The Polysemy of PO in Mandarin Chinese

Harvey Hsin-chang Ho
National Taiwan Normal University
harveyhcho@gmail.com

Abstract

The present paper notes that the lexical item PO, literally meaning ‘to break’, bears multiple semantic imports in Mandarin Chinese. Given the lack of well-documented research on the semantics of the lexical item, this paper aims to explore the various meanings of PO. By examining the collocations of PO, thirteen meanings are identified, with predicative and attributive senses. It is proposed that the manifold meanings are interrelated with each other and that several meanings are derived from the core verbal meaning of the lexical item. Three generalized metaphors are observed to assume a mediating role in the semantic extensions of PO. In light of the semantic relatedness of the various meanings, the polysemous nature of the lexical item PO is substantiated.

Key words: PO, polysemy, semantic extension, lexical semantics

1. Introduction

Since a growing number of psychological studies shed new light on human cognition in 1970s, the field of semantics has witnessed flourishing cognitive-oriented approaches to semantic representations of lexicon and grammar—especially lexical semantics and cognitive semantics (Rosch 1973, 1977, 1978, Lakoff and Johnson 1980, Lakoff 1987, 2002, Johnson 1987, Langacker 1987, 1990, 1999, Geerearts 1993, Talmy 1985, 2000a,b, Taylor 1989, 2002a,b, 2003, among others). These cognitive-theoretic proposals have spawned a voluminous literature pertaining to conceptualization, categorization, semantic extension, and grammaticalization of polysemous lexical items in Mandarin Chinese, such as *guo* ‘to cross’ (Wang 2002, Hsiao 1997, 2003, Wu 2003), *gei* ‘to give’ (Huang 2004), and *kai* ‘to open’ (Tsai 2006).

When it comes to the issues of polysemy, one point meriting our note is the distinction between homonymy and polysemy. Homonymy refers to the relation between different lexical entries which have unrelated meanings but accidentally exhibit an identical linguistic form, orthographic or phonetic (Ravin and Leacock 2000). A polysemous word, in contrast, is one single lexical item which bears different, but etymologically related, meanings (Lyons 1995, Ravin and Leacock 2000). The English word *break* is a case of polysemy (Tang 2004), and *breken* ‘to break’ in Dutch also has multiple meanings (Kellerman 1978). The present paper observes that the lexical item PO, literally meaning 'to break', seems to bear versatile semantic imports in Mandarin Chinese. A question arises as to whether PO is a polyseme or two or more homonyms in Mandarin Chinese. It is noted that studies on the semantics of the lexical item PO, if any, are underrepresented, or even undocumented. Hence, this study aims to probe into the manifold meanings of PO. A cognitive approach will be drawn on to explicate the relations between different semantics of PO and to substantiate the polysemous nature of the lexical item.

This paper is organized as follows. Section 2 is concerned with the research background of the present analysis. Section 3 deals with the various senses of PO and proposes a possible account for the semantic relatedness of the manifold meanings. Section 4 concludes this paper.

2. Research Background

In linguistics, the theory of prototypes has exerted a momentous impact on lexical semantics and cognitive linguistics (e.g. Rastier 1999, Chu and Chi 1999, Ravin and Leacock 2000). The prototypical category framework has laid theoretical foundations for research on polysemy, and mechanisms for meaning extension have also derived much inspiration from prototypes. The prototypical theory and apparatus for semantic extension are reviewed below.

2.1 The Prototypical Category Theory

The human kind seems to have an innate ability for categorization; for example, our brain divides the world into two primary types of entities, things that exist and situations that take place (Huang, Li, and Li 2006). Frameworks for human's categorization include the classical approach, the prototypical approach, and the relational approach.[1] Among them, the notion of prototypes is adopted in this paper.

Prototypes are amenable to two interpretations. The concept of prototypes is reminiscent of the renowned American psychologist Eleanor Rosch (1973, 1977, 1978). Rosch introduces the role of prototypes to elucidate human's categorization. People categorize objects on the basis of the resemblance between the objects and the prototypical members of the category. According to Rosch (1978:36), prototypes can be defined as the 'clearest cases of category membership defined operationally by people's judgments of goodness of membership in the category'. A prototype of a category is thus viewed as a salient exemplar of the category. Some instances of a category are more typical than others and hence emerge in human's mind more easily. For example, *robin* is a representative, prototypical instance of the category

[1] For detailed information of the classical approach, please refer to Katz and Fodor (1963), and for the relational approach, please see Evens (1988) and Fellbaum (1998).

BIRD in English,[2] while *penguin* is not a central, salient case.

Alternatively, prototypes are construed as an abstraction, a mental representation, rather than as a particular, concrete referent or instance. Lakoff (1987, idealized cognitive models (ICMs)), for instance, puts forth a prototypical concept of such a type—the cluster concept. The cluster concept consists of several cognitive models. For example, the meaning of MOTHER comprises the following cognitive models: the birth model, the genetic model, the nurturance model, the marital model, and the genealogical model. MOTHER forms a radial conceptual model; it has a central category where all the above models converge, as well as peripheral categories where fewer models congregate.

One point of categories merits our note here. Categories are not homogeneous; they are characterized by a prototype, with core and peripheral members, and fuzzy boundaries (Rosch 1973, 1977, 1978). Membership in a category is not contingent on whether an entity possesses all the attributes of a category. Rather, it is the degrees of family resemblances that link category members together (proposed by Wittgenstein 1953).

It has been maintained that our categorization hinges much on the structure of the outside world (Johnson 1987, Lakoff 1987). There are three levels of categories, i.e. basic-level, superordinate, and subordinate categories. Above and below basic-level categories exist superordiate and subordinate categories, respectively. The former are more abstract and embracing than the latter. For instance, DOG is a basic-level category. Superordinate to it is the category of ANIMAL, and the category of COLLIE, for example, is subordinate. The relations of the three levels of categories form a hierarchical structure of our language. Shifts from basic-level categories to superordinate ones lead to generalizations, while specifications are achieved from basic-level to subordinate categories.

2.2 Semantic Extension

Polysemy is a consequence of lexical semantic evolution towards different but related directions; different meanings are linked in terms of the semantic relatedness. Two traditional concepts assume a mediating role in semantic extension, viz. metaphor and metonymy, and avenues leading to polysemy involve semantic radiation and meaning chain.

Metaphors are one of the major mechanisms contributing to semantic change (e.g. Bybee and Pagliuca 1985, Sweester 1986, 1990). Metaphorical extension refers to the mappings across conceptual domains, from the source domain to the target, in which entities exhibit resemblances (Lakoff and Johnson 1980). It has been proposed that metaphors are grounded on our embodied experiences of the world and constitute part of our conceptual system (Lakoff and Johnson 1980, Johnson 1987, Lakoff 2002). In addition to metaphors, metonymy also accounts for semantic change. It refers to the process of establishing

[2] Prototypical exemplars of a category may be subject to cultural-specific differences.

associations between entities within a given conceptual structure (e.g. Taylor 1989, Hopper and Traugot 1993).

One path to metaphoric and metonymic extension is semantic radiation. Semantically, radiation is the process in which secondary meanings evolve from the central, core meaning in every possible uni-direction like rays (cf. Lakoff 1987, Langacker 1990). The core meaning is the prototype, from which different meanings are derived from. Nonetheless, the radial process cannot satisfactorily explain all the semantic change. For a number of words, the secondary meanings evolving from the core may become a hub for further semantic derivation, which may in turn undergo onward semantic evolution. Such a route to semantic extension is named meaning chain (cf. Lakoff 1987, Tayler 1989, Langacker 1990).

3. The Analysis

The present paper probes into the manifold meanings of the lexical item PO in Mandarin Chinese. PO is a productive word in Mandarin Chinese. Based on Academia Sinica Balanced Corpus of Modern Chinese[3] and Chinese GigaWordCorpus,[4] 16,448 tokens of PO were retrieved in total (219 tokens from the former corpus, and 16,229 tokens from the latter). Due to the limit of time and space, nonetheless, this paper only examines the meanings of 200 tokens, but simultaneously has recourse to Lü (1999) as reference to bridge a gap that the limited number of tokens analyzed might leave.

3.1 Multiple Meanings of PO

This paper notes that the lexical item PO has two syntactic categories, i.e. as a verb and as an adjective. The semantics of a verb and of an adjective can be identified by probing into their collocations. Based on the present data, the senses of PO as a verb can be classified into ten types, and three meanings are singled out for PO as an adjective, as exemplified below:

(1) A verb meaning ‘to damage an intact physical entity/substance’
 a. 窗戶破了 Chuanghu po-le. ‘The window broke.’
 b. 打破玻璃 dapo boli ‘to break glass’

[3] Academia Sinica Balanced Corpus of Modern Chinese (Sinica 3.0) encompasses 5 million words in Mandarin Chinese. The website for the corpus is http://www.sinica.edu.tw/SinicaCorpus/.

[4] Subsumed under the Sketch Engine (abbreviated as SkE, also known as Word Sketch Engine http://www.sketchengine.co.uk/), Chinese GigaWordCorpus covers 1 billion words in Mandarin Chinese and is the largest corpus of Modern Mandarin Chinese at present. The website for the corpus is http://corpora.fi.muni.cz/chinese_all/.

(2) A causative verb meaning ‘lit. to split an entity into two pieces’

a. 破門而入 po men er ru ‘lit. to split the door into two pieces and enter (a room); to burst into a room’

b. 破繭而出 po jian er chu ‘lit. to tear a cocoon apart and go out’

c. (乘風)破浪 po lang ‘lit. to split waves of ocean into two parts’

(3) A verb meaning ‘to scrape the intact surface (of skin)’

a. (磨)破皮 po pi ‘to scrape the skin’

b. 嘴破 zui po ‘The skin of the oral cavity is wounded; stomatitis.’

c. 腸破(肚流) chang po ‘The skin of the bowel is wounded’

(4) A causative verb meaning ‘to eradicate (an idea, belief, custom, etc.)’

a. 破除迷思 pochu misi ‘to break the myth’

b. 破舊(立新) po jiu ‘to eliminate the old custom’

(5) A causative verb meaning ‘to disobey (a rule, precedent, convention, etc.)’

a. 破戒 po jie ‘to break a religious precept’

b. 破例 po li ‘to make an exception’

(6) A causative verb meaning ‘to surpass (a checkpoint, record, etc.)’

a. 破關 po guan ‘to go through a checkpoint’

b. 破紀錄 po jilu ‘to break the record’

(7) A causative verb meaning ‘to defeat (enemies)’

a. 破敵 po di ‘to defeat enemies’

b. 破除重圍 pochu chongwei ‘to defeat a multitude of enemies’

(8) A causative verb meaning ‘to expend (money)’

a. 破費 po fei ‘to expend one’s money’

b. 破產 po chan ‘to go bankrupt’

(9) A causative verb meaning ‘to uncover (a fact)’

a. 破案 po an ‘to solve a criminal case’

(10) A causative verb meaning ‘to end a situation’

a. 破涕為笑 po ti wei xiao ‘to turn tears into smiles’

b. 打破僵局 dapo jiangju ‘to break the ice’

c. 打破沉默 dapo chenmo ‘to break the silence’

(11) An adjective meaning ‘broken, ragged’

a. 破舊的 pojiu de ‘tattered, ragged’

b. 破衣服 po yifu ‘ragged clothes’

(12) An adjective meaning ‘worthless’

破玩意兒 po wanyier ‘worthless stuff’

(13) An adjective meaning ‘lousy’

他的中文很破 Ta de zhongwen hen po. ‘His Chinese is very lousy.’

3.2 Semantic Relatedness

The various meanings of the item PO are elaborated on in the following subsections. The ten verbal meanings of PO are explicated first, followed by accounts of the semantic imports of PO as an adjective.

3.2.1 A Verb Meaning 'to Damage an Intact Physical Entity/Substance'

The verb PO bears the core meaning 'to damage an intact physical entity/substance'. The entity in this case is concrete, tangible, and most importantly, breakable. The breakable entity remains 'intact' before it is affected by the action of breaking. The action of breaking causes the entity to undergo change of state, and the degree of affectedness is high.

A clear, typical instance of a brittle object is a window, as in (14a). The collocate 窗戶 (*chuanghu* 'window') with the verb PO comes into human mind easily; the meaning 'to damage an intact physical entity/subtance' is prototypical and thus stands at the center of semantic structure of the lexical item PO. In (14b), 玻璃 (*boli* 'glass') is substance, of which a window is made. The meaning of *boli* alternates with that of a discrete object through metonymy.[5]

The expressions in (14a-b), despite containing the same lexical item PO, denote different aspectual interpretations. The word PO in (14a) is an intransitive verb and has an inchoative meaning.[6] In (14b), *dapo* 'hit-broken' is a resultative verb compound; PO is a complement of the verb 打 (*da* 'hit') and denotes the resultative state caused by the action that the predicate depicts.

(14) A verb meaning 'to damage an intact physical entity/substance' (=(13), repeated here for ease of reference and discussion)
a. 窗戶破了 Chuanghu po-le. 'The window broke.'
b. 打破玻璃 dapo boli 'to break glass'

3.2.2 A Causative Verb Meaning 'to Split an Entity into Two Pieces'

As derived from the core meaning, the verb PO can collocate with an entity that could be split into two pieces. Such a collocation bears the literal meaning 'to split an entity into two parts'. In (15a), a door can be viewed as a visible PHYSICAL BARRIER from one space to another. When one has the desire to enter another space, one has to remove the barrier before the desire can be realized. If the space/room has a door and the door is closed, one

[5] The count/mass alternation for nouns is one type of polysemous variation, a case of metonymy.
[6] For discussion of causative/inchoative alternation of the word *break* in English, the reader is referred to Holmes (1999), for example.

normal way of entering the room is to ‘open’ the door by unlocking the door and using the doorknob or handle. If one splits the door into two pieces and abruptly enters the room through the door barrier, instead of following the normal way of entering the room, then the extended meaning of 破門而入 (*po men er ru*) obtains, i.e. ‘to burst into a room’.

