

Table 7.1 Performance of the Improved MBM for Scoring Essays

human score	predictive score					
	1	2	3	4	5	6
1	20	2	1	0	0	0
2	14	37	11	2	0	0
3	9	28	49	18	1	0
4	6	11	37	45	5	0
5	1	2	9	19	15	0
6	0	0	0	2	2	0

Table 7.2 shows the performances of the proposed method using all features and three subsets of features. Feature set 1 only contains literary sememes; feature set 2 contains literary sememes and figures of speech “pi-yu”, “pai-bi” and literary “pi-yu”; feature set 3 contains literary sememes, inter-paragraph connections and intra-paragraph connections.

The results indicate the performance of the proposed method using all features is better than that using some of features. Although the accuracy rate of the proposed method using all features is less 4.3% than that using literary sememes only, the exact rate, average accurate rate and average exact rate of the proposed method using all features are respectively higher 3.8%, 4.2% and 11.6% than those using literary sememes only. By contrast Table 7.1 with Table 4.3, the results indicate the method using only literary sememes tends to increase the accuracy of scoring essays graded 3 or 4 as much as possible because these essays are the majority in training corpus. The usage of other features can improve the accuracy of scoring essays graded with other scores.

Table 7.2 Performances of the Proposed Method using Different Feature Sets

	Feature Set 1	Feature Set 2	Feature Set 3	All Features
Accuracy rate	0.916	0.812	0.864	0.873
Exact rate	0.442	0.350	0.451	0.480
Average accuracy rate	0.776	0.827	0.764	0.818
Average exact rate	0.330	0.385	0.405	0.446

## Chapter 8 Conclusions

The thesis proposed a novel methodology for scoring Chinese essays based on the extraction and analysis of conceptual frameworks in essays. Experimental results show that the performance of the methodology is quite close to that of current English AES systems. There are three characteristics in the methodology. First, it performs well based on the analysis of semantics in essays even if it does not employ surface features and syntax features. Second, the result of evaluation can be used for instructional feedback to the authors because it refers the conceptualization progress of authors to evaluate the quality of essays. Third, it overcomes the difficulties of applying current English AES systems to Chinese.

Many further studies can be developed based on the proposed methodology. First, various figures of speech should be explored and employed to increase the performance of scoring essays. This thesis has shown the usefulness of figures of speech for scoring essays. The extraction of more figures of speech may be useful for scoring essays and instructional feedback. Second, the e-learning system for writing instruction based on the proposed methodology will be very useful for education. Today, students cannot practice writing frequently because of lack of human raters. Using the interactive learning systems based on the proposed methodology, students can improve their writing skills fast and effectively.

## Reference

- [1] Y. Attali and J. Burstein, "Automated Scoring Using With e-raterV.2. The Journal of Technology," *Learning and Assessment*, vol. 4, no. 3, 2006.
- [2] M. H. Bai, C. J. Chen, and K. J. Chen, "Category Determination for Chinese Unknown Words Using Contextual Rules," *Proceedings of Research on Computational Linguistics Conference XI*, pp. 253-272, Taiwan, 1998.
- [3] S. T. Bai and A. W. Shi, "A Comparative Study of Figures of Speech between Chinese and English," *Journal of Xinzhou Teachers University*, vol. 18, no. 1, pp. 70-71, 2002.
- [4] J. Burstein, K. Kukich, S. Wolff, C. Lut, M. Chodorow, L. Braden-Harder and M. Dee Harri, "Automated Scoring Using A Hybrid Feature Identification Technique," *Proceedings of the 36<sup>th</sup> Annual Meeting of the Association of Computational Linguistics*, Montreal, Canada, 1998.
- [5] J. G. Cai, "A contrastive study of English and Chinese writing," Fudan University Press. Shanghai: China, 2001.
- [6] J. G. Cai, "Contrastive study of writing & rhetoric in English and Chinese," Fudan University Press, Shanghai: China, 2003.
- [7] T. H. Chang and C. H. Lee, "Topic Segmentation for Short Texts," *Proceedings of the 17<sup>th</sup> Pacific Asia Conference on Language, Information, and Computation*, Singapore, 2003.
- [8] T. H. Chang and C. H. Lee, "Automatic Chinese Unknown Word Extraction Using Small-Corpus-based Method," *Proceedings of IEEE International Conference on Natural language processing and knowledge engineering*, pp. 459-464, Beijing, China, 2003.
- [9] T. H. Chang and C. H. Lee, "Enhancing Automatic Chinese Essay Scoring System from Figures-of-Speech," *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pp. 28-35, Wuhan, China, 2006.
- [10] C. C. Chen, "A Study on Procedural Knowledge Learning for Mandarin Rhetoric Teaching," Master Thesis, Graduate Institute of Taiwan Languages and Language Education, National Hsinchu University of Education, Hsinchu, Taiwan, 2004.
- [11] C. J. Chen, M. H. Bai, and K. J. Chen, "Category Guessing for Chinese Unknown Words." *Proceedings of the 4<sup>th</sup> Natural Language Processing Pacific Rim Symposium*, pp. 35-40, Thailand, 1997.
- [12] H. J. Chen, "Analysis of Students Use Figure-of-speech "Pi-yu," *Newsletter for Languages and Literature Education*, vo. 15, pp. 48-55, 1997.
- [13] K. J. Chen and M. H. Bai. 1998. "Unknown Word Detection for Chinese by a Corpus-based Learning Method," *Computational Linguistics and Chinese Languages Processing*, vol. 3, no. 1, pp. 27-44, 1998.
- [14] K. J. Chen and W. Y. Ma, "Unknown Word Extraction for Chinese Documents," *Proceeding of the 19<sup>th</sup> International Conference on Computational Linguistics*, pp. 169-175, Taiwan, 2002.
- [15] CKIP, "Chinese Part-of-speech Analysis," Technical Report 93-05, Academia Sinica, Taipei, 1993.

