

# 國立交通大學

## 資訊科學與工程研究所

### 碩 士 論 文



建構一個中文對聯創作的知識評價架構

Building a Knowledge Evaluation Scheme for Chinese Couplet

Composition

研 究 生：陳紹宜

指 導 教 授：曾憲雄 教授

梁 婷 教授

中 華 民 國 九 十 九 年 六 月

# 國立交通大學

## 資訊科學與工程研究所

### 碩 士 論 文

建構一個中文對聯創作的知識評價架構

Building a Knowledge Evaluation Scheme for Chinese Couplet

Composition

研 究 生：陳紹宜

指 導 教 授：曾憲雄 教授

梁 婷 教授

中 華 民 國 九 十 九 年 六 月

# 建構一個中文對聯創作的知識評價架構

研究生: 陳紹宜

指導教授: 曾憲雄 博士

梁 婷 博士

國立交通大學

資訊科學與工程研究所

## 摘 要

對聯是一個重要的傳統中華文化之一。在這篇論文裡，我們提出了 **Couplet Analysis System (CAS)**，目的是為了能擷取並評價對聯的知識。而對聯的創作，有它固定的一些創作上的限制像是音韻、用字、以及語義上的要求，根據這些創作對聯的要求及特性，我們採用知識庫的方式定義出對聯中的知識屬性來擷取對聯的知識，然而這三個限制之中，語義的處理最為困難，所以這篇論文針對對聯語義處理的部分提出了重要的處理方法。為了針對對聯之中的各個詞彙來擷取知識，首先必須先將對聯先作準確的斷詞動作，因此我們提出了**HRWS**的對聯斷詞方法。接著在對聯語義的處理上，我們利用**E-HowNet**的架構，提出了**HBA**的方法來解決詞義歧異性問題，並且提出**EH-SSC**方法來決定上下聯間語義相似度。最後提出了知識評價機制，利用所擷取出來的對聯知識屬性來評價一幅對聯。為了評估**CAS**系統是否確實有效，我們利用了東吳大學的「全球徵聯」對聯比賽的2510篇參賽作品來做實驗，結果達到42%的對聯評價準確度。總結來說，**CAS**確實可以幫助使用者分析及評價他們的對聯。

# Building a Knowledge Evaluation Scheme for Chinese Couplet Composition

Student : Shau-Yi Chen

Advisor : Dr. Shian-Shyong Tseng  
Dr. Tyne Liang

Institute of Computer Science and Engineering  
National Chiao Tung University

## ABSTRACT

The Chinese couplet called *duì lián* is an important part of traditional Chinese culture. In this thesis, we propose Couplet Analysis System (CAS) that its goal is to extract and evaluate knowledge of a couplet. To analyze a couplet, the constraints about tone, word, and semantic meaning are concerned as important features in a couplet. We use knowledge-based approach to define the knowledge attributes that to extract knowledge of a couplet. Among these three features, the analysis of semantic meaning is the most difficult process. Therefore the thesis focuses on the semantic meaning analysis. Before processing the constraints, the word segmentation is addressed. Then Heuristic Rule-based Word Segmentation is proposed to solve this problem. In analysis of semantic, E-HowNet is employed to compute the semantic similarity. Following structure of E-HowNet, the thesis proposes *Heuristic-based approach* to solve the semantic tagging for word problem and *E-HowNet based semantic similarity approach* to compute semantic similarity value between sentences of a couplet. Finally, the thesis proposes Knowledge Evaluation mechanism by using the knowledge attributes to evaluate the couplet. The evaluation results of the system are compared with that of domain experts. The result shows that our approach yields 42% precision. To sum up, CAS can help couplet writers analyze and evaluate couplets.

# Acknowledgement

首先感謝最令我佩服的指導教授: 曾憲雄博士。在這兩年的碩士生涯中, 曾教授教導我最正確的邏輯觀念, 使得每一項論文中研究的細節都可以環環相扣, 除此之外, 老師不厭其煩的使用各種不同的思考方向來切入我所研究的主題, 讓我切確的感受到老師的用心以及指導學生做研究的熱誠。在此, 真的非常感謝曾教授的用心指導。同時, 我也要感謝我的共同指導教授: 梁婷博士。梁教授是一個非常與眾不同的教授, 對待學生就像在對待自己的兒女一般十分親切, 除了指導我作研究之外, 還時常分享一些她的生活經驗, 在面對未來出社會的做人處世上, 讓我獲益良多。接著要感謝我的口試委員, 洪宗貝教授、曾秋蓉教授, 他們給予了我相當多的寶貴意見, 讓本論文更有意義與價值。

此外, 我要感謝帶我的哲青學長, 他花了自己非常多的時間來協助我做論文, 在他身上我看見的負責與堅持的態度, 是一個非常好的榜樣! 還有實驗室裡的同學們, 嘉祥、國彰、佳榕、杰峰、金龍, 在學習的路上大家互相扶持, 很高興有機會跟你們在同一間實驗室裡彼此學習和分享心情。再來, 感謝我的好朋友, 宜詠、銘鴻, 在我最低潮的時候鼓勵我、關心我, 讓我有動力繼續完成我的學業, 能認識你們是我最幸運的事。還有 IR 實驗室的夥伴們, 有你們在我真的很開心, 其中最感謝冠熙和俊樺, 在我論文以及系統上給予我很多的幫助。還有其他無法一一詳述的朋友們, 感謝你們在我學習的路上——一路相伴。

最後要感謝我的家人和 mini, 在我求學的旅途中你們不斷為我加油打氣, 是我能完成這篇論文最主要的原動力。在未來的日子裡, 面對任何挑戰我都會勇往直前, 不負你們對我的期待。

紹宜

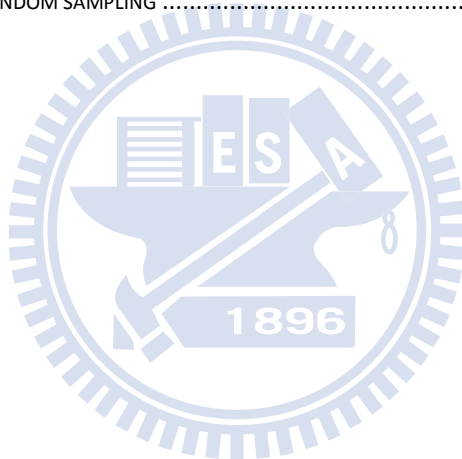
新竹/2010/7

# Content

|  |     |
|--|-----|
| Abstract (Chinese).....                                  | i   |
| Abstract (English) .....                                 | ii  |
| Acknowledgement .....                                    | iii |
| Content.....   | iv  |
| LIST OF TABLES .....                                     | v   |
| LIST OF FIGURES .....                                    | vi  |
| LIST OF ALGORITHMS.....                                  | vii |
| Chapter 1 Introduction .....                             | 1   |
| Chapter 2 Related Work.....                              | 4   |
| 2.1 Couplet Rules .....                                  | 4   |
| 2.2 Introduction to Word Segmentation.....               | 5   |
| 2.3 Word Sense Disambiguation.....                       | 5   |
| 2.4 Introduction to E-HowNet.....                        | 6   |
| 2.5 Scoring System .....                                 | 8   |
| 2.6 Related Research about Couplet.....                  | 9   |
| Chapter 3 Problem Formulation .....                      | 10  |
| 3.1 Couplet Knowledge .....                              | 10  |
| 3.2 The Assessment of Couplet Knowledge.....             | 13  |
| Chapter 4 Heuristic Rule-Based Word Segmentation.....    | 15  |
| Chapter 5 Semantic Similarity Computation .....          | 19  |
| 5.1 Semantic Tagging for Word .....                      | 19  |
| 5.2 E-HowNet Based Semantic Similarity Computation ..... | 21  |
| 5.3 Semantic Similarity for Couplet .....                | 30  |
| Chapter 6 Couplet Evaluation .....                       | 31  |
| Chapter 7 Experiment and Analysis.....                   | 34  |
| 7.1 System Implementation .....                          | 34  |
| 7.2 Experiment Result and Discussion .....               | 35  |
| Chapter 8 Conclusion.....                                | 42  |
| Reference .....  | 43  |

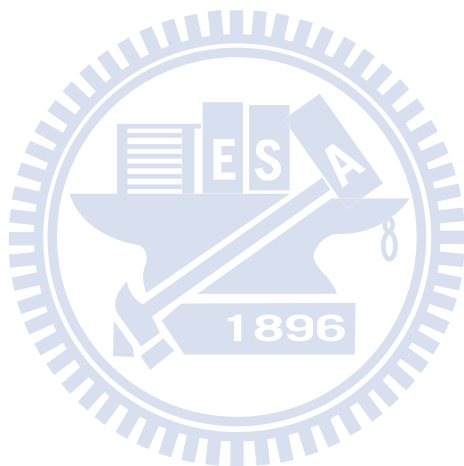
# LIST OF TABLES

|   |    |
|---|----|
| TABLE 5.1: THE MEANING AND POS OF “重/愛” .....   | 21 |
| TABLE 5.2: CORE LEXICON OF “大甲/三星” .....  | 28 |
| TABLE 5.3: ATTRIBUTES ANNOTATION OF FRAMEWORK IN “大甲/三星” .....                        | 28 |
| TABLE 5.4: ATTRIBUTE VALUES ANNOTATION OF FRAMEWORK IN “大甲/三星” .....                  | 29 |
| TABLE 6.1: TONE SCORE TABLE .....   | 32 |
| TABLE 6.2: WORD SCORE TABLE .....   | 32 |
| TABLE 6.3: MATCH CONDITION OF 大甲狀元粽 三星上將梨.....  | 33 |
| TABLE 6.4: EVALUATION SCORE RANGE WITH COUPLET LEVEL .....                            | 33 |
| TABLE 7.1: EXPERT LEVEL COMPARE WITH THE CAS SCORE .....                              | 37 |
| TABLE 7.2: PRECISION OF COUPLET ANALYSIS SYSTEM .....                                 | 37 |
| TABLE 7.3: EXPERIMENT RESULT WHEN “三星” IS NAME OF A TOWNSHIP .....                    | 37 |
| TABLE 7.4: PRECISION OF COUPLET ANALYSIS SYSTEM WHEN “三星” IS NAME OF A TOWNSHIP ..... | 38 |
| TABLE 7.5: PRECISION BY USING RANDOM SAMPLING .....                                   | 39 |