This is the case of (15b). The expression 破繭而出 (*po jian er chu*) literally means ‘to tear a cocoon apart and go out’. When a silkworm is wrapped in a cocoon, it must tear the cocoon into two parts before it transforms into a butterfly and flies out of the cocoon. According to Lakoff and Johnson (1980), our conceptual system is grounded on our embodied experiences of the world. We view our body as a PHYSICAL ENTITY, separated from the world by the surface of (the skin) of our body, like a concrete PHYSICAL CONTAINER with an inside and an outside. Likewise, a cocoon can be regarded as a concrete PHYSICAL CONTAINER, which constrains a silkworm. When one strives to get free from a container constraint, whether concrete or abstract, one has to split apart the container constraint, just as a silkworm does to the cocoon. Along the thread of thought, the metaphorical meaning of the expression is derived.

A figurative meaning can be derived as well. Human beings impose artificial boundaries onto physical phenomena and consider them to be discrete as individuals themselves are (Lakoff and Johnson 1980). Along this line, the waves of ocean can be split into two parts as well, and the extended meaning of (15c) (乘風)破浪 (*po lang* ‘lit. to split waves of ocean into two parts’) emerge.

(15) A causative verb meaning ‘to split an entity into two pieces’

a. 破門而入 po men er ru ‘lit. to split the door into two pieces and enter (a room); to burst into a room’

b. 破繭而出 po jian er chu ‘lit. to tear a cocoon apart and go out’

c. (乘風)破浪 po lang ‘lit. to split waves of ocean into two parts’

3.2.3 A Verb Meaning ‘to Scrape the Intact Surface (of Skin)’

As suggested above, our body is regarded as a concrete PHYSICAL CONTAINER with an inside and an outside. Just as the surface of a cup, which is made of brittle glass, can be damaged, our skin as the surface of the container, our body, is delicate and can be damaged as well, as in (16a) (磨)破皮 (*po pi* ‘to scrape the skin’). Since a container has an inside and an outside, the inside and the outside have a surface, both of which can be damaged. The expression 嘴破 (*zui po*) denotes that the skin of the inside, i.e. the oral cavity, is wounded, namely stomatitis. In (16c), a bowel can be viewed as a smaller container inside the larger container, our body, and 腸破(肚流) (*chang po*) means that the skin of the bowel is wounded.

Other examples involving human body are shown in (17a-b). With respect to 破相 (*po* *xiang*), a human face is regarded as ‘intact’ if there is no scar on it, just as a window is viewed as intact before it is broken or marred. When a human face is marred by a scar, the expression is used. Also, the concept of intactness applies not only to human body but also to the notion of virginality. The metaphor is VIRGINALITY IS A PHYSICAL ENTITY, which is closely linked with human body and can be damaged once one has sex, as in (17b). The meaning of PO in 破身 (*po shen*) might be extended through the avenue shown in Figure 1, i.e. from a brittle entity through human body to virginality.

(16) A verb meaning ‘to scrape the intact surface (of skin)’
a. (磨)破皮 po pi ‘to scrape the skin’
b. 嘴破 zui po ‘The skin of the oral cavity is wounded; stomatitis.’
c. 腸破(肚流) chang po ‘The skin of the bowel is wounded’
(17) a. 破相 po xiang ‘to be marred by a scar on the face’
b. 破身 po shen ‘to lose virginality’

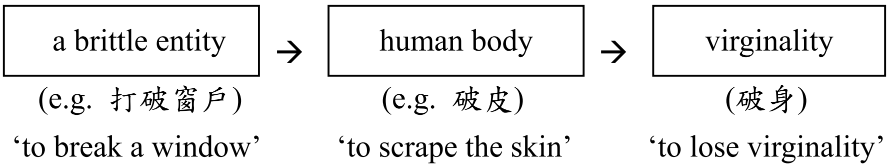


Figure 1. Semantic Extension of *po shen*

3.2.4 A Causative Verb Meaning ‘to Eradicate (an Idea, Custom, etc.)’

In addition, abstract ideas, beliefs, and customs can be metaphorically viewed as a concrete entity—AN IDEA IS A PHYSICAL ENTITY. Just as a glass, a concrete object, can be broken into pieces, a myth, imaged as a physical entity, can also be broken into pieces by the action depicted by the verb PO, as shown by 破除迷思 (*pochu misi* ‘to break the myth’) in (18a). The metaphoric extension also holds true of an old custom, as in (18b) 破舊(立新) (*po jiu* ‘to eliminate an old custom’). When an object is smashed into ‘pieces’, it undergoes tremendously high degree of affectedness. Its state is changed vastly, or even completely, and thus is different from the original shape (and/or nature); that is to say, the object is no longer what it was. Along this thread of thought, the extended meaning of PO ‘to eradicate’ is yielded.

(18) A causative verb meaning ‘to eradicate (an idea, belief, custom, etc.)’
a. 破除迷思 pochu misi ‘to break the myth’
b. 破舊(立新) po jiu ‘to eliminate the old custom’

3.2.5 A Causative Verb Meaning ‘to Disobey (a Precedent, Convention, etc.)’

When one is put in a container, one’s demeanor and action are restrained. A rule, precedent, or convention can thus be treated as a concrete container that constrains our conduct, i.e. the metaphor A RULE IS A PHYSICAL CONTAINER. For example, a religious precept constrains one’s way of living, such as interdicting one from doing something, eating something, or saying something. When one does not follow the religious precept, it is imaged that one breaks the container constraint, as in (19a) 破戒 (*po jie* ‘to break a religious precept’). For (19b) 破例 (*po li* ‘to make an exception’), if one is constantly constrained by conventions and follows them when one acts or handles matters. Once one does not act in accordance with the conventions one used to follow, one makes an exception to the conventions, i.e. the meaning of (19b).

(19) A causative verb meaning ‘to disobey (a rule, precedent, convention, etc.)’
a. 破戒 po jie ‘to break a religious precept’
b. 破例 po li ‘to make an exception’

3.2.6 A Causative Verb Meaning ‘to Surpass (a Checkpoint, Record, etc.)’

In Figure 2, there are 10 figures at the horizontal axis. The highest score among the ten figures is 75 points. The expression in (20b) 破紀錄 (*po jilu* ‘to break the record’) can be used only when a figure exceeds 75 points in this case. The noun 紀錄 (*jilu* ‘record’) can not refer to any score in the document but exclusively to the highest score recorded in the past. Hence, the expression 紀錄 (*jilu* ‘record’) has a similar meaning to 關 (*guan* ‘checkpoint’) in (20a) 破關 (*po guan* ‘to go through a checkpoint’)—a point for check or reference. By drawing on the metaphor A CHECKPOINT IS A PHYSICAL BARRIER, we can proffer a possible account for the collocations 破關 and 破紀錄 and the extended meaning of the verb PO. Since a checkpoint is regarded as a barrier, when one reaches the checkpoint and proceeds forward or upward through it,[7] one breaks through the checkpoint barrier.

(20) A causative verb meaning ‘to surpass (a checkpoint, record, etc.)’
a. 破關 po guan ‘to go through a checkpoint’
b. 破紀錄 po jilu ‘to break the record’

[7] FORWARD and UPWARD here involve an orientational metaphor MORE IS UP/FORWARD. The more upward a score goes, the higher it is.

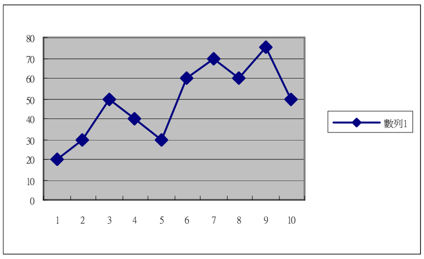


Figure 2. PO ‘to surpass’

3.2.7 A Causative Verb Meaning ‘to Defeat (Enemies)’

Enemies can form a line, a wall, or even a circle. Thus, they can be viewed as a line of barrier or a wall of barrier as well, as shown by the expression 四面圜敵 (*si mian huan di* ‘to be surrounded by enemies’). The metaphor is ENEMIES ARE A PHYSICAL WALL OF THE BARRIER. Along this line, to break through the wall of the barrier constituted by the bodies of enemies means ‘to defeat the enemies’, as in (21a) 破敵 (*po di* ‘to defeat enemies’) and (21b) 破除重圍 (*pochu chongwei* ‘to defeat a multitude of enemies’).

(21) A causative verb meaning ‘to defeat (enemies)’

a. 破敵 po di ‘to defeat enemies’

b. 破除重圍 pochu chongwei ‘to defeat a multitude of enemies’

3.2.8 A Causative Verb Meaning ‘to Expend (Money)’

The verb PO can collocate with money, as in (22a) 費 (*fei* ‘money’) and (22b) 產 (*chan* ‘property’). Money is considered to be a concrete object—the metaphor MONEY IS A PHYSICAL ENTITY. The core meaning of the verb PO is ‘to damage an intact physical entity/substance’. When one’s money is expended, the sum of the money decreases and thus does not remain intact anymore. The collocation of PO and money derives the meaning of expending, as in (22a). As a concrete entity is broken into ‘pieces’, the degree of affectedness is tremendously great, and one’s property, as imaged as the object, goes to pieces. Hence, the expression 破產 (*po chan*) derives the meaning ‘to go bankrupt’ through metaphoric extension.

(22) A causative verb meaning ‘to expend (money)’

a. 破費 po fei ‘to expend one’s money’

b. 破產 po chan ‘to go bankrupt’

3.2.9 A Causative Verb Meaning ‘to Uncover (a Fact)’

As exemplified in (23), a criminal case can be also viewed as a concrete container—the metaphor A CRIMINAL CASE IS A PHYSICAL CONTAINER. Just as a cup contains certain liquid, the criminal container contains unknown facts. As one breaks a cup, the liquid contained in the cup will flow out. In the same vein, one has to break the criminal container so as to disclose things or facts that are covered. The derived meaning of PO ‘to uncover’ obtains accordingly.

(23) A causative verb meaning ‘to uncover (a fact)’
破案 po an ‘to solve a criminal case’

3.2.10 A Causative Verb Meaning ‘to End a Situation’

Human beings impose an artificial bound on a situation and view a situation as a concrete container—the metaphor A SITUATION IS A PHYSICAL CONTAINER. As a concrete container is broken, the state of the liquid in the container is changed, and the original state does not exist anymore. When the verb PO collocates with a state of affairs, that situation is hence put to an end. As in (24a) 破涕為笑 (*po ti wei xiao* ‘to turn tears into smiles’), it signifies that one stops crying, the prior state, and turns into a smile, a new state. Along the thread of thought, the extended meaning of PO ‘to end a situation’ holds true of (36b) 打破僵局 (*dapo jiangju* ‘to break the ice’) and (24c) 打破沉默 (*dapo chenmo* ‘to break the silence’) as well.

(24) A causative verb meaning ‘to end a situation’
a. 破涕為笑 po ti wei xiao ‘to turn tears into smiles’
b. 打破僵局 dapo jiangju ‘to break the ice’
c. 打破沉默 dapo chenmo ‘to break the silence’

3.2.11 Attributive and Predicative Adjectives

In addition to functioning as a verb, the lexical item PO can also serve as an adjective, as illustrated below:

(25) An adjective meaning ‘broken, ragged’
a. 破舊的 pojiu de ‘tattered, ragged’
b. 破衣服 po yifu ‘ragged clothes’

(26) An adjective meaning ‘worthless’
破玩意兒 po wanyier ‘worthless stuff’

(27) An adjective meaning ‘lousy’
他的中文很破 Ta de zhongwen hen po. ‘His Chinese is very lousy.’

It is speculated that the adjectival meanings of PO may evolve from the core semantics of the verb PO, i.e. ‘to damage an intact physical entity’. Functioning as a verb, PO has three types of syntactic status, namely as a transitive verb, an intransitive verb, and a verb complement, as shown in (28).

(28) Syntactic status of the verb PO
a. 打破玻璃 dapo boli ‘to break glass’
b. 窗戶破了 Chuanghu po-le. ‘The window broke.’
c. 破窗(而入) po chuang ‘to split the window into two pieces’

As a transitive verb, PO imparts a causative meaning to the sentence, as in (28c). PO as an intransitive usually bears an inchoative meaning, as in (28b). As an RVC, it denotes the resultative state brought about by the action that the predicate describes, as in (28a). Both the inchoative meaning and the RVC delineate a state, which might lead to an attributive meaning of PO. The core meaning of PO as a verb might account for the core adjectival meaning ‘broken, damaged’, as in (25) 破舊的 (*pojiu de* ‘tattered, ragged’). As an entity is broken or worn-out to a great extent, that object will be regarded as useless and worthless. The extended meaning ‘worthless’ is thus derived, as in (26) 破玩意兒 (*po wanyier* ‘worthless stuff’). When it is used for predication, PO delineates the quality of the entity referred to by the subject, as in (27) 他的中文很破 (*Ta de zhongwen hen po.* ‘His Chinese is very lousy.’). In this case, the language proficiency is regarded as a concrete entity that can be measured and assessed—low proficiency. Figure 3 provides a possible account for the semantic evolution of PO as an adjective. From the core verbal meaning of PO, the lexical item derives the more concrete adjectival sense, ‘ragged’, and the abstract semantic imports, ‘worthless’ and ‘lousy’.