- [16] S. Dikli, "An Overview of Automated Scoring of Essays," *The Journal of Technology, Learning, and Assessment*. vol. 5, no. 1, 2006.
- [17] S. M. Elliot, "IntelliMetric: From Here to Validity," *Proceedings of the annual meeting of the American Educational Research Association*, Seattle, WA, 2001.
- [18] M. Hearst, "The Debate on Automated Essay Grading," *IEEE Intelligent Systems*, vol. 15, no. 5, pp. 22-37. 2003.
- [19] S. He and J. Zhu, "Bootstrap method for Chinese new words extraction," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 581-584, Salt Lake City, Utah, USA, 2001.
- [20] H. J. Hsu and C. C. Wang, "The Effect of Reading-and-Writing Rhetoric Instructive Method on the Ability of Elementary Students' Writing Rhetoric Manner," *Proceedings of Conference on Children's Languages and Literature Education*, pp. 341-362, Taitung, Taiwan, 1999.
- [21] Z. L. Hu, "Application of Discourse Analysis in Language Teaching," *Foreign Language Education*. vol. 22, no. 1, pp. 3-10, 2001.
- [22] L. C. Huang, "Figure-of-speech "Pai-bi" (I)," *Chinese Language Monthly*. vol. 491, pp. 19-24, 1998.
- [23] T. T. Huang, "Ten Figures-of-Speech Instruction Using Syntax," *Elementary Education Century*, vol. 215, pp. 79-88, 2005.
- [24] C. Y. Jiao, "A syntactic comparison and transformation between English and Chinese," *Journal of Yancheng Teachers College*. vol. 22, no. 2, pp. 83-87, 2002.
- [25] R. B. Kaplan, "Cultural Thought Patterns in Intercultural Education," *Language Learning*, vol. 16, pp. 1-20, 1966.
- [26] H. W. Ko, "The Evaluation Criteria of Expository and Narrative Writings," *Journal of Chinese Language Teaching*, vol. 1, no. 2, pp. 15-32, 2004.
- [27] Y. Ko, J. Park and J. Seo, "Improving text categorization using the importance of sentences," *Information Processing and Management*, vol. 40, no. 1, pp. 65-79, 2004.
- [28] Y. S. Lai and C. H. Wu, "Meaningful term extraction and discriminative term selection in text categorization via unknown-word methodology," *ACM Trans. on Asian Language Information Processing*, vol. 1, no. 1, pp. 34-64, 2002.
- [29] W. Lam and K. S. Ho. "FIDS: an intelligent financial Web news articles digest system," *IEEE Trans. on Systems, Man and Cybernetics: Part A*, vol. 31, no. 6, pp. 753-762, 2001.
- [30] T. K. Landauer, D. Laham and P. W. Foltz, "The Intelligent Essay Assessor," *IEEE Intelligent System*, vol. 15, pp. 27-31, 2000.
- [31] L. S. Larkey and W.B. Croft, "A Text Categorization Approach to Automated Essay Grading. Automated Essay Scoring: A Cross-Disciplinary Perspective," *Lawrence Erlbaum Associates Inc.*, Mahwah New Jersey, pp. 55-69, 2003.
- [32] G. N. Lee, "Contrastive Studies of Figures of Speech in English and Chinese," *Fujian People's Publishing House*: Fuzhou, China, 1999.
- [33] X. L. Lee and K. Zeng, "Heterogeneity and Homogeneity of Sentence Structure in English and