# LIST OF FIGURES

|  |    |
|--|----|
| FIGURE 2.1: THE FRAMEWORK STRUCTURE OF “三星” .....  | 8  |
| FIGURE 3.1: THE STRUCTURE OF CONTEXT INFORMATION .....   | 11 |
| FIGURE 3.2: COUPLET ANALYSIS SYSTEM STRUCTURE.....   | 14 |
| FIGURE 5.1: SEMEME STRUCTURE IN E-HOWNET.....  | 23 |
| FIGURE 5.2: THE FRAMEWORK OF P, Q TREE.....  | 26 |
| FIGURE 5.3: THE FRAMEWORK OF A1, A2 TREE .....   | 27 |
| FIGURE 5.4: TREE OF FRAMEWORK IN CORRESPONDING PAIR “大甲”AND “三星” .....   | 29 |
| FIGURE 5.5: TREE OF FRAMEWORK VALUE IN CORRESPONDING PAIR “大甲”AND “三星” .....   | 30 |
| FIGURE 7.1: INTERFACE OF COUPLET ANALYSIS SYSTEM WHICH COULD BE LINKED FROM<br><a href="http://skywalf.no-ip.info:8080/couplet/analysis_of_5.jsp">HTTP://SKYWALF.NO-IP.INFO:8080/COUPLET/ANALYSIS_OF_5.JSP</a> ..... | 34 |
| FIGURE 7.2: RESULT OF COUPLET ANALYSIS SYSTEM.....   | 35 |
| FIGURE 7.3: TREE STRUCTURES OF DIFFERENT SENSES ON “三星” .....  | 38 |





# LIST OF ALGORITHMS

|  |    |
|--|----|
| ALGORITHM1: HEURISTIC RULE-BASED APPROACH ALGORITHM.....                             | 17 |
| ALGORITHM2: FRAMEWORK STRUCTURE SIMILARITY IDENTIFICATION ALGORITHM (FSSI ALG) ..... | 25 |



# Chapter 1 Introduction

The couplet is an important part of traditional Chinese culture. A Chinese proverb states that, “*The sight strikes a chord in one's heart.*” People usually compose a couplet to express their emotion in that moment. However there are some constraints of couplet composition such as tone, word, and semantic meaning. The composers must follow these couplet constraints to compose their couplets. It is not easy to compose a couplet, and therefore the system that can analyze these couplets with constraints automatically is expected. In this thesis, **Couplet Analysis System (CAS)** is proposed to analyze a couplet and evaluate it.

To analyze a couplet, the constraints about tone, word, and semantic meaning are concerned as important features in a couplet. Therefore, a set of attributes are defined in the **Context Information**. Through the **Context Information**, the thesis can analyze couplets. However, the analysis of tone and word is not difficult, but the analysis of semantic meaning is difficult. Previous researches only deal with the analysis of tone and word. The thesis focuses on the semantic meaning analysis.

Before processing the constraints, the word segmentation is addressed. Due to the harmonious characteristic, a couplet can be segmented based on the sentence patterns. Furthermore, several heuristics can be used to determine the sentence pattern. Therefore, **Heuristic Rule-based Approach** is proposed to segment word for a couplet.

The semantic constraints in couplet are that the semantic meaning between two sentences must be related, that is, the semantic similarity is high. The semantic similarity computation has two major approaches: corpus-based [1] and

distance-based [2] . The corpus-based approach uses a large number of couplets as training data to compute the semantic similarity which is the probability of two words occur in a couplet. The distance-based approach uses thesaurus that is tree or net structure to compute the distance (number of edges) between two nodes as semantic similarity. However, a couplet consists of very few words. Besides, good couplet corpus is unavailable now. Since the Extended-HowNet (E-HowNet) are well-defined and the latest tree structure thesaurus, the thesaurus are employed to compute the semantic similarity. To solve this problem, two sub-problems occurred: Semantic Tagging for Words, and Semantic Similarity Computation. Therefore, the thesis proposes **Heuristic-Based Approach** to solve the first sub-problem and **E-HowNet based Semantic Similarity Approach** to solve the second sub-problem.

For evaluating the compositions, previous studies developed scoring systems to grade literature compositions such as e-rater [3] . They use several features as the evaluation criteria. Therefore, the thesis proposes **Couplet Evaluation** by using the **Context Information** to evaluate the couplet.

To evaluate the applicability of the experiment, we use “7<sup>th</sup> global couplet composition” [16] as our testing data that contains 2510 couplets. The evaluation results of the system are compared with that of domain experts. The result shows that our approach yields 42% precision.

In Chapter 2, we introduce related works: *Couplet Rules*, *Introduction to Word Segmentation*, *Semantic tagging for words*, *Scoring System*, *E-HowNet*, and *Related Research about Couplet*. In Chapter 3, we define the Context Information and the sub-problems. In Chapter 4, the **Heuristic Rule based words segmentation approach** is proposed, and its idea uses *sentence pattern*, *Known word*, *Longest word*, *Proper noun*, and *Allusion* to address the words segmentation in a sentence. In Chapter 5, this thesis proposes **Heuristic based approach** to tag semantics for words

and **E-HowNet based semantic similarity computation** to calculate semantic similarity value between two sentences in a couplet. In Chapter 6, the **Couplet Evaluation** is proposed to assess the couplet. In Chapter 7, we introduce the **Couplet Analysis System** and the experiment results. Finally, the contribution of this thesis is discussed in Chapter 8.



## Chapter 2 Related Work

The goal of this thesis is to extract and evaluate knowledge of a couplet. First, the background information of a couplet is introduced. For extracting knowledge, word segmentation problem and semantic tagging for words need to be solved. In order to compute the semantic similarity of a couplet, E-HowNet is used to assist in calculating. To evaluate the knowledge of a couplet, related researches about scoring system is addressed. At last, the thesis introduces related researches about couplet.

### 2.1 Couplet Rules

There are some constraints about the couplet rules such as tone, word, and semantic meaning. We survey related books or researches [17] [18], and five basic constraints are listed as follows.

1. In Chinese, each character is pronounced either “Ping” (平) or “Ze” (仄). The tone of the last character in first sentence must be “Ze”(仄); The tone of the last character in second sentence must be “Ping”(平).
2. The word number of first sentence and second sentence must be the same.
3. If the same word appears twice in first sentence, then the word appears twice in the same position of second sentence.
4. Same character cannot be used in the difference sentence.
5. The contents of the two sentences should be related, but not duplicated.

## 2.2 Introduction to Word Segmentation

Previous studies indicate that there are two primary methods to solve word segmentation problem: Statistic based approach and Rule based approach.

Statistic based approach uses statistic information, probability information, and mathematic model to determine the results of word segmentation. For example, W. Andi [4] uses modified maximal matching to segment words and then using the segmentation results to build up a parsing tree. However, this approach needs a large amount of couplets as training data, and a couplet consists of few words. It's hard to construct a word segmentation model using the couplets.

Rule based approach uses well-designed rules to segment words. For example, K.J. Chen [20] found out the possible words by using thesaurus and then filter impossible words by using word structure and word combination rules. Since couplet composition must follow constraints, it is suitable to use rule based approach to segment words. Therefore, the thesis develops a rule based word segmentation approach.

## 2.3 Word Sense Disambiguation

The Word Sense Disambiguation (WSD) is the task of determining which meaning of a polysemous word is intended in a given context. Different kinds of training data, features, and learning algorithms have been proposed in the computational literature of WSD.

Supervised methods [5] [11] are corpus-based supervised learning methods for WSD. It uses a sense-tagged training data to build a word sense classifier. R. Bruce and J. Wiebe [5] use multiple contextual features for word-sense disambiguation, without requiring untested assumptions regarding the form of the model. In this

approach, the joint distribution of all variables is described by only the most systematic variable interactions, thereby limiting the number of parameters to be estimated, supporting computational efficiency, and providing an understanding of the data. Supervised methods usually have good accuracy but building the sense-tagged corpora must spend a lot of time.

Unsupervised methods [6] [12] are corpus-based unsupervised learning methods for WSD. It doesn't use sense-tagged training data, but using large of texts to find feature of words. C. Leacock, M. Chodrow, and GA. Miller [6] use a statistical classifier that combines topical context with local cues to identify a word sense. The classifier is used to disambiguate a noun, a verb, and an adjective. Besides, WordNet's lexical relations are used to automatically locate training examples in a general text corpus.

Knowledge-based methods disambiguate word sense by matching context with information from a prescribed knowledge source such as well-defined thesaurus. For example, E. Agirre, O.L.D Lacalle, and A. Soroa [7] use WordNet information to solved WSD in Sports and Finance Domain.

Since it's hard to collect a large amount of quality couplets as training data, the thesis does not use supervised and unsupervised methods. Extended-HowNet (E-HowNet) is a lexical knowledge base, which consists of definitions for lexical senses and an ontology. Therefore, knowledge-based method using E-HowNet is proposed to solve WSD problem.

## **2.4 Introduction to E-HowNet**

The Sinica CKIP group and Professor Dong build a HowNet for traditional Chinese in a cooperative project. They use the HowNet-based meaning representation

mechanism to define the word meanings of over 90,000 lexical entries in the CKIP Chinese Lexical Knowledge Base called Extended-HowNet [21] [8] . But it creates new representation mechanism named Concept representation. Therefore, E-HowNet has two representation, Sememe representation and Concept representation.