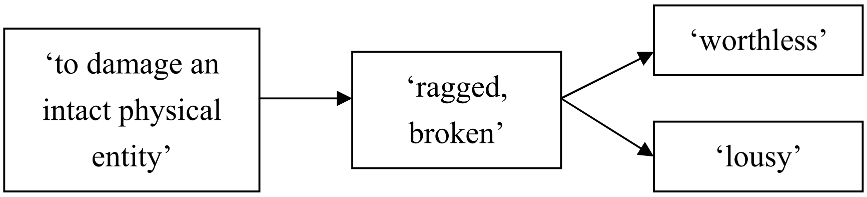


Figure 3. Semantic Extensions of PO as an Adjective

4. Conclusion

This paper has discussed the multiple semantics of the lexical item PO and identities ten verbal meanings and three adjectival meanings. As a verb, PO bears the core meaning 'to damage an intact physical entity'. The other extended meanings are interrelated to the core meaning; several abstract meanings are derived from the core directly, and some meanings evolve through secondary semantic extension. It is noted that metaphoric extension assumes an indispensable role in accounting for the evolution of the series of semantic imports. It is generalized that the semantic derivations of PO involve the following three metaphors:

X IS A PHYSICAL ENTITY.
X IS A PHYSICAL CONTAINER.
X IS A PHYSICAL BARRIER.

These metaphors are interrelated. A container is a physical entity, and so is a physical barrier. A container usually has at least five facets, and each facet can be regarded as a barrier in some sense. A concrete entity, a physical container, and a physical barrier exhibit the nature of brittleness/breakability and therefore can undergo the action of breaking. Through metaphors, the verb PO can collocate with abstract entities, including ideas, rules, records, enemies, money, unknown facts, and situations, and thus develops the extended meanings. With respect to the attributive meanings of PO, the core verbal meaning of PO gives rise to the derived core attributive meaning, which in turn lends impetus to the further evolution of adjectival meanings. Since the manifold meanings of PO, predicative or attributive, are interrelated, the polysemous nature of PO is evident. Hence, PO is not homonymous but a lexical item which bears different, but semantically related, meanings, viz. a polyseme.

References

Bybee, J. L., and W. Pagliuca. 1985. Cross-linguistic comparison and the development of grammatical meaning. *Historical Semantics, Historical Word Formation*, ed. by J. Fisiak, 59-83. Berlin: Mouton.

Chu, C. C., and T.-J. Chi. 1999. *A Cognitive-Functional Grammar of Mandarin Chinese*. Taipei: Crane.

Evens, M. W. 1988. *Relational Models of the Lexicon: Representing Knowledge in Semantic Network*. Cambridge: Cambridge University Press.

Fellbaum, C. 1998. *WordNet: A Lexical Reference System and Its Application*. Cambridge, Mass.: MIT Press.

Geerearts, D. 1993. Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*

4:223-272.

Holmes, J. 1999. The syntax and semantics of causative verbs. *UCL Working Papers in Linguistics* 11.

Hopper, P. J., and E. C. Traugott. 1993. *Grammaticalization: A Conceptual Framework*. Chicago: University of Chicago Press.

Hsiao, Y. E. 2003. Conceptualizations of GUO in Mandarin. *Language and Linguistics* 4:279-300.

Huang, C.-T. J., Y.-H. A. Li, and Y. Li. 2006. *Syntax of Chinese*, to be published by Cambridge University. Downloadable at http://www.people.fas.harvard.edu/~ctjhuang/

Huang, S. 2004. On deriving complex polysemy: Mandarin GEI in spoken and written corpora. Paper presented at the 16th American Conference on Chinese Linguistics, University of Iowa, USA.

Johnson, M. 1987. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago: University of Chicago Press.

Katz, J. J., and J. A. Fodor. 1963. The structure of a semantic theory. *Language* 39:170-210.

Kellerman, E. 1978. Giving learners a break: Native language intuitions as a source of predictions about transferability. *Working Papers on Bilingualism* 15:59-92.

Lakoff, G., and M. Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Lakoff, G. 1987. *Women, Fire, and Dangerous Things*. Chicago: University of Chicago Press.

Lakoff, G. 2002. Why cognitive linguistics requires embodies realism. *Cognitive Linguistics* 13:245-263.

Langacker, R. W. 1982. Space grammar, analyzability, and the English passive. *Language* 58:22-80.

Langacker, R. W. 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Stanford: Stanford University Press.

Langacker, R. W. 1990. *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Berlin; New York: Mouton de Gruyter.

Langacker, R. W. 1999. *Grammar and Conceptualization*. Berlin; New York: Mouton de Gruyter.

Lü, S. 1999. *Xiandai Hanyu Ba Bai Ci [Eight Hundred Words in Modern Chinese]*. Beijing: Commercial Press.

Lyons, J. 1995. *Linguistic Semantics: An Introduction*. Cambridge, England; New York: Cambridge University Press.

Rastier, F. 1999. Cognitive semantics and diachronic semantics. *Historical Semantics and Cognition*, ed. by A. Blank, and P. Koch, 109-144. Berlin: Mouton de Gruyter.

Ravin, Y., and C. Leacock. 2000. Polysemy: An overview. *Polysemy: Theoretical and*

Computational Approaches, ed. by Y. Ravin, and C. Leacock, 1-29. Oxford: Oxford University Press.

Rosch, E. 1973. On the internal structure of perceptual and semantic categories. *Cognitive Development and Acquisition of Language*, ed. by T. E. Moore. New York: Academic Press.

Rosch, E. 1977. Human categorization. *Advances in Cross-Cultural Psychology*, ed. by N. Warren. London: Academic Press.

Rosch, E. 1978. Principles of categorization. *Cognition and Categorization*, ed. by E. Rosch, and B. Lloyd, 27-48. Hillsdale, NJ: Erlbaum.

Talmy, L. 1985. Lexicalization patterns: Semantic structure in lexical forms. *Language Typology and Syntactic Descriptions*, ed. by T. Shopen, 36-149. Cambridge: Cambridge University Press.

Talmy, L. 2000a. *Toward a Cognitive Semantics*: *Concept Structuring System*. Cambridge: MIT Press.

Talmy, L. 2000b. *Toward a Cognitive Semantics*: *Typology and Process in Concept Structuring.* Cambridge: MIT Press.

Tayler, J. R. 1989. *Linguistic Categorization: Prototypes in Linguistic Theory*. Oxford: Oxford University Press.

Tayler, J. R. 2002a. *Cognitive Grammar*. Oxford; New York: Oxford University Press.

Tayler, J. R. 2002b. Polysemy's paradoxes. *Language Sciences* 25:637-655.

Tayler, J. R. 2003. *Linguistic Categorization*. 3rd edition. New York: Oxford University Press.

Tang, S. 2004. Putting BREAK to use: Prototypes and meaning extension. MA thesis, Huazhong University of Science and Technology.

Tsai, P.-T. 2006. Hsiendai hanyu *kai* yu qi yanshen yanchiu [A study on semantic extension of *kai* in Modern Chinese]. MA thesis, National Taiwan Normal University.

Wang, L. F. 2002. From a motion verb to an aspect marker: A study of *guo* in Mandarin Chinese. *Concentric: Studies in English Literature and Linguistics* 28:57-84.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Basil Blackwell and Mott.

Wu, H.-C. 2003. A case study on the grammaticalization of GUO in Mandarin Chinese—Polysemy of the motion verb with respect to semantic changes. *Language and Linguistics* 4:857-885.

One-Sample Speech Recognition of Mandarin Monosyllables using Unsupervised Learning

By

Tze Fen Li

Institute of Management, Ming Dao University, Chang-Hua, Taiwan, ROC

and

Shui-Ching Chang

Department of Information Management, The Overseas Institute of Technology, Taichung, Taiwan, ROC

Abstract

In the speech recognition, a mandarin syllable wave is compressed into a matrix of linear predict coding cepstra (LPCC), i.e., a matrix of LPCC represents a mandarin syllable. We use the Bayes decision rule on the matrix to identify a mandarin syllable. Suppose that there are K different mandarin syllables, i.e., K classes. In the pattern classification problem, it is known that the Bayes decision rule, which separates K classes, gives a minimum probability of misclassification. In this study, a set of unknown syllables is used to learn all unknown parameters (means and variances) for each class. At the same time, in each class, we need one known sample (syllable) to identify its own means and variances among K classes. Finally, the Bayes decision rule classifies the set of unknown syllables and input unknown syllables. It is an one-sample speech recognition. This classifier can adapt itself to a better decision rule by making use of new unknown input syllables while the recognition system is put in use. In the speech experiment using unsupervised learning to find the unknown parameters, the digit recognition rate is improved by 22%.

Key words and phrases: classification, dynamic processing algorithm, EM (estimate maximize) algorithm, empirical Bayes, maximum likelihood estimation, speech recognition.

Corresponding author address: Tze Fen Li, Institute of Management, Ming Dao University, 369 Wen-Hua Road, Pee-Tow, Chang-Hua (52345), Taiwan, ROC.

email address(Tze Fen Li): tfli@mdu.edu.tw

1. Introduction

A speech recognition system in general consists of feature extractor and classification of an utterance [1-5]. The function of feature extractor is to extract the important features from the speech waveform of an input speech syllable. Let x denote the measurement of the significant, characterizing features. This x will be called a feature value. The function performed by a classifier is to assign each input syllable to one of several possible syllable classes. The decision is made on the basis of feature measurements supplied by the feature extractor in a recognition system. Since the measurement x of a pattern may have a variation or noise, a classifier may classify an input syllable to a wrong class. The classification criterion is usually the minimum probability of misclassification [1].

In this study, a statistical classifier, called an empirical Bayes (EB) decision rule, is applied to solving K-class pattern problems: all parameters of the conditional density function $f(x \mid \omega)$ are unknown, where ω denotes one of K classes, and the prior probability of each class is unknown. A set of n unidentified input mandarin monosyllables is used to establish the decision rule, which is used to separate K classes. After learning the unknown parameters, the EB decision rule will make the probability of misclassification arbitrarily close to that of the Bayes rule when the number of unidentified patterns increases. The problem of learning from unidentified samples (called unsupervised learning or learning without a teacher) presents both theoretical and practical problems [6-8]. In fact, without any prior assumption, successful unsupervised learning is indeed unlikely.

In our speech recognition using unsupervised learning, a syllable is denoted by a matrix of features. Since the matrix has 8x12 feature values, we use a dynamic processing algorithm to estimate the 96 feature parameters (means and variances). Our EB classifier, after unsupervised learning of the unknown parameters, can adapt itself to a better and more accurate decision rule by making use of the unidentified input syllables after the speech system is put in use. The results of a digit speech experiment are given to show the recognition rates provided by the decision rule.

2. Empirical Bayes Decision Rules for Classification

Let X be the present observation which belongs to one of K classes $c_i, i = 1, 2, \cdots, K$. Consider the decision problem consisting of determining whether X belongs to c_i. Let $f(x \mid \omega)$ be the conditional density function of X given ω, where ω denotes one of K classes and let $\theta_i, i = 1, 2, \cdots, K$, be the prior probability of c_i with $\sum_{i=1}^{K} \theta_i = 1$. In this study, both the parameters of $f(x \mid \omega)$ and the θ_i are unknown. Let d be a decision rule. A simple loss model is used such that the loss is 1 when d makes a wrong decision and the loss is 0 when d makes a correct decision. Let $\theta = \{(\theta_1, \theta_2, \cdots, \theta_K); \theta_i > 0, \sum_{i=1}^{K} \theta_i = 1\}$ be the prior probabilities. Let $R(\theta, d)$ denote the risk function (the probability of misclassification) of d. Let $\Gamma_i, i = 1, 2, \cdots, K$, be K

regions separated by d in the domain of X, i.e., d decides c_i when $X \in \Gamma_i$. Let ξ_i denote all parameters of the conditional density function in class c_i, $i = 1, ..., K$. Then

$$R(\theta, d) = \sum_{i=1}^{K} \int_{\Gamma_i^c} \theta_i \, f(x \mid \xi_i) dx \tag{1}$$

where Γ_i^c is the complement of Γ_i . Let D be the family of all decision rules which separate K pattern classes. For θ fixed, let the minimum probability of misclassification be denoted by

$$R(\theta) = \inf_{d \in D} R(\theta, d). \tag{2}$$

A decision rule d_θ which satisfies (2) is called the Bayes decision rule with respect to the prior probability vector $\theta = (\theta_1, \theta_2, \cdots, \theta_K)$ and given by Ref.[1]

$$d_\theta(x) = c_i \quad \text{if} \quad \theta_i \, f(x \mid \xi_i) > \theta_j \, f(x \mid \xi_j) \quad for\ all\ \ j \neq \ i. \tag{3}$$

In the empirical Bayes (EB) decision problem [9], the past observations (ω_m, X_m), $m = 1, 2, \cdots, n$, and the present observation (ω, X) are i.i.d., and all X_m are drawn from the same conditional densities, i.e., $f(x_m \mid \omega_m)$ with $p(\omega_m = c_i) = \theta_i$. The EB decision problem is to establish a decision rule based on the set of past observations $\mathbf{X}_n = (X_1, X_2, \cdots, X_n)$. In a pattern recognition system with unsupervised learning, $\mathbf{X}_n$ is a set of unidentified input patterns. The decision rule can be constructed using $\mathbf{X}_n$ to select a decision rule $t_n(\mathbf{X}_n)$ which determines whether the present observation X belongs to c_i. Let $\xi = (\xi_1, ..., \xi_K)$. Then the risk of t_n, conditioned on $\mathbf{X}_n = \mathbf{x}_n$, is $R(\theta, t_n(\mathbf{x}_n)) \geq R(\theta)$ and the overall risk of t_n is

$$R_n(\theta, t_n) = \int R(\theta, t_n(\mathbf{x}_n)) \prod_{m=1}^{n} p(x_m \mid \theta, \xi) \, dx_1 \cdots dx_n \tag{4}$$

where $p(x_m \mid \theta, \xi)$ is the marginal density of X_m with respect to the prior distribution of classes, i.e., $p(x_m \mid \theta, \xi) = \sum_{i=1}^{K} \theta_i f(x_m \mid \xi_i)$. The EB approach has been recently used in many areas including classification [10,11], sequential estimation [12], reliability [13-15], multivariate analysis [16,17], linear models [18,19], nonparametric estimation [20,21] and some other estimation problems [22,23]. Let

$$S = \{(\theta, \xi); \theta = (\theta_1, ..., \theta_K), \ \xi = (\xi_1, ..., \xi_K)\} \tag{5}$$

define a parameter space of prior probabilities θ_i and parameters ξ_i representing the i-th class, $i = 1, ..., K$. Let P be a probability distribution on the parameter space S. In this study, we want to find an EB decision rule which minimizes

$$\hat{R}_n(P, t_n) = \int R_n(\theta, t_n) dP(\theta, \xi). \tag{6}$$

Similar approaches to constructing EB decision rules can be found in the recent literature [11,15,24]. From (1) and (4), (6) can be written as

$$\hat{R}_n(P,t_n) = \int \sum_{i=1}^{K} \int_{\Gamma^c_{i,n}} \big[\int f(x \mid \xi_i)\theta_i \prod_{m=1}^{n} p(x_m \mid \theta,\xi)dP(\theta,\xi)\big]dx\,dx_1\cdots dx_n \tag{7}$$

where, in the domain of X, $\Gamma_{i,n},\ i=1,2,\cdots,K$, are K regions, separated by $t_n(\mathbf{X}_n)$, i.e., $t_n(\mathbf{X}_n)$ decides c_i when $X \in \Gamma_{i,n}$ and hence they depend on the past observations $\mathbf{X}_n$. The EB decision rule which minimizes (7) can be found in Ref[24]. Since the unsupervised learning in this study is based on the following two theorems given in Ref[24], both theorems and their simple proofs are provided in this paper.