- Chinese,” *Journal of Shenyang University*. vol. 12, no. 1, pp. 52-55, 2001.
- [34] W. S. Li and D. Agrawal, “Supporting web query expansion efficiently using multi-granularity indexing and query processing,” *Data and Knowledge Engineering*, vol. 35, no. 3, pp.239-257, 2000.
- [35] G. C. Ling, M. Asahara, and Y. Matsumoto, “Chinese Unknown Word Identification Using Character-based Tagging and Chunking,” *Proceedings of the 41<sup>st</sup> Annual Meeting of the Association for Computational Linguistics*, p.p. 197-200, Sapporo, Japan, 2003.
- [36] L. J. Liu, “A contrastive study of Discourse Structure in English and Chinese,” *Modern Foreign Languages*, vol. 86, no. 4, pp. 408-419, 1999.
- [37] K. Y. Liu and J. H. Zheng, “Research of Automatic Chinese Word Segmentation,” *Proceedings of the 1<sup>st</sup> International Conference on Machine Learning and Cybernetics*, vol. 2, pp. 805-809, Beijing, China, 2002.
- [38] W. Y. Ma and K. J. Chen, “Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff,” *2<sup>nd</sup> SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan, 2003.
- [39] R. Mandala, T. Tokunaga, and H. Tanaka, “Query expansion using heterogeneous thesauri,” *Information Processing & Management*, vol. 36, no. 3, pp. 361-378, 2000.
- [40] C. D. Manning and H. Schutze, “Foundations of statistical natural language processing,” Massachusetts: MIT Press, 1999.
- [41] A. Mikheev, “Automatic rule induction for unknown word guessing,” *Computational Linguistics*, Vol. 23, no. 3, pp. 405-423, 1997.
- [42] M. F. Moens and R. de Busser, “Generic Topic Segmentation of Document Texts,” *Proceedings of the 24th ACM SIGIR Annual International Conference on Research and Development in Information Retrieval*, pp. 418-419, 2001.
- [43] T Nakagawa, T. Kudoh, and Y. Matsumoto, “Unknown word guessing and part-of-speech tagging using support vector machines,” *Proceedings of the 6<sup>th</sup> Natural Language Processing Pacific-Rim Symposium*, pp. 325-331, Tokyo, Japan, 2001.
- [44] E. B. Page, “Computer Grading of Student Prose, Using Modern Concepts and Software,” *Journal of Experimental Education*, vol. 67, pp. 127-142, 1994.
- [45] J. M. Ponte and W. B Croft. “Text segmentation by topic,” *Proceedings of the 1<sup>st</sup> European Conference on Research and Advanced Technology for Digital Libraries*. pp. 120-129. Pisa, Italy, 1997.
- [46] L. M. Runder and T. Liang, “Automated Essay Scoring Using Bayes’ Theorem,” *The Journal of Technology, Learning, and Assessment*, vol. 1, no.2, 2002.
- [47] R. Scollon, S. W. Scollon, and A. Kirkpatrick, ”Contrastive Discourse in Chinese and English: A Critical Appraisal,” *Foreign Language Teaching and Research Press*: Shaihai, China, 2000.
- [48] M. D. Shermis, and J. C. Burstein (Eds), “Automated Essay Scoring: A Cross Disciplinary Perspective,” *Lawrence Erlbaum Associates Inc.*: Mahwah New Jersey, 2003.
- [49] M. S. Sun, D. Y. Shen, and C. N. Huang, “CSEg&Tag1.0: A practical word segmenter and POS tagger for Chinese texts,” *Proceedings of the 5<sup>th</sup> Conference on Applied Natural Language Processing*, pp.

- 119-126, Washington DC, USA, 1997.
- [50] S. M. Thede, "Predicting part-of-speech information about unknown words using statistical methods," *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 1505-1507, Montréal, Canada, 1998.
- [51] M. F. Tsai, "Figures-of-speech pi-yu, bi-ni and zhuan-hua", *World of Chinese Language and Literature*, vol. 16, no. 9, pp. 81-87, 2001.
- [52] T. Y. Tsai, "Rhetoric Theory and Writing Teaching," *Newsletter for Teaching the Humanities and Social Sciences*, vol. 9, no. 3, pp. 52-62, 1998.
- [53] H. H. Tseng, C. L. Liu, Z. M. Gao, and K. J. Chen, "A hybrid approach for automatic classification of Chinese unknown verbs," *Computational Linguistics and Chinese Language Processing*, vol. 7, no. 1, p.p. 1-28, 2002.
- [54] S. Valenti, F. Neri, and A. Cucchiarelli, "An overview of current research on automated essay grading," *Journal of Information Technology Education*, vol. 2, pp. 319-330, 2003.
- [55] J. P. Yamron, I. Carp, L. Gillick, S. Lowe and P. van Mulbregt, "A hidden Markov model approach to text segmentation and event tracking," *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 333-336, 1998.
- [56] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*. ACM Press, New York, 1999.
- [57] H. P. Zhang, Q. Liu, H. Zhang, and X. Cheng, "Automatic recognition of Chinese unknown words based on roles tagging," *Proceedings of the 1<sup>st</sup> SIGHAN Workshop on Chinese Language Processing*, pp. 71-77, Taiwan, 2002.
- [58] X. H. Zeng, "Enhancing English Writing Ability by Comparing the Difference of Organization in Paragraphs between English and Chinese," *Journal of Nanchang Vocation-technical Teachers College*, vol. 4, pp 75-77. 1997.