Sememe representation inherits HowNet representation. A sememe denotes an unit of transmitted or intended meaning; it is atomic or indivisible. This representation uses fixed and limited sememes to represent a concept of a word. Concept representation uses one or more simple concepts to represent a complex concept, which could shorten the length in Sememe representation and could be understood easily. The simple concepts come from domain experts or the sememes.

Each word has these two representations, and each representation comprises **Core lexicon**, **Attribute**, and **Attribute value**. Core lexicon is the most important semantic meaning in a word and attributes are used to modify core lexicon. In general, the likely concept has the same attribute set.

#### EXAMPLE 2.1:

三星

<Concept represent> : { 公司:quantifier={ definite|定指},name={"三星"},location={ 韓國}}

<Sememe represent>: { InstitutePlace|場所:quantifier={ definite|定指},location={ country|國家:location={ continent|大陸:quantifier={ definite|定指},name={"亞洲"}},quantifier={ definite|定指},name={"韓國"}},domain={ economy|經濟},name={"三星"}}



*Core lexicon* is first semantic element in the representation. Take “三星” as an example, Concept representation is “公司” and Sememe representation is “場所”. Attributes in Concept representation are “qualification, name, location” and their values are “定指”, “三星”, “韓國”. Attributes in Sememe representation are “quantifier, location, domain, name, location, quantifier, name, quantifier, name” and their values are “定指, 國家, 經濟, 三星, 大陸, 定指, 亞洲, 定指, 韓國”.

Furthermore, the attribute set of Sememe representation can be represented as a hierarchical structure as shown in Figure 2.1, which is called Framework structure (Tree structure). The root of this tree is *core lexicon*. The higher level in the structure denotes the more correlation with *core lexicon*.

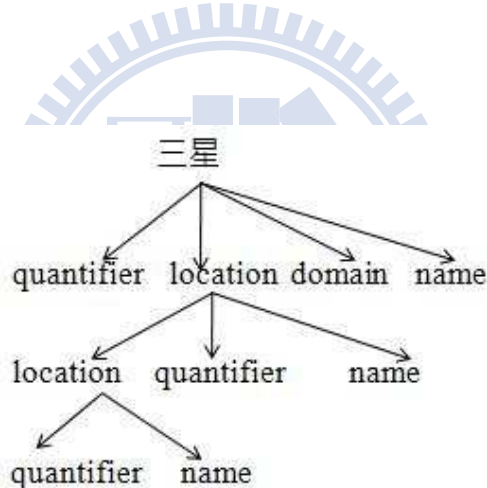


Figure 2.1: The Framework structure of “三星”

## 2.5 Scoring System

There are several organizations using automatic scoring systems to score or grade the essay or literature because scoring by computer is fair. For instance, the TOEFL uses e-rater [3] [10] that measures multitudinous features of writing in its training essays. Then it uses a stepwise linear regression procedure to choose the features that are most predictive of essay score. Otherwise, the Basic Competence Test for Junior High School Students in Taiwan will use Automatic Composition

Evaluation System (ACES) [13] [14] [15] to evaluate the composition of student. In this thesis, the couplet scoring mechanism is proposed to evaluate a couplet. It can be used on couplet competition in the future.

## **2.6 Related Research about Couplet**

Previous studies show that their couplet system can generate the second sentence based on the first sentence such as Microsoft couplet system [9] . It uses a statistical MT approach to generate Chinese couplets. First, the composer input the first sentence in their system, and then it uses a phrase-based SMT decoder to generate an N-best list of proposed second sentences as output. Next, a set of filters is used to delete some candidates violating couplet rules by their definition. Finally, it re-arranges the candidates. Otherwise, couplet system of China Tsinghua University [22] uses forward maximum matching and first-order Markov model (FMM) to generate couplets. First, they segment the first sentence of a couplet on a scroll using FMM. Next, they find matched candidates from the corpus, and then the dynamic programming technique is used to give a second sentence of a couplet.

These two systems are used to generate the best second sentence of couplet when author inputs the first sentence. However, it doesn't analyze the inputted couplets and grade them. In this thesis, our couplet analysis system generates the analysis list and evaluation table which can assist couplet composers to understand advantages and disadvantages in their couplets.

## Chapter 3 Problem Formulation

In order to analyze a couplet, this thesis proposes knowledge-based approach to extract couplet knowledge which uses a set of attributes to extract knowledge. When the knowledge is extracted, there are three problems occurred. Therefore, the attributes and problems are described in more detail.

### 3.1 Couplet Knowledge

The knowledge-based approach uses a set of attributes to extract knowledge which is context information. Before defining the context information, we introduce what and why these attributions are selected.

According to couplet books and related researches [17] [18] , couplets have six types [18] – spring festival, wedding, funeral, birthday, funny, and others. When couplet composers feel frustration, happy, angry, and etc. in a specific location, they compose a type of couplet. Therefore, mood, location, and type are important attributes in a couplet. Through these attributes, people can understand the background information of the couplet.

Since couplet composition needs to follow some restrictions, e.g., tone, word, and semantic meaning, these information are important. Therefore, a set of retrieved words, POS, pronunciation, and semantic meaning are taken as essential attributes for a couplet. Based on the analysis, the structure of context information is shown in Figure 2.1.

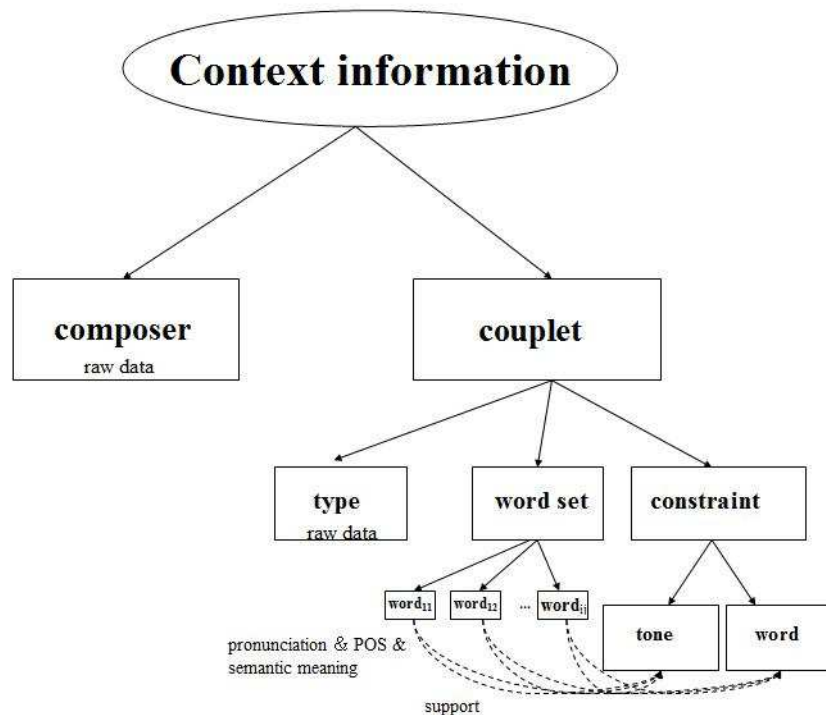


Figure 3.1: The structure of context information

### Definition: Context Information

- composer={name, mood, location }
- couplet={text, type, word set, tone constraint, word constraint }
  - type={spring festival, wedding, funeral, birthday, funny, others }
  - word set ={word<sub>11</sub>, word<sub>12</sub>, ... ,word<sub>ij</sub>}, i is sentence number and j is word number
    - word<sub>ij</sub>={POS, pronunciation, semantic meaning }
  - tone constraint ={ZP, opposite}
    - ZP ={match, non-match }
    - opposite ={match, non-match }
  - word constraint={repetition of words, use of words, POS consistency}
    - repetition of words = {match, non-match }
    - use of words= {different, same }

- POS consistency = {match, non-match }

**Definition: ZP**

**ZP** is a tone constraint. The tone of the last character in first sentence must be “Ze”; and the tone of the last character in second sentence must be “Ping”.

**Definition: Opposite**

**Opposite** is a tone constraint. The character of the same position in the first and the second sentences are pronounced oppositely.

**Definition: Repetition of words**

**Repetition of words** is a word constraint. If the same word appears twice in first sentence, then the word appears twice in the same position of second sentence.

**Definition: Use of words**

**Use of words** is a word constraint. Same character cannot be used in the difference sentence.

**Definition: POS consistency**

**Pos consistency** is a word constraint. The POS of word in the same position between two sentences is the same.

**EXAMPLE 3.2:**

Take a couplet “七夕情人果，三星上將梨。” as an example, which comes from 7<sup>th</sup> global couplet composition (全球徵聯) [16] . The context information is as follows.

- composer={陳莉莉, null, null}
- couplet={text, type, word set, tone constraint, word constraint }
  - text=七夕情人果，三星上將梨。
  - type= funny
  - word set ={七夕, 情人, 果, 三星, 上將, 梨}
    - Word<sub>11</sub> = { N, Ze Ze, 節}, Word<sub>12</sub> = { N, Ping Ping, 人}, Word<sub>13</sub> = { N, Ze, 水果}
    - Word<sub>21</sub> = { N, Ping Ping, 公司}, Word<sub>22</sub> = { N, Ze Ze, , 軍官}, Word<sub>23</sub> = { N, Ping, 水果}
  - tone constraint ={match, match}
  - word constraint={non-match, different, match }

## 3.2 The Assessment of Couplet Knowledge

Based on the defined attributes, the couplet knowledge can be extracted from a couplet. However, there are three sub-problems occurred in assessment of couplet knowledge process. The three sub-problems are listed below:

### Subproblem<sub>1</sub>: Word segmentation sub-problem

When given a couplet and outer thesauruses, the problem is how to segment word for a couplet correctly.

### Subproblem<sub>2</sub>: Semantic Similarity sub-problem

When given the results of word segmentation and outer thesaurus, the problem is how to compute the semantic similarity for a couplet. Furthermore, the problem contains the problem of sense tagging for words.