Theorem 1 [24]. The EB decision rule $\hat{t}_n$ with respect to P which minimizes the overall risk function (7) is given by

$$\hat{t}_n(\mathbf{x}_n)(x) = c_i \qquad \text{if} \qquad \begin{aligned} &\int f(x \mid \xi_i)\,\theta_i \prod_{m=1}^{n} p(x_m \mid \theta,\xi)dP(\theta,\xi) > \\ &\int f(x \mid \xi_j)\,\theta_j \prod_{m=1}^{n} p(x_m \mid \theta,\xi)dP(\theta,\xi) \end{aligned} \tag{8}$$

for all $j \neq i$, i.e., $\Gamma_{i,n}$ is defined by the definition of the inequality in (8).

Proof. To minimize the overall risk (7) is to minimize the integrand

$$\sum_{i=1}^{K} \int_{\Gamma^c_{i,n}} \big[\int f(x|\xi_i)\theta_i \prod_{m=1}^{n} p(x_m|\theta,\xi)dP(\theta,\xi)\big]dx$$

of (7) for each past observations $\mathbf{x}_n$. Let the past obervations $\mathbf{x}_n$ be fixed and let i be fixed for $i = 1,...,k$. Let

$$g_i(x) = \int f(x|\xi_i)\theta_i \prod_{m=1}^{n} p(x_m|\theta,\xi)dP(\theta,\xi).$$

Then the integrand of (7) can be written as

$$\sum_{i=1}^{K} \int_{\Gamma^c_{i,n}} g_i(x)dx = \int_{\Gamma^c_{i,n}} g_i(x)dx + \sum_{j\neq i}[\int g_j(x)dx - \int_{\Gamma_{j,n}} g_j(x)dx]$$

$$= \sum_{j\neq i} \int g_j(x)dx + \sum_{j\neq i} \int_{\Gamma_{j,n}} [g_i(x) - g_j(x)]dx \quad (\Gamma^c_{i,n} = \sum_{j\neq i} \Gamma_{j,n})$$

which is minimum since $\Gamma_{j,n} \subset \{x|g_j(x) > g_i(x)\}$ for all $j \neq i$ by the definition of $\Gamma_{j,n}$.

In applications, we let the parameters $\xi_i,\ i = 1,...,K$, be bounded by a finite numbers M_i. Let $\rho > 0$ and $\delta > 0$. Consider the subset S_1 of the parameter space S defined by

$$\begin{aligned} S_1 =&\{(n_1\rho, n_2\rho, ..., n_K\rho, n_{K+1}\delta, n_{K+2}\delta, ..., n_{2K}\delta); \quad integer\ n_i > 0,\ i = 1, ..., K, \\ &\sum_{i=1}^{K} n_i\rho = 1, |n_i\delta| \leq M_i,\ integer\ n_i,\ i = K+1, ..., 2K\} \end{aligned} \tag{9}$$

where $(n_1\rho, ..., n_K\rho)$ are prior probabilities and $(n_{K+1}\delta, ..., n_{2K}\delta)$ are the parameters of K classes. In order to simplify the conditional density of (θ, ξ), let P be a uniform distribution on S_1 so that the conditional density can later be written as a recursive formula. The boundary for class i relative to another class j as separated by (8) can be represented by the equation

$$E[f(x \mid \xi_i)\theta_i \mid \mathbf{x}_n] = E[f(x \mid \xi_j)\theta_j \mid \mathbf{x}_n] \tag{10}$$

where $E[f(x \mid \xi_i)\theta_i \mid \mathbf{x}_n]$ is the conditional expectation of $f(x \mid \xi_i)\theta_i$ given $\mathbf{X}_n = \mathbf{x}_n$ with the conditional probability function of (θ, ξ) given $\mathbf{X}_n = \mathbf{x}_n$ equal to

$$h(\theta, \xi \mid \mathbf{x}_n) = \frac{\prod_{m=1}^{n} p(x_m \mid \theta, \xi)}{\sum_{(\theta'\xi') \in S_1} \prod_{m=1}^{n} p(x_m \mid \theta', \xi')} \tag{11}$$

The actual region for class i as determined by (8) is the intersection of the regions whose borders are given by (10), relative to all other classes.

The main result in Ref[24] is that the estimates $E[\theta_i \mid \mathbf{X}_n]$ converge almost sure (a.s.) to a point arbitrarily close to the true prior probability and $E[\xi_i|\mathbf{X}_n]$ will converge to a point arbitrarily close to the true parameter in the conditional density for the i-th class. Let $\lambda = (\theta_1, ..., \theta_K, \xi_1, ..., \xi_K)$ in the parameter space S. Let λ^o be the true parameter of λ.

Lamma 1 (Kullback, 1973 [25]). Let

$$H(\lambda^o, \lambda) = \int ln\, p(x|\lambda)p(x|\lambda^o)dx.$$

Then the Kullback-Leibler information number $H(\lambda^o, \lambda^o) - H(\lambda^o, \lambda) \geq 0$ with equality if and only if $p(x|\lambda) = p(x|\lambda^o)$ for all x, i.e., $H(\lambda^o, \lambda)$ has an absolutely maximum value at $\lambda = \lambda^o$.

Let $\lambda' = (\theta', \xi') \in S_1$ such that $H(\lambda^o, \lambda') = max_{\lambda \in S_1} H(\lambda^o, \lambda)$. Since S_1 has a finite number of points, $H(\lambda^o, \lambda') - H(\lambda^o, \lambda) \geq \epsilon$ for some $\epsilon > 0$ and for all $\lambda \in S_1$. Since $H(\lambda^o, \lambda)$ is a smooth (differentiable) function of $\lambda \in S$, the maximum point λ' in S_1 is arbitrarily close to the true parameter λ^o in S if the increments δ and ρ are small.

Theorem 2 [24]. Let λ^o be the true parameter of λ. Let $\lambda = (\theta, \xi)$ in S. The conditional probability function $h(\lambda|\mathbf{x}_n)$ given $\mathbf{X}_n = \mathbf{x}_n$ in (11) has the following property: for each $\lambda \in S_1$,

$$\begin{aligned} \lim_{n \to \infty} h(\lambda \mid \mathbf{x}_n) &= 0 \qquad \text{if } \lambda \neq \lambda' \\ &= 1 \qquad \text{if } \lambda = \lambda' \end{aligned} \tag{12}$$

and hence $E[\lambda \mid \mathbf{X}_n]$ converges to λ' with probability 1.

Proof. $H(\lambda^o, \lambda)$ has an absolutely maximum value at $\lambda = \lambda'$ on S_1. Let $\lambda \in S_1$ and $\lambda \neq \lambda'$. Consider

$$\frac{1}{n} ln \frac{\prod_{m=1}^{n} p(X_m|\lambda)}{\prod_{m=1}^{n} p(X_m|\lambda')} = \frac{1}{n} \sum_{m=1}^{n} ln\, p(X_m|\lambda) - \frac{1}{n} \sum_{m=1}^{n} ln\, p(X_m|\lambda')$$

which converges almost sure to $H(\lambda^o, \lambda) - H(\lambda^o, \lambda') < -\epsilon$ by a theorem (the strong law of large numbers, Wilks, (1962) [26]), i.e., there exists a $N > 0$ such that for all $n > N$,

$$\frac{1}{n} ln \frac{\prod_{m=1}^{n} p(X_m|\lambda)}{\prod_{m=1}^{n} p(X_m|\lambda')} < -\frac{\epsilon}{2}.$$

Hence, for all $n > N$, $\frac{1}{n} ln\, h(\lambda|\mathbf{X}_n) < -\frac{\epsilon}{2}$, i.e., for all $n > N$, $ln\, h(\lambda|\mathbf{X}_n) < -n\frac{\epsilon}{2}$. This implies that $lim_{n\to\infty} ln\, h(\lambda|\mathbf{X}_n) = -\infty$ and $lim_{n\to\infty} h(\lambda|\mathbf{X}_n) = 0$ for $\lambda \neq \lambda'$ almost sure. Obviousy, $\sum_{\lambda \in S_1} h(\lambda|\mathbf{X}_n) = 1$ implies $lim_{n\to\infty} h(\lambda'|\mathbf{X}_n) = 1$ almost sure.

3. Feature Extraction

The measurements of features made on the speech waveform include energy, zero crossings. extrema count, formants, LPC cepstrum (LPCC) and the Mel frequency cepstrum coefficient (MFCC). The LPCC and MFCC are most commonly used for the features to represent a syllable. The LPC method provides a robust, reliable and accurate method for estimating the parameters that characterize the linear, time-varying system which is recently used to approximate the nonlinear, time-varying system of the speech wave. The MFCC method uses the bank of filters scaled according to the Mel scale to smooth the spectrum, performing a processing that is similar to that executed by the human ear.

3.1. Preprocessing Speech Signal

In the real world, all signals contain noise. In our speech recognition system, the speech data must contain noise. We propose two simple methods to eliminate noise. One way is to use the sample variance of a fixed number of sequential sampled points of a syllable wave to detect the real speech signal, i.e., the sampled points with small variance does not contain real speech signal. Another way is to compute the sum of the absolute values of differences of two consecutive sampled points in a fixed number of sequential speech sampled points, i.e., the speech data with small absolute value does not contain real speech signal. In our speech recognition experiments, the latter provides slightly faster and more accurate speech recognition.

3.2. Linear Predict Coding Cepstrum (LPCC)

For speech recognition, the most common features to be extracted from a speech signal are Mel-frequency cepstrum coefficient (MFCC) and linear predict coding cepstrum (LPCC). The MFCC was proved to be better than the LPCC for recognition [27], but we have shown [28] that the LPCC has a slightly higher recognition rate. Since the MFCC has to compute the DFT and inverse DFT of a speech wave, the computational complexity is much heavier than that of the LPCC. The LPC coefficients can be easily obtained by Durbin's recursive procedure [2,29,30] and their cepstra can be quickly found by another recursive equations [2,29,30]. The LPCC can provide a robust, reliable and accurate method for estimating the parameters that characterize the linear and time-varying system like speech signal [2,4,29-30]. Therefore, in this study, we use the LPCC as the feature of a mandarin syllable. The following is a brief discussion on the LPC method:

It is assumed [2-4] that the sampled speech wave $s(n)$ can be linearly predicted from the past p samples of $s(n)$. Let

$$\hat{s}(n) \ = \sum_{k=1}^{p} a_k s(n-k) \tag{13}$$

and let E be the squared difference between $s(n)$ and $\hat{s}(n)$ over N samples of $s(n)$, i.e.,

$$E \ = \sum_{n=0}^{N-1} [s(n) - \hat{s}(n)]^2. \tag{14}$$

The unknown $a_k,\ k = 1, ...p$, are called the LPC coefficients and can be solved by the least square method. The most efficient method known for obtaining the LPC coefficients is Durbin's recursive procedure [31]. Here in our speech experiment, $p = 12$, because the cepstra in the last few elements are almost zero.

3.3. Feature Extraction

Our feature extraction from LPCC is quite simple. Let $x(k) = (x(k)_1, ..., x(k)_p),\ k = 1, .., n$, be the LPCC vector for the k-th frame of a speech wave in the sequence of n vectors. Normally, if a speaker does not intentionally elongate pronunciation, a mandarin syllable has 30-70 vectors of LPCC. After 50 vectors of LPCC, the sequence does not contain significant features.

Since an utterance of a syllable is composed two parts: stable part and feature part. In the feature part, the LPCC vectors have a dramatic change between two consecutive vectors, representing the unique characteristics of syllable utterance and in the stable part, the LPCC vectors do not change much and stay about the same. Even if the same speaker utters the same syllable, the duration of the stable part of the sequence of LPCC vectors changes every time with nonlinear expansion and contraction and hence the duration of the stable parts and the duration of the whole sequence of LPCC vectors are different every time. Therefore, the duration of stable parts is contracted such that the compressed speech waveforms have about the same length of the sequence of LPCC vectors. Li [32] proposed several simple techniques to contract the stable parts of the sequence of vectors. We state one simple technique for contraction as follows:

Let $x(k) = (x(k)_1, ..., x(k)_p),\ k = 1, ..., n$, be the k-th vector of a LPCC sequence with n vectors, which represents a mandarin syllable. Let the difference of two consecutive vectors be denoted by

$$D(k) = \sum_{i=1}^{p} |x(k)_i - x(k-1)_i|, \quad k = 2, ..., n. \tag{15}$$

In order to accurately identify the syllable utterance, a compression process must first be performed to remove the stable and flat portion in the sequence of vectors. A LPCC vector $x(k)$ is removed if its difference $D(k)$ from the previous vector $x(k-1)$ is too small. Let $x'(k),\ k = 1, ..., m(< n)$, be the new sequence of LPCC vectors after deletion. We think that the first part (about 40 vectors or less) of an utterance of a mandarin syllable contains main features which can most represent the syllable and the rest of the sequence contains the "tail" sound, which has a variable length. If a speaker intentionally elongates pronunciation of a syllable, the speaker only increases the tail part of the sequence and the length of the feature part stays about the same. We partition the feature part (the first 40 vectors of the new sequence) into 6 equal segments since the feature part of LPCC vectors has a dramatic change and partition the tail part into 2 equal segments. If the whole length of the new sequence is less than 40, we neglect the tail sound and partition the new sequence into 8 equal segments. The average value of the LPCC in each segment is used as a feature value. Note that the average values of samples tend to have a normal distribution [26]. This compression produces 12x8 feature values for each mandarin syllable.

4. Stochastic Approximation

Stochastic approximation [1,2,33,34] is an iterative algorithm for random environments, which is used for parameter estimation in pattern recognition. Its convergence is guaranteed under very general circumstances. Essentially, a stochastic approximation procedure [1,2,33,34] should satisfy: (1) the successive expression of the estimate of a parameter can be written as an estimate calculated from the old n patterns and the contribution of the new $(n+1)$-st pattern and (2) the effect of the new pattern may diminish by using a decreasing sequence of coefficients. The best known of the stochastic approximation procedures are the Robbins-Monro procedure [1,33,34] and the Kiefer-Wolfowitz procedure [1,34].

For the unsupervised learning, (11) can be written in the recursive form

$$h(\lambda|\mathbf{x}_{n+1}) = \frac{p(x_{n+1}|\lambda)h(\lambda|\mathbf{x}_n)}{\sum_{\lambda' \in S_1} p(x_{n+1}|\lambda')h(\lambda'|\mathbf{x}_n)} \quad for\ n = 0, 1, 2, ... \tag{16}$$

where $h(\lambda|\mathbf{x}_n) = 1$, if $n = 0$. Equ. (16) is different from the above two types of procedures. It does not have a regression function or an obvious decreasing sequence of coefficients, but it appears to be a weighted product of the estimates calculated from the old patterns and the contribution of the new pattern. In each step of evaluation, (16) multiplies a new probability factor with the old conditional probability $h(\lambda|\mathbf{x}_n)$ based on the new pattern x_{n+1}. The convergence of (16) is guaranteed by Theorem 2.

5. A Dynamic Processing Algorithm

As in Section 3, a mandarin syllable is represented by a 12x8 matrix of feature values, which tend to be normally distributed. Let $\mathbf{x}_n = (x_1, ..., x_n)$ denote n unidentified syllables, where each x_m, $m = 1, ..., n$, denotes a 12x8 matrix of feature values, which are used to learn the means μ_{kij}, variances σ^2_{kij}, $i = 1, ..., 12$, $j = 1, ..., 8$, $k = 1, ..., K$, of normal distributions of 12x8 feature values and the prior probabilities θ_k (the probability for a syllable to appear) for K classes of syllables. For large number of classes, the stochastic approximation procedure in Section 4 is not able to estimate the means and variances, because the recursive procedure (16) needs tremendous size of computer memory. For simplicity, we let $\theta_k = 1/K$, i.e., each syllable has an equal chance to be pronounced. Let λ denote all parameters, i.e., Kx12x8 means and variances for K classes of syllables. Let λ^o be the true parameters. From Theorem 2 in Section 2, the conditional probability $h(\lambda|\mathbf{x}_n)$ has the maximum probability at $\lambda = \lambda^o$ for large n, i.e., the numerator

$$F(\mathbf{x}_n|\lambda) = \prod_{m=1}^{n} p(x_m|\lambda) \tag{17}$$

is maximum at $\lambda = \lambda^o$ for large n, where x_m, $m = 1, ..., n$, is the 12x8 matrix. Therefore, to search the true parameter λ^o by the recursive equation (16) is to find the MLE of λ.

To find the MLE of unknown parameters is a complicated multi-parameter optimization problem. First one has to evaluate the likelihood function F on a coarse grid to locate roughly the global maximum and then apply a numerical method (Gauss method, Newton-Raphson or some gradient-search iterative algorithm). Hence the direct approach tends to be computationally complex and time consuming. Here, we use a simple dynamic processing algorithm to find the MLE, which is similar to an EM [35,36] algorithm.

5.1. The Log Likelihood Function

A syllable is denoted by a matrix of feature values X_{ij}, $i = 1, ..., 12$, $j = 1, ..., 8$. For simplicity, we assume that the 12x8 random variables X_{ij} are stochastically independent (as a matter of fact, they are not independent). The marginal density function of an unidentified syllable X_m with its matrix denoted by $x_m = (x^m_{ij})$ in (17) can be written as

$$p(x_m|\lambda) = \sum_{k=1}^{K} \theta_k \prod_{ij} f(x^m_{ij}|\mu_{ijk}, \sigma_{ijk}) \tag{18}$$

where $f(x_{ij}^m|\mu_{ijk}, \sigma_{ijk})$ is the conditional normal density of the feature value X_{ij}^m in the matrix if the syllable $X_m = (X_{ij}^m)$ belongs to the k-th class. The log likelihood function can be written as

$$ln\, F(\mathbf{x}_n|\lambda) = \sum_{m=1}^{n} ln\, \{ \sum_{k=1}^{K} \theta_k \prod_{i=1}^{12} \prod_{j=1}^{8} \frac{1}{\sqrt{2\pi}\sigma_{kij}} e^{-\frac{1}{2}(\frac{x_{ij}^m - \mu_{kij}}{\sigma_{kij}})^2} \}. \tag{19}$$

5.2. A Dynamic Processing Algorithm

From the log likelihood function (19), we present a simple dynamic processing algorithm to find the MLE of unknown parameters μ_{ijk} and σ_{ijk}. Our algorithm is an EM algorithm [35,36], more and less like the Viterbi algorithm [2-4]. We state the our dynamic processing algorithm as follows:

1. In the matrix, pick up an initial value of $(\mu_{kij}, \sigma_{kij})$, $k = 1, ..., K$, for K classes.

2. For $k = 1$ and for each $i = 1, ..., 12$ and $j = 1, ..., 8$, pick up a point $(\hat{\mu}_{1ij}, \hat{\sigma}_{1ij})$ such that $ln\, F$ in (19) is maximum.

3. Continue step 2 for $k = 2, ..., K$.

4. If (19) continues increasing, go to step 2, otherwise, stop the dynamic processing and the final estimates $(\hat{\mu}_{kij}, \hat{\sigma}_{kij})$ are the MLE of $(\mu_{kij}, \sigma_{kij})$ for all K classes and are saved in a database.

5.3. Finding the Means and Variances for each Syllable by a Known Sample

For each element (i, j) in the matrix, we have found the MLE $(\hat{\mu}_{kij}, \hat{\sigma}_{kij})$ for each syllable. There are totally K matrices of MLE representing K different syllables, but we do not know which matrix of MLE belongs to the syllable c_i, $i = 1, ..., K$. We have to use one known sample from each syllable to identify its own matrix of MLE. In this paper, we simply use the distance to select a matrix of MLE among K matrices for the known sample.

5.4. Classification by the Bayes Decision Rule

After each syllable obtains its means and variances which are identified by a known sample of the syllable, the Bayes decision rule (3) with the estimated means and variances (MLE) classifies the set of all unidentified syllables. After simplification [32], the Bayes decision rule (3) can be reduced to

$$l(c_k) = \sum_{ij} ln(\hat{\sigma}_{kij}) + \frac{1}{2} \sum_{ij} (\frac{x_{ij} - \hat{\mu}_{kij}}{\hat{\sigma}_{kij}})^2 \tag{20}$$

where $\{x_{ij}\}$ denotes the matrix of LPCC of an input unknown syllable. The matrix of LPCC of an unknown syllable is compared with each known syllable c_k represented by $(\hat{\mu}_{kij}, \hat{\sigma}_{kij})$. The Bayes rule (20) selects a syllable c_k with the least value of $l(c_k)$ from K known syllables to be the input unknown syllable.

Note that new input unidentified syllables can update the estimated means and variances (MLE) which are closer to the true unknown means and variances, and hence the Bayes decision rule will become a more accurate classifier.

6. Speech Experiment on Classification of Digits

Our speech recognition is implemented in a classroom. The data of 10 mandarin digits are created by 10 different male and female students, each pronouncing 10 digits (0-9) once. The mandarin pronunciation for 1 and 7 is almost the same. It is hard to classify these two syllables.

6.1. Speech Signal Processing.

The speech signal of a mandarin monosyllable is sampled at 10k Hz. A Hamming window with a width of 25.6 ms is applied every 12.8 ms for our study. A Hamming window with 256 points is used to select the data points to be analyzed. In this study, the 12x8 unknown parameters of features representing a digit are estimated by unsupervised learning. After learning the parameters, there are 10 12x8 matrices of estimates representing 10 digits. For each digit, use one known sample to identify a 12x8 matrix of estimates to represent the digit.

In our speech experiments, we use this database to produce the LPCC and obtain a 12x8 matrix of feature values for each syllable. There are totally 100 matrices of feature values.

6.2. To Learn Means and Variances using Unsupervised Learning

The simple dynamic processing algorithm in Section 5 produces 10 matrices of MLE (estimated means and variances). After a known sample of each digit (0,1,...,9) picks up its own matrix of MLE, the 10 matrices are ranked in order from 0 to 9 as fellows: $(\hat{\mu}_{kij}, \hat{\sigma}_{kij})$, $i = 1, ..., 12$, $j = 1, ..., 8$, for $k = 0, ..., 9$. One of 10 students pronounces 10 digits which are considered as 10 known samples (each for one digit) and the other 9 students pronounce 10 digits (90 samples), which are considered as unknown samples. The total 100 samples (10 known samples and 90 unknown samples) are used for finding the matrices of MLE of the means and variances for 10 digits. This experiment is implemented five times, each time for one of five different students whose 10 digit pronunciations are considered as known samples. Note that the only training samples are the only one sample for each digit pronounced by a student and note that the testing samples are the mixed 90 unknown samples of 10 digits pronounced by the other 9 students. Actually, the experiment is a speaker-independent speech recognition. The 10 training samples and the 90 testing samples (90 mixed unknown samples also used for unsupervised learning of parameters) are totally separated.

6.3. Speech Classification on the Mixed Samples

In this study, two different classifiers are used to classify 90 unknown mixed digital samples since 10 digital samples pronounced by one student are already known.

(a). Bayes Decision Rule.

The estimated means and variances of each digit obtained in (6.2) are placed into the Bayes decision rule (20). The Bayes decision rule classifies 90 mixed samples (except 10 known samples for 10 digits (0-9)). The recognition rates are listed in Table 1.

(b). Distance Measure from 10 Known Samples

The known sample of a digit (0-9) identifies 90 other mixed unknown samples using distance measure from the known sample, i.e., to classify an unknown sample, we select a known sample from 10 known samples which is the closest to the unknown sample to be the unknown sample. Its recognition rates are also listed in Table 1. From Table 1, the Bayes decision rule using unsupervised learning gives the higher recognition rate 79%, 22% more than the rate 57% given by the distance measure using one known sample.

Table 1. Recognition rates for 10 digits given by the Bayes decision rule with unsupervised learning to classify 90 unknown samples as compared with the distance measure without unsupervised learning.

	student 1	student 2	student 3	student 4	student 5	average
Bayes rule with	72	70	69	68	76	71.0
unsupervised learning	.80	.78	.77	.76	.84	.79
distance measure	55	51	39	54	58	54.4
	.61	.57	.43	.60	.64	.57

Discussions and Conclusion

This paper is the first attempt to use an unsupervised learning for speech recognition. Actually, this paper presents an one-sample speech recognition. An unsupervised learning needs a trmendous amount of unknown samples to learn the unkown parameters of syllables. From Theorem 2, the estimates using unsupervised learning will converge the true parameters and hence, our classifier can adapt itself to a better decision rule by making the use of unknown input syllables for unsupervised learning and will become more and more accurate after the system is put in use. Theoretically, from Theorem 2, our one-sample speech recognition rate will approach to the rate given by supervised learning classifiers if a syllable does not have too many unknown parameters. In our experiments, we only have 9 samples for each syllable (a total of 90 unknown samples after 90 samples are mixed) for unsupervised learning of 96 parameters for each syllable and hence we only obtain 79% accuracy, 22% more than the rate without unsupervised learning.

Acknowledgments

The authors are grateful to the editor and the referees for their valuable suggestions to improve this paper.

References

[1]. K. Fukunaga, Introduction to Statistical Pattern Recognition, New York: Academic Press, 1990.

[2]. Sadaoki Furui, Digital Speech Processing, Synthesis and Recognition, Marcel Dekker, Inc., New York and Basel, 1989.

[3]. L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, PTR, Englewood Cliffs, New Jersey, 1993.

[4]. X. D. Huang, A. Acero, and H. W. Hon, Spoken Language Processing - A guide to theory, algorithm, and system development, Prentice Hall, PTR, Upper Saddle River, New Jersey, USA, 2001.

[5]. L. Deroye, L. Gyorfi, and G. Lugosi, A Probabilistic Theory of Pattern Recognition, Elsevier, New York, 1996.

[6]. R. L. Kasyap, C. C. Blayton, and K. S. Fu, Stochastic Approximation in Adaptation, Learning and Pattern Recognition Systems: Theory and Applications, J. M. Mendel and K. S. Fu. Eds., New York, Academic, 1970.

[7]. T. Y. Young and T. W. Calvert, Classification, Estimation and Pattern Recognition, New York: Elsevier, 1974.