### Subproblem<sub>3</sub>: Couplet Evaluation sub-problem

When given the attribute values, the problem is how to grade couplet knowledge.

According to these three subproblems, the system architecture is designed in Figure 3.2. The *Couplet Analysis System* (CAS) is divided into five parts: word segmentation, semantic similarity recognition, tone recognition, word recognition, and couplet evaluation. The input of CAS is a couplet and the outputs are a knowledge table and an evaluation table. Finally, CKD stores the couplet knowledge.

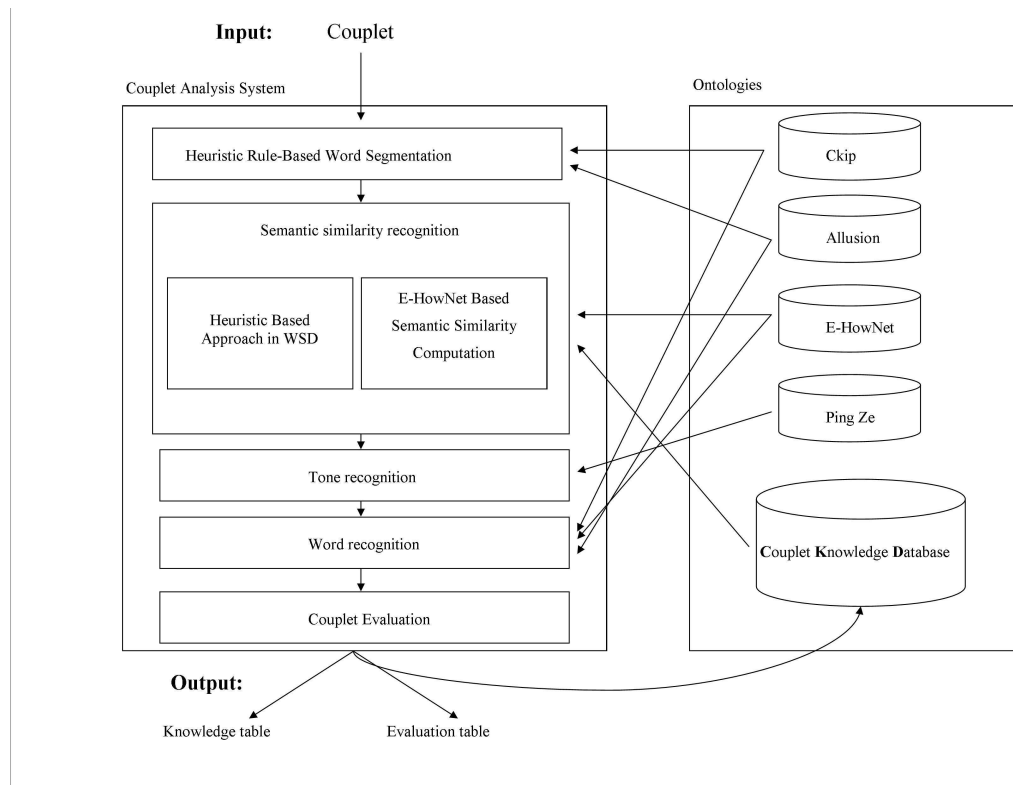


Figure 3.2: Couplet Analysis System structure

# Chapter 4 Heuristic Rule-Based Word Segmentation

Before semantic similarity analysis, words in a couplet must be identified, which is word segmentation problem. Rule-based approach is suitable for solving this problem, because the couplet composers follow the harmonious characteristic. The characteristic results in the result that each sentence in couplet can be segmented to several sentence patterns.

In couplet composition, the sentence patterns are fixed. The problem is how to segment word by selecting a correct sentence pattern. There are five significant characteristics that can influence the selection of sentence pattern: sentence patterns priority, known words, longest word, proper noun, and allusion. The heuristic rules are defined below. The thesis takes five characters in a sentence as an illustration.

## **Heuristic 1: Sentence patterns.**

Dr. Shiu [19] pointed out there are nine sentence patterns that are used in a sentence of couplet: The priority of the sentence patterns is decreasing “ $2/3$ ,  $2/2/1$ ,  $2/1/2$ ,  $1/2/2$ ,  $3/2$ ,  $4/1$ ,  $1/4$ ,  $1/3/1$ ,  $1/1/3$ ”. For example, the sentence pattern of this sentence “三星上將梨” is  $\{2/2/1\}$ , that is, the segmentation result is “三星/上將/梨”. Besides, according to the usage frequency,  $\{2/3, 2/2/1, 2/1/2, 1/2/2, 3/2\}$  is popular patterns and  $\{4/1, 1/4, 1/3/1, 1/1/3\}$  is unpopular patterns.

## **Heuristic 2: Known word.**

Known word denotes that the words can be found in thesaurus. If all tokens in a



sentence pattern are known words, then the candidate sentence pattern has the highest priority to be the final sentence pattern.

### **Heuristic 3: Longest word.**

Longest word denotes that the word has the longest characters over whole possible sentence pattern. If the longest word is a known word, then the candidate sentence pattern has the highest priority to be the final sentence pattern.

### **Heuristic 4: Proper noun.**

Proper nouns are nouns representing unique entities (such as *London* or *John*), as distinguished from common nouns which describe a class of entities (such as *city* or *person*). Proper nouns usually express specific semantic meanings, and therefore they have higher priority than common nouns. If one of the sentence patterns contains a proper noun, the sentence pattern has higher priority than the sentence patterns with common nouns.

### **Heuristic 5: Allusion.**

Poets usually use allusion to express profound semantic meaning; therefore allusion also has higher priority than common nouns. If one of the sentence patterns contains allusions, the sentence pattern has higher priority than the sentence patterns with common nouns.

To segment words for a couplet, the thesis proposes Heuristic Rule-Based Word Segmentation (HRBWS) based on five heuristics. The number of known words is the most important feature to determine the correct sentence pattern. Since couplet composers used to compose couplets by referring to noted couplets, the usage priority

of sentence pattern is an important characteristic to segment word. The sentence pattern can be a candidate pattern when all tokens in a sentence are known words. If there are unknown words in all sentence patterns, longest words can be used to determine the sentence pattern especially when the sentence pattern is unpopular. However, the previous steps may not work. The number of known words, allusions and proper nouns can be the decisive factor to determine the candidate sentence pattern. In the last, the sentence patterns of the two sentences in couplet must be the same, and therefore the highest priority of the candidate sentence patterns is selected as the sentence pattern. The algorithm is shown as follows.

Algorithm1: Heuristic Rule-Based Approach Algorithm

**Input:** A couplet

**Output:** The result of *word segmentation*

**Method:**

Step 1: User inputs a couplet.

Step 2: For each sentence:

Step2.1: Segment word by using sentence patterns.

Step2.2: If all tokens in a sentence pattern are known words, the sentence pattern is candidate sentence pattern.

Step2.3: If there is a longest word in an unpopular sentence pattern, the sentence pattern is candidate sentence pattern.

Step2.4: Compare the number of *known word*, *proper noun*, and *allusion* and compare the *sentence pattern priority* to decide candidate popular sentence pattern.

Step3: Select higher priority between sentence patterns of first sentence and second sentence.

Step4: Output the final sentence pattern.

**EXAMPLE 4.1:**

大甲狀元粽，三星上將梨。

First, the couplet is segmented by HRBWS. Based on the sentence pattern and

known word heuristics, the candidate sentence pattern in first sentence and second sentence is “2/2/1” and “2/2/1”. Since the candidate sentence patterns of the two sentences are the same, the final sentence pattern is “2/2/1”. Therefore, the segmentation results is “大甲, 狀元, 粽, 三星, 上將, 梨”.

**EXAMPLE 4.2:**

山窮水盡人，柳暗花明村。

First, the couplet is segmented by HRBWS. Since not all tokens in each sentence pattern are known words, the longest word heuristic is used to determine sentence pattern. In this couplet, the candidate sentence pattern in first sentence and second sentence are “4/1”. Therefore, the final sentence pattern is “4/1” and the segmentation results is “山窮水盡, 人, 柳暗花明, 村”.

**EXAMPLE 4.3:**

故人江海別，幾度隔山川。

First, the couplet is segmented by HRBWS. Based on the sentence pattern and known word heuristics, the candidate sentence pattern in first sentence is “2/2/1”. In second sentence, not all tokens in each sentence pattern are known words and no longest word are found in unpopular sentence pattern. Based on comparing number of known words, number of allusion, number of proper noun, and sentence pattern priority, the candidate sentence pattern in second sentence is “2/1/2”. Since the sentence pattern of first sentence and second sentence must be the same in a couplet, the final sentence pattern is “2/2/1”. Finally, the segmentation results is “故人, 江海, 別, 幾度, 隔山, 川”.

## Chapter 5 Semantic Similarity Computation

The semantic similarity computation method is introduced in this Chapter. Since it's hard to collect a large amount of quality couplets as training data, the thesis does not use corpus-based method. E-HowNet [21] is a lexical knowledge base, which consists of definitions for lexical senses and an ontology. Besides, E-HowNet has tree structure that supports distance-based method to compute similarity between two words. Therefore, E-HowNet is selected to tag semantics for word and to compute semantic similarity in this thesis.

### 5.1 Semantic Tagging for Word

The semantic tagging for word means to identify sense for a word. It can be divided into two parts. When a word has only one sense, we extract a sense from E-HowNet immediately. When a word is a polysemous word, it is called Word sense disambiguation problem. Word sense disambiguation (WSD) is the process of assigning a sense to a polysemous word based on the context in which it occurs. Since there are polysemous words in E-HowNet, the problem of WSD need to be solved. In a couplet, the word in first sentence and the word in second sentence are in the same position, which is called corresponding pair. Previous researches [2] indicate that the WSD problem in corresponding pair can be solved by using maximum similarity value between two words because semantic meaning of the two words must be similar. However, POS is an important feature for a couplet. The POS in corresponding pair should be the same. Besides, the historic record can be used to solve WSD. Therefore, the thesis proposes *Heuristic Based approach* to solve WSD. The heuristics are listed

as follows.