[8]. A. G. Barto and P. Anandan, Pattern recognizing stochastic learning automata, IEEE Trans. Syst., Man, Cybern., Vol. SMC-15(May 1985) 360-375.

[9]. H. Robbins, An empirical Bayes approach to statistics, Proc. Third Berkeley Symp. Math. Statist. Prob., Vol. 1, University of California Press, (1956), 157-163.

[10]. Y. Lin, A note on margin-based loss function in classification, Statist. and Pro. Letters, 68(1)(2004), 73-81.

[11]. T.F. Li and S.C. Chang, Classification on defective items using unidentified samples, Pattern Recognition, 38(2005), 51-58.

[12]. R. J. Karunamuni, Empirical Bayes sequential estimation of the means, Sequential Anal., 11(1)(1992),

37-53.

[13]. A. Sarhan, Non-parametric empirical Bayes procedure, Reliability Engineering and System, 80(2)(2003), 115-122.

[14]. A. Sarhan, Empirical Bayes estimation in exponential reliability model, Applied Math. and Computation, 135(2)(2003), 319-332.

[15]. T. F. Li, Bayes empirical Bayes approach to estimation of the failure rate in exponential distribution, Commu.-Stat. Meth., 31(9)(2002), 1457-1465.

[16]. M. Ghosh, Empirical Bayes minimax estimators of matrix normal means, J. Multivariate Anal., 38(2)(1991), 306-318.

[17]. S. D. Oman, Minimax hierarchical empirical Bayes estimation in multivariate regression, J. Multivariate Anal., 80(2)(2002), 285-301.

[18]. R. Basu, J. K. Ghosh, and R. Mukerjee, Empirical Bayes prediction intervals in a normal regression model: higher order asymptotics, Statist. and Pro. Letters, 63(2)(2003), 197-203.

[19]. L. Wei and J. Chen, Empirical Bayes estimation and its superiority for two-way classification model, Statist. and Prob. Letters, 63(2)(2003), 165-175.

[20]. M. Pensky, Nonparametric empirical Bayes estimation of the matrix parameter of the Wishart distribution, J. Multivariate Anal., 69(2)(1999), 242-260.

[21]. M. Pensky, A general approach to nonparametric empirical Bayes estimation, Statistics, 29(1)(1997), 61-80.

[22]. S. Majumder, D. Gilliland, and J. Hannan, Bounds for robust maximum likelihood and posterior consistency in compound mixture state experiments, Statist. and Prob. Letters, 41(3)(1999), 215-227.

[23]. Y. Ma, Empirical Bayes estimation for truncation parameters, J. Statistical Planning and Inference, 84(1)(2000), 111-120.

[24]. T. F. Li and T. C. Yen, A Bayes Empirical Bayes decision rule for classification, Communications in Statistics-Theory and Methods, 34(2005), 1137-1149.

[25]. S. Kullback, Information Theory and Statistics, Gloucester, MA: Peter Smith, 1973.

[26]. S.S. Wilks, Mathematical Statistics, New York: John Wiley and Son, 1962.

[27]. S. B. Davis and P. Mermelstein, Comparison of parametric representation for monosyllabic word recognition in continously spoken sentences, IEEE. Trans. Acoust., Speech, Signal Processing, 28(4)(1980), 357-366.

[28]. T. F. Li, A note on Mel frequency cepstra in speech recognition, Department of Applied Mathematics, Chung Hsing University, Taichung, Taiwan, (2006).

[29]. J. Makhoul and J. Wolf, Linear Prediction and the Spectral Analysis of Speech, Bolt, Baranek, and Newman, Inc., Cambridge, Mass., Rep. 2304, 1972.

[30]. J. Makhoul, Linear prediction: a tutorial review, Proc. IEEE, 63(4)(1975), 561-580.

[31]. J. Tierney, A study of LPC analysis of speech in additive noise, IEEE Trans. Acoust. Speech Signal Process., 28(4)(1980), 389-397.

[32]. T. F. Li, Speech recognition of mandarin monosyllables, Pattern Recognition, 36(2003), 2712-2721.

[33]. H. Robbins and S. Monro, A stochastic approximation method, Ann. Math. Statist., 22(1951), 400-407.

[34]. A. Abert and L. Gardner, Stochastic Approximation and Nonlinear Regression, Cambridge, MA, M.I.T., 1967.

[35]. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Ann. R. Stat. Soc. 39(1977), 1-35.

[36]. C. F. J. Wu, On the convergence properties of the EM algorithm, Ann. Stat., 11(1983), 95-103.

Examining the Lexical Effect on Categorical Perception of Stops in Taiwan Southern Min

Yunglin Tai
Institute of Linguistics
National Chung Cheng University
yunglin2@gmail.com

Abstract

The goal of this study is to examine if there is a word superiority effect on perception of three-way contrast of stops in Taiwan Southern Min (TSM). Based on Ganong's (1980) findings that English participants showed a significant lexical effect in phonetic perception, I hypothesize that there exists a difference in perception between real words and nonwords in TSM. The prediction is that the categorical boundary shifts as lexical status plays a role in perception of stops. Experiment 1 was conducted as a neutral set of perception of TSM bilabial stops *b-p-p^h*. In experiment 2, cases of "nonword-word-nonword" set along the *b-p-p^h* continuum were conducted. Results showed that real words corresponded to a wider range of VOT in the continuum compared to the neutral pattern. The categorical boundaries (both between /b/ and /p/ and between /p/ and /p^h/) were found to shift away from the real word sides towards the nonword sides. The lexical effect may be explained by parallel processing in which a higher level of processing (lexical level) interacts with a lower level (phonemic level) in speech perception.

Keywords: perception, categorical boundary, VOT, lexical effect, processing

1. Introduction

Categorical perception, which represents that sounds within a phonemic category are perceived as indistinguishable regardless of the correct identification of each sound, supports the modular view that speech is perceptually special. In literature, consonants were found to be perceived categorically, but not gradual; that is, although listeners were able to distinguish the consonant stimuli between different phonetic categories with ease, mostly they failed to distinguish the stimuli within categories.

Sometimes the perceptual difference between stops attributes only to the initial voicing feature. For instance, the difference between [b], [p] and [p^h] is due to voice onset time (VOT): the interval between the stop release and the beginning of vocal cord vibrations. A positive VOT value means such a lag exists (e.g. the aspirated [p^h]); a VOT value of zero represents no delay in voicing; a negative VOT value refers to the phenomenon where the vocal fold vibrations begin before the articulatory release of the stop (the prevoiced [b]). As VOT varies gradually, the identification changes abruptly from one stop to another. This categorical perception drives how the modular view regards speech—as a modular system.

However, this raises an interesting question—is it modular in such a way that consonants are always perceived without the affection by other processing information?

There are some parallel models proposed for linguistic processing. One of those is the Trace model of speech perception (McClelland & Elman 1986). The model assumes that different levels of processing—features, phonemes, and words—are activated simultaneously during speech perception, which contradicts with the modularity view that phonemic processing in unaffected by higher levels of processing.

Ganong (1980) investigated whether auditory word perception affected phonetic categorization. He constructed acoustic continua varying in voice onset time, each with an end being a real word and the other end a nonword (e.g., *dash—tash vs. dask—task*). The results showed a categorical boundary shift. Thus he argued that the lexical effect must arise at a processing stage sensitive to both lexical and auditory information.

Based on previous studies on categorical perception and Ganong's (1980) findings on the lexical effect that English participants showed a significant lexical effect in phonetic perception, it is hypothesized that there exists a lexical effect—a difference in perception between real words and nonwords. The present study examines a language of a three-way contrast in initial stops, Taiwan Southern Min, to see if there is a word superiority effect on speech perception. The prediction is that real words will correspond to a wider range of VOT in the continuum either comparing to the neutral pattern or to the nonword situation. The logic behind it is that listeners tend to perceive an ambiguous sound as a real word rather than a nonword. If the prediction is true, it should be found that the phoneme boundaries (both between /b/ and /p/ and between /p/ and /p^h/) shift away from the real word sides towards the nonword sides.

The current research questions are:

(1) Where is the categorical boundary between *b* and *p* and that between *p* and *p^h* in Taiwan Southern Min (TSM)?

(2) Does lexical status (a real word or nonword) of a sound sequence affect the perception of TSM stops? If it does, how does the categorical boundary shift?

(3) Categorical perception may be better explained by the modular theory or parallel processing models?

The pinyin system (including consonants, vowels and tones) adopted in this paper follows "台灣閩南語音標系統" released by Ministry of Education in 1998. Example lexemes were obtained through *臺灣閩南語辭典* by 董 et al. (2001).

Table 1. Tones used in the current study

Tone category	Tone value	Example
陰上	53	飽 /pa53/ 'full'
陰去	11	富 /pu11/ 'rich'
陽去	33	密 /ba33/ 'closely'

2. Experiment 1—identification of a neutral set

Experiment 1 is an identification task designed to test the categorical boundaries of bilabial stops *b, p,* and p^h in TSM. Serving as a neutral set, stimuli were not informed to participants beforehand about their lexical or non-lexical status in TSM. Rather, they were instructed to pay attention to any change of the consonant category only.

2.1 Methods

2.1.1 Participants

Ten native speakers of TSM participated. All of them were graduate students from National Chung Cheng University, with self-reported fluency in TSM.

2.1.2 Materials

21 stimuli from the bilabial neutral continuum pair *ba—pa—*p^h*a*, with VOTs of -100, -90, -80, -70, -60, -50, -40, -30, -20, -10, 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 msec manipulated in Praat (Boersma & Weenink 2007). Voice onset time (VOT) in msec of each syllable-initial stop was measured from the release of the stop closure to the beginning of the oscillating line which demonstrates voicing in the following vowel.

These stimuli are adjusted from a TSM male speaker's natural speech. Among the 21 stimuli, all attributes of sounds are identical except the VOT in syllable initials.

The manipulation of prevoicing or aspiration is through deletion or addition from a range of repetitive circular noise lines. One thing to be careful is that the beginning and ending point in selection must match each other in terms of their amplitude position. Otherwise, low pitch noise or unnatural burst would be created during manipulation.

Here is what some examples of stimuli look like:

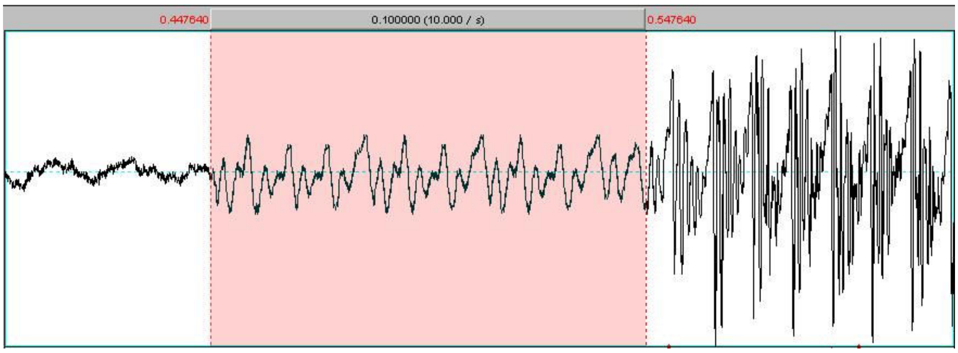


Figure 1. A Stimulus with -100ms in VOT (prevoicing)

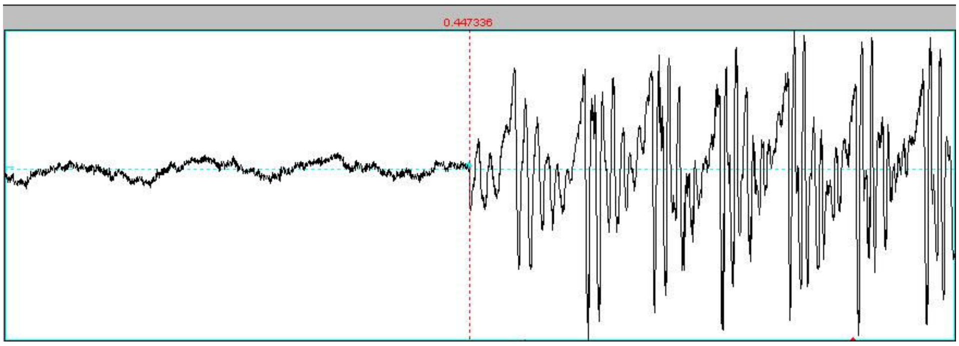


Figure 2. A Stimulus with 0ms in VOT

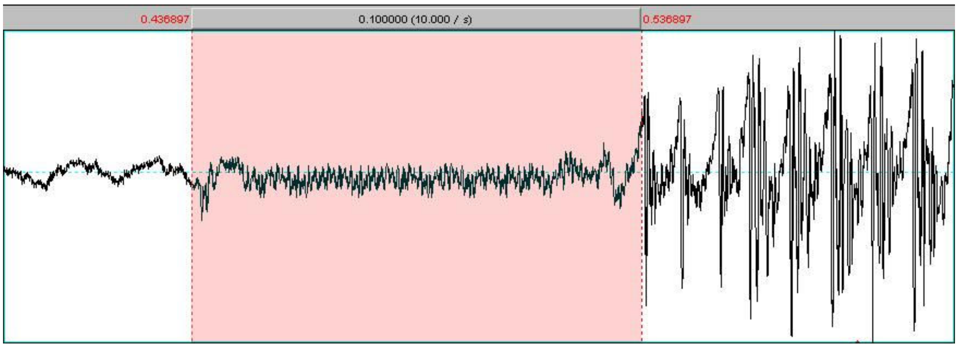


Figure 3. A Stimulus with 100ms in VOT (aspiration)

2.1.3 Procedure

Stimuli were saved as a WAV file and played in a notebook computer. Participants were given a piece of paper where there are three columns (*ba, pa,* and *pha*) along the 21 stimuli. They are instructed to judge whether the CV syllable they are listening begins with *b*, *p* or p^h and then fill in the corresponding column with a check.