### **Heuristic 1: Historical record**

If corresponding pairs in a couplet were stored in *historical record*, the WSD could be solved by using the confirmed sense. For example, A is a polysemous word and {A, B} is a corresponding pair. In this case, the sense of A is  $\alpha$ . If the corresponding pair {A, B} appears again, the sense of A is  $\alpha$ .

### **Heuristic 2: POS agreement**

The POS of corresponding pair in a couplet must be the same. For this reason, WSD could be solved by reference POS of the other one of the corresponding pair. For example, A is a polysemous word which represents distinct POS and {A, B} is a corresponding pair. In this case, the sense of B is N and therefore the sense of A is N.

### **Heuristic 3: Max similarity**

If a polysemous word cannot be solved by heuristic 1 and heuristic 2, WSD could be solved by E-HowNet Semantic Similarity Computation (Eq 1) and retrieve the maximum. For example, A is a polysemous word and {A, B} is a corresponding pair. The sense of A is  $\alpha$  when the sense has the maximum semantic similarity with B.

#### **EXAMPLE 5.1:**

Take a couplet “男人重義氣，女人愛交心。” for example. After word segmentation, the segmentation result is “男人/重/義氣 女人/愛/交心”. Three corresponding pairs are “{男人, 女人}”, “{重, 愛}”, “{義氣, 交心}”. The “{重, 愛}” are polysemous words.

Table 5.1: The meaning and POS of “重/愛”

| word | entry   | 1  | 2  | 3 | 4 | 5  | 6 |
|------|---------|----|----|---|---|----|---|
| 重    | meaning | 重  | 注意 | 再 | 疊 | 嚴重 | 強 |
|      | POS     | V  | V  | D | N | V  | V |
| 愛    | meaning | 愛戀 | 情感 |   |   |    |   |
|      | POS     | V  | N  |   |   |    |   |

The corresponding pair “重/愛” in a couplet were not stored in *historical record*. Therefore, POS is checked. The POS in these two words are not unique, so it can't use the characteristic of POS to assign sense to them. Finally, based on the semantic similarity computation mentioned in Chapter 5.2, the meaning of “重” is “注意” and the meaning of “愛” is “愛戀”

## 5.2 E-HowNet Based Semantic Similarity Computation

The goal of **E-HowNet Based Semantic Similarity Computation (EH-SSC)** is to compute the semantic similarity of corresponding pair between two sentences in a couplet. E-HowNet is a common sense knowledge base annotating inter-conceptual relations and inter-attribute relations of concept as signified in lexicons of the Chinese and their English equivalents. Based on the structure of E-HowNet, EH-SSC contains three approaches: *Core Lexicon Computation*, *Tree structure computation*, and *Attribute Value Set Similarity*.

## Core lexicon computation

$$\text{Sim}(W_1, W_2) = \begin{cases} 5, & \text{if first concept and first sememe are same} \\ 4, & \text{if only first concept are same} \\ 3, & \text{if only first sememe are same} \\ S(W_1, W_2), & \text{otherwise} \end{cases} \quad \text{Eq 1}$$

$$S(W_1, W_2) = \begin{cases} \frac{1}{2}(3 \times C(W_1, W_2) + T(\text{tree1}, \text{tree2})), & \text{if } T(\text{tree1}, \text{tree2}) < 3 \\ \frac{1}{3}(3 \times C(W_1, W_2) + T(\text{tree1}, \text{tree2}) + F(A_1, A_2)), & \text{Otherwise} \end{cases} \quad \text{Eq 2}$$

Where

$C(W_1, W_2)$ : similarity between core lexicons of  $W_1$  and  $W_2$

$T(\text{tree1}, \text{tree2})$ : similarity between tree structures of  $W_1$  and  $W_2$

$F(A_1, A_2)$ : similarity between attribute sets of  $W_1$  and  $W_2$

The core lexicon is the first semantic elements in the two representations of E-HowNet structure. In Eq 1, if core lexicons (first lexicon in representation) of concept representation and sememe representation in corresponding pair are the same, the semantic similarity value of corresponding pair is 5; if only core lexicons of concept representation in corresponding pair are the same, the value is 4; if only core lexicons of sememe representation in corresponding pair are the same, the value is 3. If none of the cases existed, Eq 2 is used to compute the semantic similarity value of corresponding pair, which can be divided into three sub-functions: *C function*, *T function*, and *F function*. *C function* is used to compute semantic similarity value between core lexicons of two words in a corresponding pair. Moreover, to alleviate the deviation by using core lexicons, *T function*, and *F function* is used in this study. *T function* is used to compute the similarity value between tree structures of two words in a corresponding pair. *F function* is used to compute the semantic similarity value between attribute value sets of two words in a corresponding pair. If the value of *T function* is over than threshold, the tree structures are similar and *F function* is used to

further compute the similarity between attribute sets. On the contrary, *F function* is not used.

As shown in Fig.5.1, the higher sememe in E-HowNet Sememe Structure is more abstract. On the contrast, the lower sememe is more specific. Therefore, the similarity degree between sememe pairs is high when the pairs are in the lower level. For example, the distance between *beast* and *animal* is the same as that between *thing* and *entity*. However, the pair *beast* and *animal* are in the lower level than the pair *thing* and *entity* in E-HowNet Sememe Structure. The pair *beast* and *animal* has higher similarity degree than the other pair. Therefore, the *C function* [2] follows this idea to compute the similarity of core lexicon between two words.

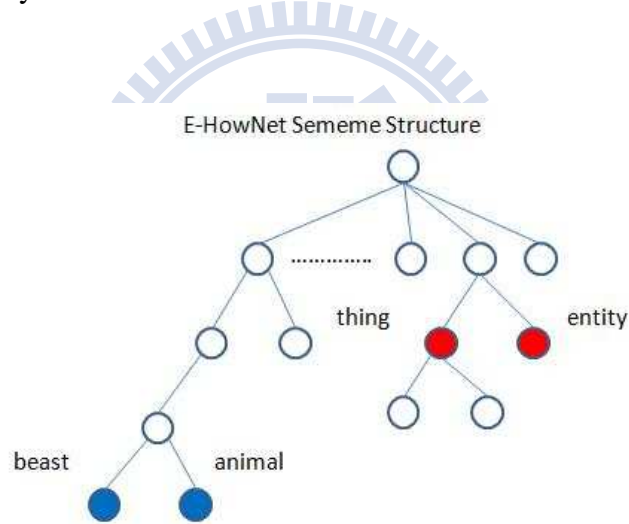


Figure 5.1: Sememe Structure in E-HowNet

$$C(w_1, w_2) = e^{-\alpha d} \times \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \quad \text{Eq 3}$$

Where

- e: Euler' s number
- $\beta$  and  $\alpha$ : smoothing factor
- h: common parent\_level for  $\langle W_1, W_2 \rangle$
- d: minimum distance between  $W_1$  and  $W_2$



## Tree structure computation

The framework structure (tree structure) is the attribute sets in E-HowNet, and they can be used to build up a tree as shown in Figure 2.1. Since similar things can be described by similar framework structure, their structures can be compared to compute the similarity degree. For example, books usually use the following attributes: title, author, publisher, and publication date. In a tree structure, a node in a higher level denotes more important for core lexicon (root is in level 0). Therefore, the similarity degree is decreased based on its level and the number of sibling nodes when the node is only existed in one tree. As shown in Eq 4,  $T(\text{tree1}, \text{tree2})$  compares the similarity of tree structure between two words. To explain the steps precisely, the study proposes Framework Structure Similarity Identification Algorithm (FSSI Alg) as follows.

$$T(\text{tree1}, \text{tree2}) = 5 - \sum_i \frac{5}{2^{L-1} \times |U|} \quad \text{Eq 4}$$

Where

i: tree node i at level L

U: node set includes i and its sibling nodes.

Algorithm2: Framework Structure Similarity Identification Algorithm (FSSI Alg)

**Symbols Definition:**

$C_{px}$ : the set of child nodes belongs to parent node  $x$  in tree  $P$ .

$C_{qy}$ : the set of child nodes belongs to parent node  $y$  in tree  $Q$ .

$U$ : uncheck set.

$i$ : tree node  $i$  at level  $L$ .

**Input:** Concept trees  $P, Q$ .

**Output:** Dissimilarity Tree Value (DTV).

Step 1: Find the smaller set between  $|C_{px}|$  and  $|C_{qy}|$  as  $U$ .

Step 2: For each node  $i$  in  $U$ ,

If  $i$  does not exist in the same level of the other tree, add  $\frac{5}{2^{L-1} \times |U|}$  to DTV

and terminate the process.

Else

Run FSSI Alg.

Step 3: Output Dissimilarity Tree Value.