2.2 Results and discussion

Stimuli and the corresponding VOT values are shown below:

Table 2. Stimuli & Corresponding VOT values

Stimulus number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
VOT (ms)	-100	-90	-80	-70	-60	-50	-40	-30	-20	-10	0	10	20	30	40	50	60	70	80	90	100

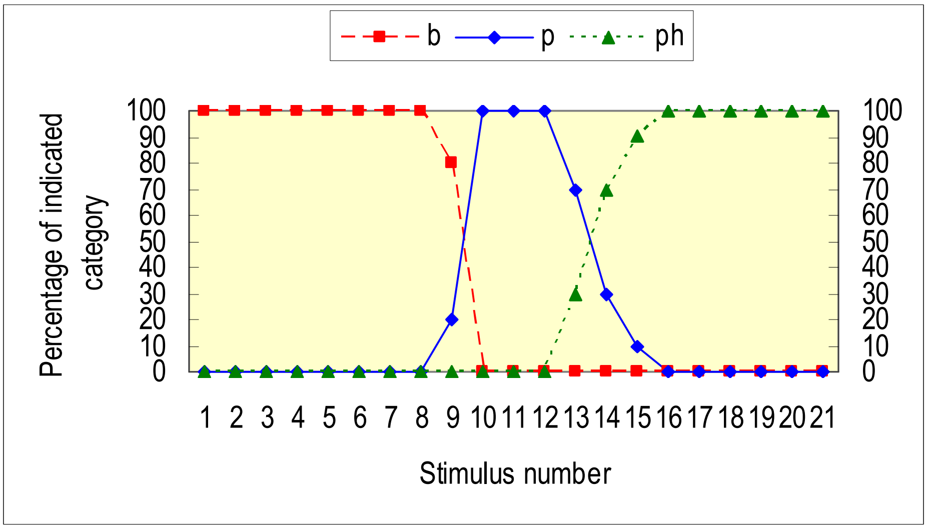


Figure 4. Perception of the Neutral Set

Inter-participant variation (standard deviation) is 0.422 for *b*, 1.16 for *p* and 0.994 for *ph*.

As for the categorical boundaries, their locations are to refer to the intersection points of each line. Thus, according to the graph above, the categorical boundary is -15ms between *b* and *p* and 25ms between *p* and *ph*.

3. Experiment 2—identification of nonword-word-nonword sets

Experiment 2 contains two identification tasks conducted to examine if there is a difference comparing with the result in Experiment 1, by adding a linguistic variable: the lexical status of a sound sequence along a continuum. In this experiment, combination of "nonword-word-nonword"[1] is chosen to examine this effect.

3.1 Methods

3.1.1 Participants

Ten native speakers of TSM (different from the participants in Experiment 1) participated. All of them were graduate students from National Chung Cheng University, with self-report fluency in TSM.

3.1.2 Materials

There are two groups^2 of stimuli used. The first group is 21 stimuli from the bilabial pair "*ba53* - *pa53* 飽'full' - *pha53*"; the second group is 21 stimuli from the bilabial pair "*bu11* - *pu11* 富 'rich' - *phu11*". VOTs are -100, -90, -80, -70, -60, -50, -40, -30, -20, -10, 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 msec along the continuum respectively.

3.1.3 Procedure

Materials were in WAV file format and played in a notebook computer. All the participants do both groups of stimuli. This experiment was conducted with a counterbalance design in order. Half of the participants did the first group of stimuli first, and the other half did the second group first. Participants were given a piece of paper where there were three columns of *ba53, pa53* 飽 and *pha53* along the 21 stimuli and another piece of paper where there are three columns *bu11, pu11* 富 and *phu11* along the 21 stimuli (Or the reversed order). They were instructed to judge whether the CV syllable they were listening belonged to *ba53, pa53* 飽 or *pha53* and to *bu11, pu11* 富 or *phu11*, in order to fill in the corresponding column with a check.

3.2 Results and discussion

3.2.1 The "*ba53* - *pa53* 飽'full' - *pha53*" set

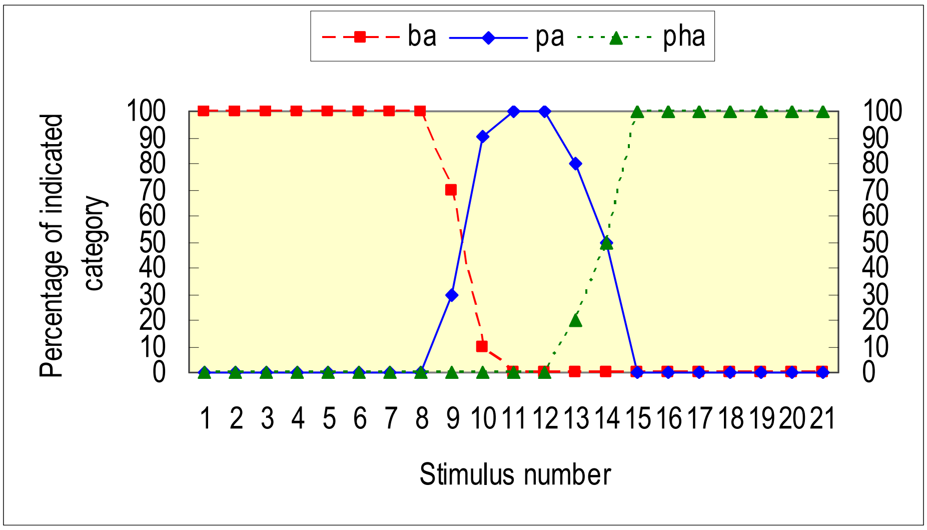


Figure 5. Perception of the "*ba53* - *pa53* 飽 'full' - *pha53*" set

Inter-participant variation (standard deviation) is 0.632 for *ba*, 1.08 for *pa* and 0.823 for *pha*.

As for the categorical boundaries, their locations are to refer to the intersection points of each line. According to each of the solution of two formulas in mathematics, the number of intersection points in x axis is 28/3 and 14 respectively. Thus, after conversion to the corresponding VOTs, the categorical boundary is -16.667ms between *ba* and *pa* and 30ms between *pa* and *pha*.

We can see that comparing to those of the neutral set (-15ms and 25ms), the two categorical boundaries of this nonword-word-nonword set shift towards the nonword sides. Hence, the real word *pa53* 飽'full' corresponds to a wider range of VOT in the continuum comparing to the neutral pattern.

3.2.2 The "*bu11 - pu11* 富 'rich' - *phu11*" set

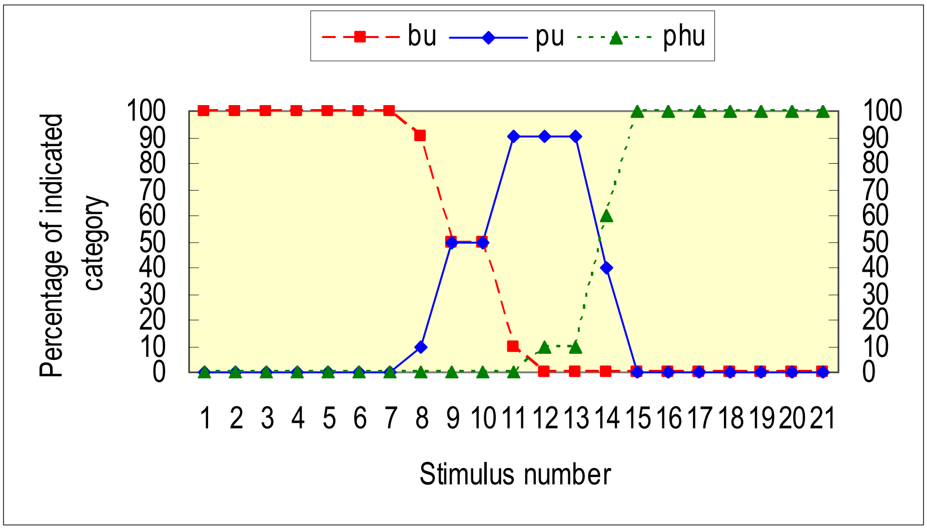


Figure 5. Perception of the "*bu11 - pu11* 富 'rich' - *phu11*" set

Inter-participant variation (standard deviation) is 1.333 for *bu*, 1.033 for *pu* and 0.919 for *phu*. Note that the variation of *bu* is higher than both *b* in the neutral set and *ba*. One of the possible factors may be due to the more marked CV co-occurrence patterns[3] (Davis & MacNeilage 1995) of *bu* than *ba*, which further affects the salience in speech perception.

As for the categorical boundaries, the number of the intersection points of each line in x axis is 9 to 10 (the common part is actually a line segment) and 13.8 respectively. Thus, after conversion to the corresponding VOTs, the categorical boundary is -15ms between *bu* and *pu* and 28ms between *pu* and *phu*.

Comparing to those of the neutral set (-15ms and 25ms), the categorical boundary between *pu* and *phu* in "*bu11 - pu11* 富 'rich' - *phu11*" set shift towards the nonword sides. But it does not show the leftward spreading clearly in this case. However, the real word *pu11*

富 'rich' also corresponds to a wider range of VOT in the continuum compared to the neutral pattern.

4. General discussion & Conclusion

To compare the three sets done in Experiment 1 and Experiment 2, see the following graphs:

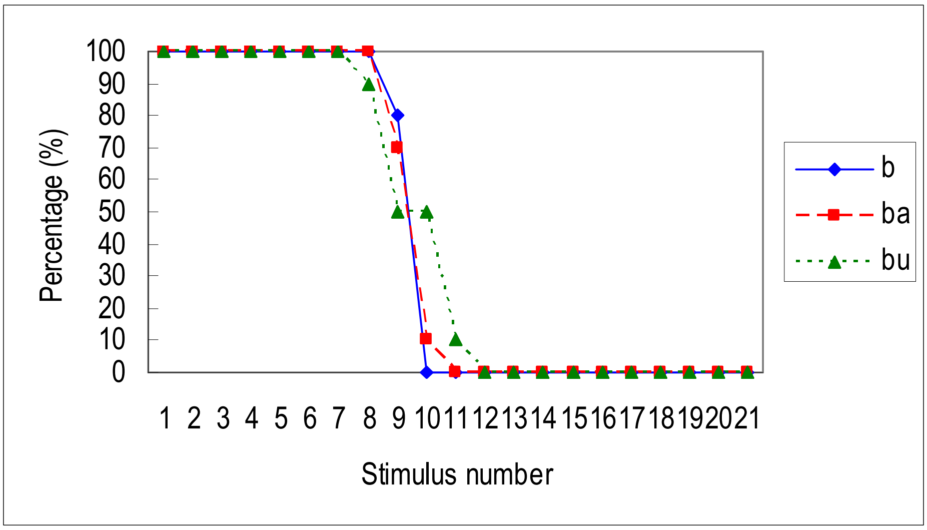


Figure 6. Comparison of the neutral /b/ with *b* in /ba/ and /bu/

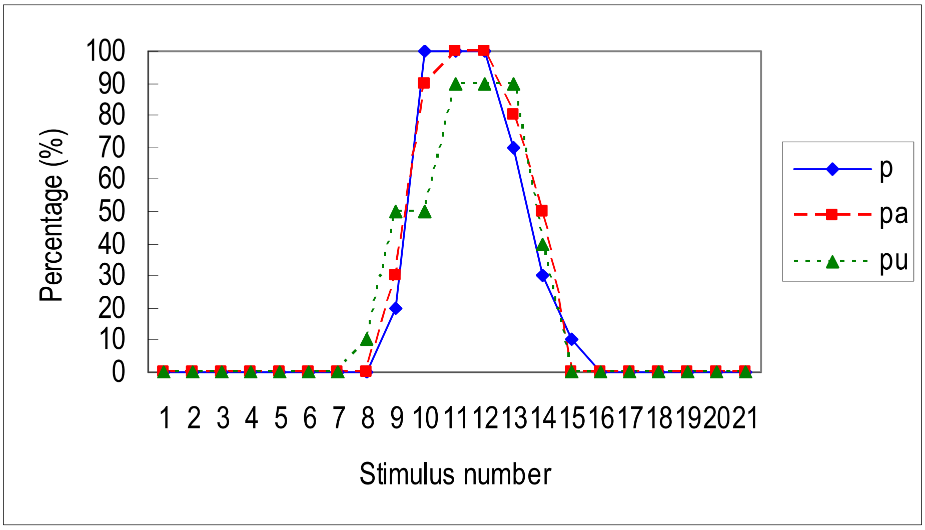


Figure 7. Comparison of the neutral /p/ with *p* in /pa/ and /pu/

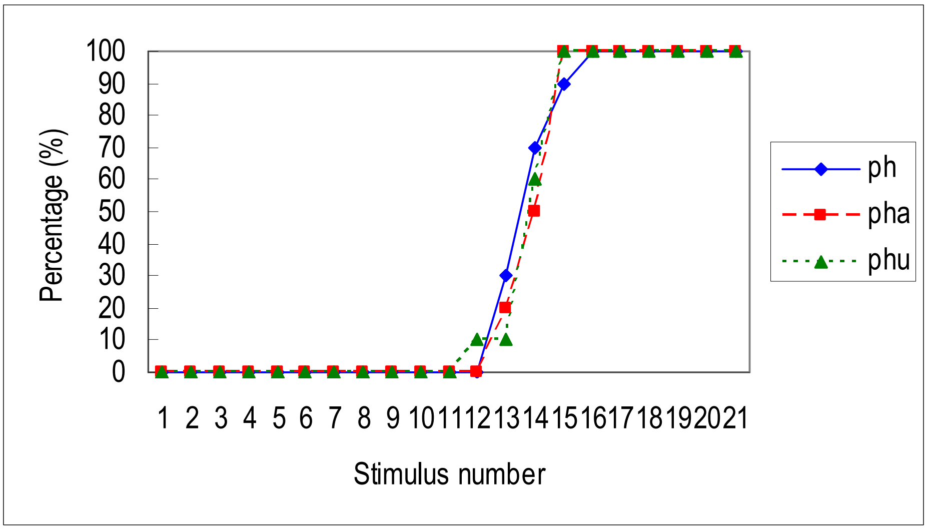


Figure 8. Comparison of the neutral /p^h/ with *p^h* in /p^{h}a/ and /p^{h}u/

If you see the broken line representing *ba* in the first graph, the location is a little bit left to the solid line representing the neutral pattern. For the *pa* line in the second graph, the location is outside the neutral solid line. In terms of the *pha* line in the third graph, it locates obviously at the right side of the neutral line. All of these consistent patterns suggest that there exists a lexical effect from the word *pa53* 飽'full'—its lexical status affects the perception of consonants, making categorical boundaries shift towards the nonword sides.