**EXAMPLE 5.2:**

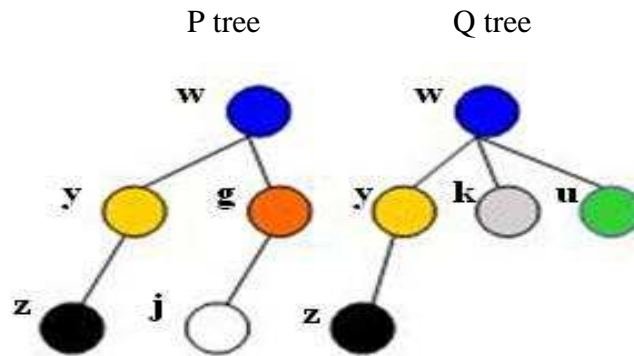


Figure 5.2: The framework of P, Q tree

As shown in Figure 5.2, P tree has fewer nodes than Q tree in the level 1. The nodes in the level 1 of P tree will be checked whether the nodes appear in the level 1 of Q tree. Node y appears in the level 1 of both trees. Second, there is only one leaf node in the second level. Node z appears in the level 2 of both trees. Third, node g does not appear in the second level of Q tree, and therefore the process is terminated.

$$T(P,Q) = 5 - \frac{5}{2^{1-1} \times 2} = 2.5$$

**Attribute set value computation**

As shown in Eq 5, the *F function* computes the semantic similarity between two words by using a set of attribute values. In Eq 5, the tree  $A_1$  has fewer nodes than the other tree  $A_2$ . Since the attribute set may be located in different levels, level weight is used as shown in Eq 6. If level of  $a_{1m}$  is larger than  $a_{2n}$ , the value of  $\text{level}(a_{1m}, a_{2n})$  is 0.5; If level of  $a_{1m}$  and  $a_{2n}$  are equal, the value of  $\text{level}(a_{1m}, a_{2n})$  is 1; If level of  $a_{1m}$  is smaller than  $a_{2n}$ , the value of  $\text{level}(a_{1m}, a_{2n})$  is 2.

$$F(A_1, A_2) = 5 \times \frac{\sum_{0 \leq m \leq |A_1|} (\text{Level}(a_{1m}, a_{2n}) \times (\text{Max}_{0 \leq n \leq |A_2|} (C(a_{1m}, a_{2n}))))}{|A_1|} \quad \text{Eq 5}$$

Where

$A_1, A_2$ : attribute set of  $W_1$  and  $W_2$  respectively

$\text{Level}(a_{1m}, a_{2n})$ : level weight of a certain attribute  $a_{1m}, a_{2n}$  for  $A_1$  and  $A_2$  sets

$$\text{Level}(a_{1m}, a_{2n}) = \begin{cases} 0.5, & \text{if the level of } a_{1m} > \text{ the level of } a_{2n} \\ 1, & \text{if the level of } a_{1m} = \text{ the level of } a_{2n} \\ 2, & \text{if the level of } a_{1m} < \text{ the level of } a_{2n} \end{cases} \quad \text{Eq 6}$$

**EXAMPLE 5.3:**

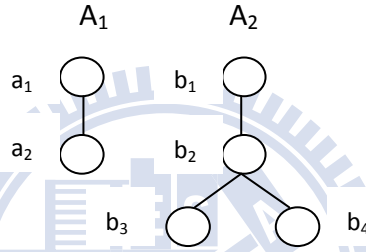


Figure 5.3: The framework of A1, A2 tree

As shown in Fig.5.3, the similarity values of  $a_1$  and each node  $b_x$  in  $A_2$  is calculated by Eq 3 and then  $C(a_1, b_1)$  is the maximum value among  $C(a_1, b_x)$  ( $x$  is from 1~4). Similarly,  $C(a_2, b_3)$  is the maximum value among  $C(a_2, b_x)$  ( $x$  is from 1~4). Since  $C(a_1, b_1)$  is the maximum value, and level of  $a_1$  and  $b_1$  are the same,  $\text{Level}(a_1, b_1)$  is 1. Different from  $\text{Level}(a_1, b_1)$ ,  $\text{Level}(a_2, b_3)$  is 0.5 because the level of  $a_2$  is higher than  $b_3$ .

$$F(A_1, A_2) = \frac{C(a_1, b_1) \times 1 + C(a_2, b_3) \times 0.5}{2}$$

**EXAMPLE 5.4:**

大甲狀元粽 三星上將梨

In this couplet, we calculate the semantic similarity value of corresponding pair “{大甲, 三星}”. The process of *E-HowNet Semantic Similarity Computation* is shown blow.

## Core lexicon computation

Table 5.2: Core lexicon of “大甲/三星”

|    | Concept representation | Sememe representation |
|----|------------------------|-----------------------|
| 大甲 | 鄉                      | 鄉                     |
| 三星 | 公司                     | 場所                    |

Since core lexicon of Concept represent and Sememe represent of these two words are different, it belongs to Case 4 in Eq 1.  $S(w_1, w_2)$  is used to compute the semantic similarity. In this stage, the *C function* is computed. In Eq 3,  $\alpha$  is set to 1.6,  $\beta$  is set to 0.16. The core meaning of sememe representation are “鄉” and “場所” and the distance of this two sememes is 7( $d=7$ ). Their first common parent node is in level 3( $h=3$ ), and therefore value of *C function* is 0.11.

## Tree structure computation

Table 5.3: Attributes annotation of Framework in “大甲/三星”

|    | Sememe representation   |
|----|---|
| 大甲 | {village 鄉: <u>quantifier</u> = {definite 定指}, <u>location</u> = {country 國家: <u>quantifier</u> = {definite 定指}, <u>location</u> = {Asia 亞洲}, <u>name</u> = {"台灣"} }, <u>name</u> = {"大甲"}}   |
| 三星 | {InstitutePlace 場所: <u>quantifier</u> = {definite 定指}, <u>location</u> = {country 國家: <u>location</u> = {continent 大陸: <u>quantifier</u> = {definite 定指}, <u>name</u> = {"亞洲"} }, <u>quantifier</u> = {definite 定指}, <u>name</u> = {"韓國"} }, <u>domain</u> = {economy 經濟}, <u>name</u> = {"三星"} } |

The words are underlined that are attributes, and they can be used to build up a tree structure as shown in Figure 5.4.

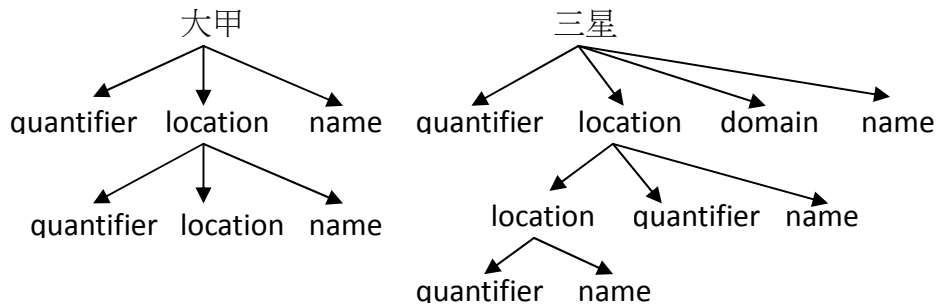


Figure 5.4: Tree of framework in corresponding pair “大甲”and “三星”

Since the tree of “三星” contains the the tree of “大甲”, value of  $T$  function is 5 (5-0). It is larger than 3, so *Attribute set value computation* is needed.

### Attribute set value computation

Table 5.4: Attribute values annotation of Framework in “大甲/三星”

|    | Sememe representation  |
|----|--|
| 大甲 | {village 鄉: quantifier = {definite  <u>定指</u> }, location = {country  <u>國家</u> : quantifier = {definite  <u>定指</u> }, location = {Asia  <u>亞洲</u> }, name = {" <u>台灣</u> " }}, name={" <u>大甲</u> "}}  |
| 三星 | {InstitutePlace 場所: quantifier = {definite  <u>定指</u> }, location = {country  <u>國家</u> : location = {continent  <u>大陸</u> : quantifier = {definite  <u>定指</u> }, name = {" <u>亞洲</u> " } }, quantifier = {definite  <u>定指</u> }, name = {" <u>韓國</u> " } }, domain = {economy  <u>經濟</u> }, name = {" <u>三星</u> " } } |

The words are underlined that are attribute values, and they can be used to build up a tree structure as shown in Figure 5.5.

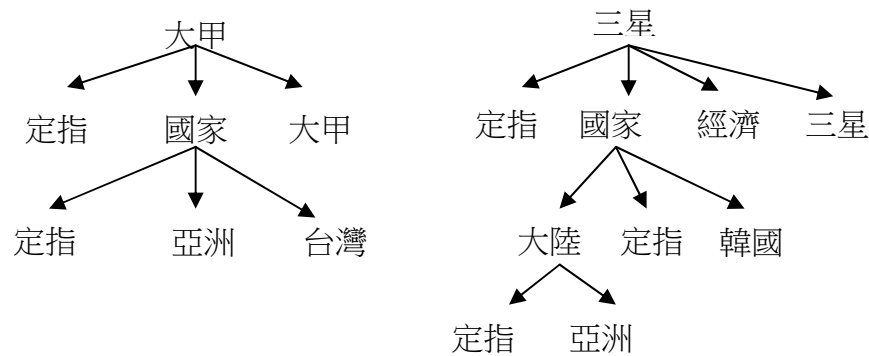


Figure 5.5: Tree of framework value in corresponding pair “大甲”and “三星”

$$F(P, Q) = \frac{(1+1+1) \times 1 + 0.444 \times 2}{4} = 4.86$$

Based on  $F$  function, its value is 4.86.

Finally, the semantic similarity value of corresponding pair “大甲” and “三星” is 3.4.

### 5.3 Semantic Similarity for Couplet

Based on semantic similarity value of corresponding pair, the semantic similarity of a couplet is computed by Eq 7.

$$\text{Semantic}(\text{line1}, \text{line2}) = \frac{1}{n} \sum_{i=1}^n \text{Sim}(W_{1,i}, W_{2,i}) \quad \text{Eq 7}$$

Where

n: number of corresponding pairs in a couplet

#### EXAMPLE 5.5:

大甲狀元粽 三星上將梨

In this couplet, the values of corresponding pairs are (大甲, 三星)= 3.4, (狀元, 上將)= 0.56, and (粽, 梨)=1.488. According to Eq 7, Semantic (line1,line2) is 1.815.

## Chapter 6 Couplet Evaluation

Based on the constraints of couplet composition, the couplet knowledge evaluation is divided into three parts: semantic meaning, tone usage, and word usage. Among these three constraints, the score of semantic meaning in a couplet is introduced in Chapter 5. In this Chapter, the tone score and word score are determined by score tables and described as follows. Finally, the couplet knowledge evaluation is defined by these scores.

In order to assist composers to analyze couplets, a couplet evaluation mechanism is proposed in this research to determine whether their composition is good or not. This idea is based on Chinese essay automatic scoring of Basic Competence Test for Junior High School Students [13] [14] [15] and English essay automatic scoring [3] in GRE or TOEFL. It is a fairer scoring method than human scoring since all the scorings are under the same condition. We adopt the same way to evaluate couplets. In this research, the Couplet Evaluation Mechanism includes semantic score, tone score, and word score. Its estimation equation is shown below:

$$\text{score}(\text{line1}, \text{line2}) = \alpha \times \text{semantic}(\text{line1}, \text{line2}) + \beta \times \text{tone}(\text{line1}, \text{line2}) + \gamma \times \text{word}(\text{line1}, \text{line2}) \quad \text{Eq 8}$$

The method for computing  $\text{semantic}(\text{line1}, \text{line2})$  is described in Chapter 5. Moreover, there are two score tables used to evaluate tone and word of couplets. The score tables are defined by the related attributes in **context information**. According to the importance of these constraints, the attributes in score tables have different weights. Moreover, in the tone constrains, the *ZP* is a basic couplet constraint and the



*opposite* is an advanced couplet constrain. Therefore we think that *ZP* is more important than *opposite*. The tone score table is defined according to the importance of each attribute and is shown below.

Table 6.1: Tone score table

**Tone Constraints**

| <b>Match Condition</b> | <b>ZP</b>        | <b>Opposite</b>  | <b>Score</b>     |
|------------------------|------------------|------------------|------------------|
|                        | <b>Non-match</b> | <b>Non-match</b> | <b>Non-match</b> |
| <b>Non-match</b>       | <b>Non-match</b> | <b>Match</b>     | <b>2</b>         |
| <b>Match</b>           | <b>Match</b>     | <b>Non-match</b> | <b>3.5</b>       |
| <b>Match</b>           | <b>Match</b>     | <b>Match</b>     | <b>5</b>         |

The word score table is a combination of *Repetition of words*, *Use of words*, and *POS consistency* attributes. In the word constrains, the “*Repetition of words*” and “*Use of words*” are more common than “*POS Consistency*” in the couplet. The more common the attribute is, the more important the attribute is.

Table 6.2: Word score table

**Word Constraints**

| <b>Match Condition</b> | <b>Repetition of words</b> | <b>Use of words</b> | <b>POS consistency</b> | <b>Score</b>     |
|------------------------|----------------------------|---------------------|------------------------|------------------|
|                        | <b>Non-match</b>           | <b>Non-match</b>    | <b>Same</b>            | <b>Non-match</b> |
| <b>Non-match</b>       | <b>Non-match</b>           | <b>Different</b>    | <b>Match</b>           | <b>1</b>         |
| <b>Non-match</b>       | <b>Match</b>               | <b>Different</b>    | <b>Non-match</b>       | <b>2</b>         |
| <b>Non-match</b>       | <b>Match</b>               | <b>Different</b>    | <b>Match</b>           | <b>3</b>         |
| <b>Match</b>           | <b>Match</b>               | <b>Same</b>         | <b>Non-match</b>       | <b>2</b>         |
| <b>Match</b>           | <b>Match</b>               | <b>Same</b>         | <b>Match</b>           | <b>3</b>         |
| <b>Match</b>           | <b>Match</b>               | <b>Different</b>    | <b>Non-match</b>       | <b>4</b>         |
| <b>Match</b>           | <b>Match</b>               | <b>Different</b>    | <b>Match</b>           | <b>5</b>         |

According to these tables, the tone score and word score can be determined by matching condition. In next section, these two score and semantic score are consolidated by a mechanism to evaluate knowledge couplet.

**EXAMPLE 6.1:**

## 大甲狀元粽 三星上將梨

To evaluate the couplet, the score concerning tone and word constraints are listed in Table 6.3. According to Tone table and Word table, tone(line1, line2) is 3.5 and word(line1, line2) is 5.

Table 6.3: Match Condition of 大甲狀元粽 三星上將梨

| Match Condition | ZP    | Opposite  | Repetition of words | Use of words | POS consistency |
|-----------------|-------|-----------|---------------------|--------------|-----------------|
|                 | Match | Non-match | Match               | Different    | Match           |

Finally, the Evaluation Score of a couplet can help us to distinguish level of a couplet. The couplet can be classified into five levels. The reason why we use this five-level scoring mechanism is that it is usually used in proficiency test such as The Japanese-Language Proficiency Test and General English Proficiency Test. We define level 5 as the highest, and level 0 is the lowest. Their score range is shown below:


Table 6.4: Evaluation Score range with couplet level

| Couplet Level | Evaluation Score range       |
|---------------|------------------------------|
| 5             | $4 \leq \text{score} \leq 5$ |
| 4             | $3 \leq \text{score} < 4$    |
| 3             | $2 \leq \text{score} < 3$    |
| 2             | $1 \leq \text{score} < 2$    |
| 1             | $0 \leq \text{score} < 1$    |

# Chapter 7 Experiment and Analysis

## 7.1 System Implementation

In this thesis, we implement the *Couplet Analysis System* based on the knowledge-based approach to generate couplet knowledge list and couplet evaluation table. The CAS is implemented with JAVA/JSP, MySQL, and drama 2.5. The interface of *Couplet Analysis System* (CAS) is shown in Figure 7.1. The composer inputs his/her composition and profile to this system. Then, in Figure 7.2 the system shows the list of couplet knowledge and couplet score.



KDE Lab NCTU

請輸入五言對聯:

中秋玉兔殿,七夕情人橋.

填寫使用者基本資料

使用者姓名 小陳

使用者心情 過節愉快

使用者地點 冬山河鵲橋上

開始分析

Figure 7.1: Interface of Couplet Analysis System which could be linked from [http://skywalf.no-ip.info:8080/Couplet/analysis\\_of\\_5.jsp](http://skywalf.no-ip.info:8080/Couplet/analysis_of_5.jsp)

對聯知識列表

| 對聯名稱        |         |    |    |       |    |    |
|-------------|---------|----|----|-------|----|----|
| 中秋玉兔殿 七夕情人橋 |         |    |    |       |    |    |
| 詞彙紀錄        |         |    |    |       |    |    |
| 上聯第一個詞彙     | 名稱      | 中秋 | 詞性 | Ndabd | 聲韻 | 平平 |
| 上聯第二個詞彙     | 名稱      | 玉兔 | 詞性 | Nba   | 聲韻 | 仄仄 |
| 上聯第三個詞彙     | 名稱      | 殿  | 詞性 | Nab   | 聲韻 | 仄  |
| 下聯第一個詞彙     | 名稱      | 七夕 | 詞性 | Ndabd | 聲韻 | 仄仄 |
| 下聯第二個詞彙     | 名稱      | 情人 | 詞性 | Nab   | 聲韻 | 平平 |
| 下聯第三個詞彙     | 名稱      | 橋  | 詞性 | Nab   | 聲韻 | 平  |
| 對聯規則        |         |    |    |       |    |    |
| 上仄下平        | 符合      |    |    |       |    |    |
| 平仄相對        | 符合      |    |    |       |    |    |
| 重字相對        | 符合      |    |    |       |    |    |
| 上下聯不同字      | 符合      |    |    |       |    |    |
| 詞性相對        | 符合      |    |    |       |    |    |
| 對聯語義架構      |         |    |    |       |    |    |
| 上聯語義        | 節,月亮,房屋 |    |    |       |    |    |
| 下聯語義        | 節,人,橋樑  |    |    |       |    |    |

對聯評分表

|      |                    |
|------|--------------------|
| 對聯名稱 | 中秋玉兔殿 七夕情人橋        |
| 聲韻評分 | 5.0                |
| 用字評分 | 5.0                |
| 語義評分 | 2.5271075394393128 |
| 總評分  | 4.010843015775725  |
| 等級評比 | 5                  |

Figure 7.2: Result of Couplet Analysis System

## 7.2 Experiment Result and Discussion

The *E-HowNet Semantic Similarity Computation* doesn't take the semantic relations between words in a sentence into consideration. In contrast, it considers semantic relations of correspond pairs. Therefore, we assume the semantic meanings of words in a sentence are correlated in this research.

To evaluate the experiment, Ping-Ze database is used to find out the tone of

words. However, some words have multiple pronunciations. The problem is solved by choosing a tone of a word that has the highest usage rate in *Chinese Dictionary of Ministry of Education*.

The test data in the experiment is from “the 7<sup>th</sup> global couplet composition(全球徵聯)”. The purpose of the 7<sup>th</sup> global couplet composition [16] is giving a second sentence “三星上將梨”, and composer compose a corresponding first sentence. Three stages of these couplets are collected: passing the primary selection, passing the double selection and passing the final selection. The total number of collected couplets is 2510. After deleting repeated couplets which are composed by different composers, the number of couplets is 1778. There are 1213 couplets passing the primary selection; 494 couplets passing the double selection; 71 couplets passing the final selection. Owing to the characteristics of collected couplets, the collected couplets are categorized into three parts by CAS: low(n=1213), middle(n=494), top(n=71). The evaluation criterion is precision rate. For example, the top-71 in CAS is compared to the couplets passing the final selection. The comparison result between CAS and expert is shown in Table7.1. The results show that 814 couplets of low-1213 in CAS are passing the primary selection; 164 couplets of middle-494 in CAS are passing the double selection; 11 couplets of top-71 in CAS are passing the final selection. The precision of CAS is shown in Table 7.2.

Table 7.1: Expert level compare with the CAS score

| <b>system<br/>expert</b> | Low | Middle | High |
|--------------------------|-----|--------|------|
| 1                        | 814 |        |      |
| 2                        |     | 164    |      |
| 3                        |     |        | 11   |

Table 7.2: Precision of Couplet Analysis System

| <b>Level of expert grading</b> | <b>Precision</b> |
|--------------------------------|------------------|
| 1                              | 67.1%            |
| 2                              | 33.2%            |
| 3                              | 15.5%            |

In this thesis, we found that the meaning of the word “三星” in the second line is a Korean company name in E-HowNet. However, the meaning of this word should be the name of a township in Taiwan. For evaluating the applicable of the experiment, this research modified the meaning of the word and reran the experiment. The result is shown in Table 7.4 and Table 7.5.

Table 7.3: Experiment result when “三星” is name of a township

| <b>system<br/>expert</b> | Low | Middle | High |
|--------------------------|-----|--------|------|
| 1                        | 872 |        |      |
| 2                        |     | 154    |      |
| 3                        |     |        | 7    |

Table 7.4: Precision of Couplet Analysis System when “三星” is name of a township

| Level of expert grading | Precision |
|-------------------------|-----------|
| 1                       | 71.9%     |
| 2                       | 31.2%     |
| 3                       | 9.9%      |

Contrasting with these two results, we discover when “三星” is a Taiwan township name rather than a Korean company name, most of the couplets are evaluated higher. It means that most authors know the meaning of “三星” should be a Taiwan township.

Contrasting with these two results, we discover their precisions are similar. The reason is the tree structures are similar as shown in Figure 7.3. Moreover, the attribute sets are similar. Therefore, values of *T function* and *F function* are similar.

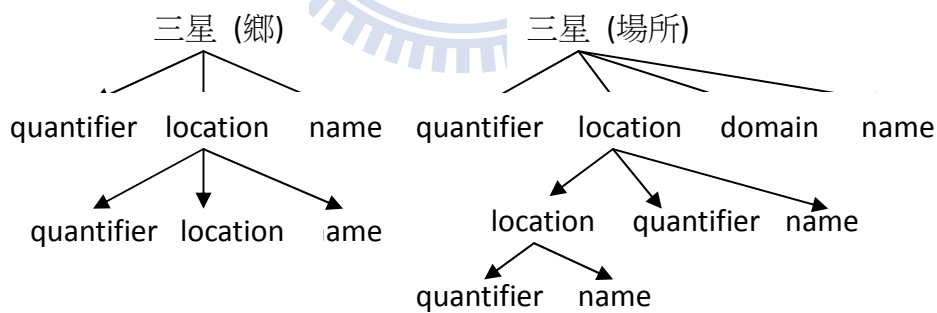


Figure 7.3: tree structures of different senses on “三星”

Since each level has different number of couplets, the results may have some problems. Therefore, we repeat the experiment 10 times by using random sampling. The samples in each time are 71 (same with level 3). The results are shown in Table 7.5. The number in parentheses is the correct number between CAS and domain

experts. Total precision by using random sampling is 42%, which is better than the precision by using overall sampling (37%).

Table 7.5: Precision by using random sampling

| Run     | Level 1   | Level 2   | Level 3   |
|---------|-----------|-----------|-----------|
| 1       | 38% (27)  | 39% (28)  | 46% (33)  |
| 2       | 37% (26)  | 41% (29)  | 46% (33)  |
| 3       | 37% (26)  | 37% (26)  | 46% (33)  |
| 4       | 42% (30)  | 45% (32)  | 48% (34)  |
| 5       | 39% (28)  | 37% (26)  | 48% (34)  |
| 6       | 30% (21)  | 35% (25)  | 48% (34)  |
| 7       | 38% (27)  | 37% (26)  | 44% (31)  |
| 8       | 39% (28)  | 45% (32)  | 48% (34)  |
| 9       | 44% (31)  | 42% (30)  | 48% (34)  |
| 10      | 34% (24)  | 38% (27)  | 51% (36)  |
| Total   | 38% (268) | 40% (281) | 47% (336) |
| Average | 42%       |           |           |

## Error analysis

The reason of error in CAS can be divided into two categories: Sense tagging error and Ontology limitation.

### Sense tagging error:

#### 1) Non-related semantic meaning

This research doesn't consider the inner sentence word relationships. For example: In couplet “七堵雙人枕 三星上將梨.” Word “七堵” is not related to word “雙人枕”. This couplet should have a low score because of the non-related



semantic meaning. But we get a high score because we do not consider the inner sentence word relationships..

2) Different corresponding pair score

Example: 石碇桂花釀 三星上將梨. In this couplet, the semantic meaning of “石碇/三星” and “釀/梨” are similar, but the semantic meaning of “桂花” and “上將” are different. The semantic score of this couplet in our system is not bad because it is calculated by averaging the corresponding pairs. However this couplet is estimated lower, if it is estimated by domain experts.

3) Low semantic score

The tone score and word scores are high, but semantic score is very low. For Example: 寶島財神廟 三星上將梨. This couplet matches all the tone and word constraints but the semantics meanings are not correlated.

**Ontology limitation:**

4) Unknown word

Example: 一品大員外 三星上將梨. The words, “一品” and “外”, are not specified in E-HowNet, therefore the semantic score of this couplet is calculated by a pair “大員” / “上將”.

5) Unknown tone.

Example: 六甲神仙煲 三星上將梨. In this example, the tone of the word “煲” cannot be identified. Since it is the last character of the first sentence, the system cannot identify the value of FZSP or Reverse FS of context information. The tone score is 0 in our system.

In our future work, the sense tagging error can be solved by using weight

learning scheme for features. Besides, web-mining approach can solve unknown word and unknown tone issues.



## Chapter 8 Conclusion

For knowledge extraction and evaluation, the thesis develops *Couplet Analysis System*, which contains three processes: word segmentation, semantic similarity computation, and knowledge evaluation. To deal with word segmentation, the *Heuristic Rule-Based Word Segmentation* is proposed, which uses the *sentence pattern, known word, longest word, proper noun, and allusion* to decide the word segmentation in a sentence. For semantic similarity computation, the *Heuristic-Based Approach* is proposed to solve WSD problem, and the *E-HowNet Based Semantic Similarity Computation* is proposed to calculate couplet semantic score between two sentences in a couplet. At last, the couplet knowledge about tone, word and semantic meaning are extracted and graded. Couplet composers can understand the advantage and disadvantage about their compositions through this system.

The experiment results show that the system can analyze the couplet knowledge well. The precision of this system is 42%. In the near future, the system can be further improved by identifying the semantic coincidence in a sentence and identifying unknown word.

## Reference

- [1] A. Islam and D. Inkpen. “Semantic text similarity using corpus-based word similarity and string similarity” ACM Transactions on Knowledge Discovery from Data(2008):1–25.
- [2] L.L. Dai, B. Liu, Y. Xia, and S.K. Wu. “Measuring Semantic Similarity between Words Using HowNet” International Conference on Computer Science and Information Technology(2008):601-605.
- [3] Y. Attali and J. Burstein. “Automated Essay Scoring With e-raterV.2” The Journal of Technology, Learning, and Assessment(2006)
- [4] A. Wu and Z. Jiang. “Word segmentation in sentence analysis” Proceedings of the 1998 International Conference on Chinese Information Processing (1998):169-180.
- [5] R. Bruce and J. Wiebe. “Word-sense disambiguation using decomposable models” Proceedings 32nd Annual Meeting of the Association for Computational Linguistics(1994):139-146.
- [6] C. Leacock, M. Chodrow, and GA. Miller. “Using corpus statistics and WordNet relations for sense identification” Computational Linguistics(1998):147 – 165.
- [7] E. Agirre, O.L.D. Lacalle, and A. Soroa. “Knowledge-Based WSD on Specific Domains: Performing Better than Generic Supervised WSD” Proceedings of the 21st international joint conference on Artificial intelligence(2009):1501-1506.
- [8] K.J. Chen, S.L Huang, Y.Y Shih, Y.J. Chen. “Extended-HowNet- A Representational Framework for Concepts” Ontologies and Lexical Resources IJCNLP-05 Workshop(2005).
- [9] L. Jiang and M. Zhou. “Generating Chinese Couplets using a Statistical MT Approach” Proceedings of the 22nd International Conference on Computational Linguistics(2008):377-384.
- [10] R. Navigli. “Consistent validation of manual and automatic sense annotations with the aid of semantic graphs” Computational linguistics(2006):273–281.
- [11] Y.S. Chan and H.T. Ng. 2007. “Domain adaptation with active learning for word sense disambiguation” Proceedings of the 45th annual meeting on Association for Computational Linguistics(2007):49-56.
- [12] C. Kruengkrai, K. Uchimoto, J. Kazama, Y. Wang, K. Torisawa, and H. Isahara, “An error-driven word-character hybrid model for joint chinese word segmentation and pos tagging” Proceedings of ACLIJCNLP(2009).
- [13] 蔡沛言, “Automated Chinese Essay Scoring System : Generation 、 Selection 、

- Evaluation” 國立交通大學資訊科學系研究所, 碩士論文(June 2005).
- [14] 粘志鵬, “Automated Chinese Essay Scoring System Based on Support Vector Machine” 國立交通大學資訊科學與工程研究所, 碩士論文(June 2006).
- [15] 林信宏, “A Bayesian Based Chinese Essay Scoring System” 國立交通大學資訊科學與工程研究所, 碩士論文(June 2006).
- [16] 國立東吳大學中文系, 全球徵聯, < <http://art.pch.scu.edu.tw/cc34.htm>>.
- [17] 嚴恩萱, 嚴考亮, 輕鬆學對聯, 廣州:暨南大學出版社(2005).
- [18] 陸偉廉, 對聯知識導讀, 安徽:黃山書社(1989).
- [19] 許清雲, 近體詩創作理論, 台北:洪葉文化(2003).
- [20] 彭載衍, 張俊盛, “中文辭彙歧異之研究 – 斷詞與詞性標示”, 第六屆中華民國計算語言學研討會論文集(1993):173-194.
- [21] 陳克健, 黃淑齡, 施悅音, 陳怡君, “多層次概念定義與複雜關係表達 – 繁體字知網的新增架構” 漢語詞彙語義研究的現狀與發展趨勢國際學術研討會 (2005).
- [22] 鄭亞斌, 曹嘉偉, 劉知遠, “Couplet System Based on Maximum Matching and Markov Model” 第四屆全國學生計算語言學研討會會議論文集(2008).