As for the dotted line representing *bu* in the first graph and *pu* in the second graph, the categorical boundary does not shift clearly; that is, the lexical effect is not found in that the categorical boundary does not spread leftward. However, *phu* in the third graph shows a similar pattern with *pha*, indicating a shift of the categorical boundary between *p* and *ph*. These suggest there is still a lexical effect found in the real word *pu11* 富 'rich'.

However, since the results of the two sets in Experiment 2 show some difference themselves, there may be other factors other than only the lexical status to affect perception of consonants. Newman et. al. (1997) examined the lexical neighborhood effect and found that it affected word recognition in much the same way as lexical status. Difference in CV combinations, tone or neighborhood density of a possible word may be taken into account.

All in all, results in the present study show the pattern as what is predicted—real words correspond to a wider range of VOT in the continuum comparing to the neutral pattern. The categorical boundaries (both between /b/ and /p/ and between /p/ and /p^h/) are found to shift away from the real word sides towards the nonword sides. Besides, the phenomenon that the lexical status affects the phonetic categorization suggests that there exists an interactive processing between different levels, consistent with the Trace model. Hence the lexical effect may be explained by parallel processing in which a higher level of processing (lexical level) interacts with a lower level (phonemic level) in speech perception.

Other combination of lexical status (nonword-word-word and word-word-nonword) must be examined in future study to determine the lexical effect of mono-syllables in TSM. Besides, disyllabic word pairs in TSM are possible materials to examine the lexical effect in the future since disyllabic words in spoken TSM are more frequent compared to monosyllabic words.

References

[1] P. Boersma and D. Weenink, *Praat: doing phonetics by computer (Version 5.0.02)* [Computer program]. Retrieved December 27, 2007, from http://www.praat.org/.

[2] B. L. Davis and P. F. MacNeilage, "The articulatory basis of babbling" in *Journal of Speech and Hearing Research*, Vol. 38, 1995, pp. 1199-1211.

[3] W. F. Ganong, "Phonetic categorization in auditory word perception" in *Journal of Experimental Psychology: Human Perception Performance*, Vol. 6, 1980, pp. 110–125.

[4] J. L. McClelland and J. L. Elman, "The TRACE model of speech perception" in *Cognitive Psychology*, Vol. 18, 1986, pp. 1-86.

[5] R. S. Newman, J. R. Sawusch, and P. A. Luce, "Lexical neighborhood effects in phonetic processing" in *Journal of Experimental Psychology: Human Perception Performance* , Vol. 23, 1997, pp. 873–889.

[6] 董忠司、城淑賢、張屏生, *臺灣閩南語辭典*. 五南, 2001.

[1] The combination of "word-nonword-word" is not chosen because there is only one case in TSM: "*bu53-pu53-phu53*".

[2] Another group "*bu55-pu55-phu55*" is not used as materials for *pu55* serves as an onomatopoeic word only.

[3] Davis & MacNeilage (1995) proposed a universal unmarked CV co-occurrence patterns to be [bilabial+central], [alveolar+front] and [velar+back].

Automatic labeling of troponymy for Chinese verbs

羅巧珊　Chiao-Shan Lo*+　陳怡蓉　Yi-Rung Chen+
jcland0407@hotmail.com　yrsmile117@gmail.com

林芝佑　Chih-Yu Lin+　謝舒凱　Shu-Kai Hsieh*+
jyyow@msn.com　shukai@gmail.com

*Lab of Linguistic Ontology, Language Processing and e-Humanities,
+Graduate School of English/Linguistics,
National Taiwan Normal University

Abstract

以同義詞集與詞彙語意關係架構而成的詞彙知識庫，如英語詞網 (Wordnet)、歐語詞網 (EuroWordnet)等，已有充分的研究，詞網的建構也已相當完善。基於相同的目的，中研院語言所亦已建立大規模之中文詞彙網路 (Chinese Wordnet,CWN)，旨在提供完整的中文辭彙之詞義區分。然而，在目前之中文詞彙網路系統中，由於目前主要是採用人爲判定來標記同義詞集之間的語意關係，因此這些標記之數量尙未達成可行應用之一定規模。因此，本篇文章特別針對動詞之間的上下位詞彙語意關係 (Troponymy)，提出一種自動標記的方法。我們希望藉由句法上特定的句型 (lexical syntactic pattern)，建立一個能夠自動抽取出動詞上下位的系統。透過詞義意判定原則的評估，結果顯示，此系統自動抽取出的動詞上位詞，正確率將近百分之七十。本研究盼能將本方法應用於正在發展中的中文詞網自動語意關係標記，以及知識本體之自動建構，進而能有效率的建構完善的中文詞彙知識資源。

關鍵詞: 中文詞彙網路、語義關係自動標記、動詞詞彙語義

Abstract

Synset and semantic relation based lexical knowledge base such as wordnet, have been well-studied and constructed in English and other European languages (EuroWordnet). The Chinese wordnet (CWN) has been launched by Academia Sinica basing on the similar paradigm. The synset that each word sense locates in CWN are manually labeled, however, the lexical semantic relations among synsets are not fully constructed yet. In this present paper, we try to propose a lexical pattern-based algorithm which can automatically discover the semantic relations among verbs, especially the troponymy relation. There are many ways that the structure of a language can indicate the meaning of lexical items. For Chinese verbs, we identify two sets of lexical syntactic patterns denoting the concept of hypernymy-troponymy relation. We describe a method for discovering these syntactic patterns and automatically extracting the target verbs and their corresponding hypernyms. Our system achieves satisfactory results and we beleive it will shed light on the task of automatic acquisition of Chinese lexical semantic relations and ontology learning as well.

Key word: troponymy, automatic labeling, lexical syntactic pattern

1 Introduction

In recent years, there has been an increasing focus on the construction of lexical knowledge resources in the field of Natural Language Processing, such as Thesaurus, Wordnets, Mindnet, Hownet, VerbNet, etc. Among these resources, Princeton WordNet[1], started as an implementation of a psycholinguistic model of the mental lexicon, has sparked off most interest both in theoretical and applicational sides. WordNet's growing popularity has prompted the modeling and construction of wordnet in other languages and various domains as well. However, creating a lexical semantic knowledge resource like WordNet is a time-consuming and labor-intensive task. Languages other than English and some European languages are facing with the lack of long-term linguistic supports, let alone those languages without balanced corpus available. This has motivated researches into automatic methods paralleled with manual verification, in order to ease the work.

In Chinese, constructing a semantic relation-based wordnet is comparatively difficult owing to the fuzzy definition and classification among words, morphemes, and characters. The Chinese Wordnet (CWN)[2], created by Academia Sinica, aims to provide complete senses for each word based on the theory of lexical semantics and ontology. However, the synsets of each word in CWN are manually labeled and the semantic relations among synsets are not fully constructed. In this present paper, we try to propose an algorithm which can automatically label the semantic relations among verbs, especially focused on the hypernymy-troponymy relation. According to Fellbaum [2], lexical entries in a dictionary can reflect the relatedness of words and concepts. Such relations reflect the paradigmatic organization of the lexicon. Also, there are many ways that the structure of a language can indicate the meaning of lexical items.

This paper is organized as follows: In the next section we briefly outline the main research on the automatic discovery of lexical semantic relations, which motivates the present study. Then we discuss the concept of troponymy between verbs. Section 3 introduces our proposal and experiments. Section 4 shows the results and discussion of this method; Section 5 concludes this paper with future directions.

2 Literature Review

There has been a variety of studies on the automatic acquisition of lexical semantic relations, such as hypernymy/hyponymy [6], antonymy [7], meronymy [5] and so on. In Section 2.1 we will review Hearst's approach, which most of the works on automatic labeling of word sense relations are based upon. To the best of our knowledge, there is no study targeting at troponymy extraction yet, so in Section 2.2, we first define what troponymy is, the complexity of troponymy, and discuss how we can infer troponymy motivated by Hearst's approach.

2.1 Syntactic patterns and semantic relation

The structure of a lexical entry in a dictionary reflects the relatedness of words and concepts; also, certain structures or syntactic patterns usually define the semantic relation among each other. Hearst [6] proposed a **lexico-syntactic pattern** based method for automatic acquisition of hyponymy from unrestricted texts. Basing on a text corpus, which contains terms and expressions that are not defined in Machine Readable Dictionaries, she postulates six lexico-syntactic

[1] http://wordnet.princeton.edu

[2] http://cwn.ling.sinica.edu.tw/

patterns to automatically detect hypernymy-hyponymy relation and extract these pairs from the sentences. Lexico- syntactic patterns which denote the concept of "including" or "other than" may often possibly reveal the hypernymy-hyponymy relation. The six syntactic patterns used in Hearst's algorithm are as follows: *(1) X such as Y; (2) such X as Y; (3) Y, or other X; (4) Y, and other X; (5) X, including Y; (6) X, especially Y.* For terms that are present in the above patterns, this algorithm successfully captures the relation that Ys are hyponymy of Xs.

According to Miller [4] and Fellbaum [2], the lexical database WordNet resembles a thesaurus in that it represents word meanings primarily in terms of conceptual-semantic and lexical relation. A synset, therefore, is constructed by 'assembling a set of synonyms that together define a unique sense.' If one sense of a word is the same to another word, they share the same synsets and they are synonyms (at least partial synonyms). Ramanand and Bhattacharyya [8] hence use the concept of synset to suggests that 'if a word w is present in a synset along with other words $w_1, w_2, \dots w_k$, then there is a dictionary definition of w which refers to one or more of $w_1, w_2, \dots w_k$, and/or to the words in the hypernymy of the synset.' With this assumption, he applies groups of rules to validate synonymy and hypernymy relation among corpus. For the rule which can denote hypernymy, the author defined that the definitions of words for particular senses often make references to the hypernym of the concept. Also, another rule detected partial hypernymy: many words in the wordnet are made up of more than one word, which are called 'multiwords'. In many cases, hypernyms or synonyms of such words are not entirely present in the definitions of words, but parts of them can be found in the definition.

2.2 Troponymy

As known, synsets in WordNet are connected with each other by various kinds of lexical semantic relations, such as Meronymy and Holonymy (between parts and wholes), Hypernymy and Hyponymy (between specific and more general synsets) and so on. Among them, the most important semantic relation in Wordnet is the hypernymy/hyponymy relation which links general and more specific concepts in both directions [7]. In Fellbaum's study [3], she defined the hyponymy-hypernymy relation among verbs as troponymy. Basing on this definition, troponymy, may at first sight appear like the relations of hypernymy/hyponymy among nouns: The subordinate concept contains the superordinate, but adds some additional semantic specification of its own. However, the semantic organization of verbs is more complicated than that of nouns and "the semantic relations so important for the organization of the noun lexicon are not the relations that hold among verbs" [4]. Hence, not all verbs can be placed under a single top node and verbs do not seem obviously related in a consistent manner like nouns do. According to Fellbaum and Miller [4], saying that troponymy is a particular kind of entailment involves temporal co-extensiveness for the two verbs. As known, entailment is a unilateral relation, taking *snore* and *sleep* for example, *snoring* entails *sleeping* but not the other way around. Although *snore* entails *sleep* and is included in *sleep*, we can not say that *snore* is a troponym of *sleep*; these two verbs are not in a hypernymy/ troponymy relation. Hence, for troponymy to hold, the essential factor is the co-extensiveness in time: one can sleep before or after snoring, but not necessarily happened at the same time. On the other hand, the activities denoted by the hypernym/ troponymy relation verbs must be coextensive in time. The following is an example from Lin et al [7]. *Reason* is a troponym of *think* because to *reason* is to *think* in a particular manner (logically). Therefore, the definition of *reason* naturally includes *to think "at the same time"* and thus inherits the property of *think.*

Beside the complicated distinction among verbs themselves, the troponymy relation is also different from the *is-a* relation among nouns in two ways [1]. First, the *is-a-kind-of* formula linking semantic related nouns may cause oddness when applying to verbs. For example, "(to) yodel is a kind of (to) sing." sounds odd only when changing into gerund form "yodeling is a kind of singing" will make it acceptable. Second, in the case of nouns, *kind of* can be omitted without changing the truth statement, for instance, "A donkey is a kind of animal." equals "A donkey is an animal." By contrast, the same deletion makes verbs odd as the following sentences show: " Murmuring is talking/ To murmur is to talk". These differences indicate that there is more than just a *is-a* relation among concepts expressed by verbs and the way that used to distinct nouns and adjectives is not the same as the way we distinct verbs. Rather than *kind*, troponymy seems to link verbs in a *manner* elaboration. Basing on the above properties of troponym, we postulate two syntactic patterns as the possible environments for discovering troponyms. More details will be discussed in the following section.

Our literature survey revealed that although some work had been done in automatically detection of hypernymy-hyponymy relation, none of them focus on hypernymy-troponymy relation of verbs. Therefore, in this paper, we attempt to propose a lexical pattern-based algorithm to tackle with this issue.

3 Algorithm

To automatically label the troponymy relation among verb senses, in the following, we propose an algorithm which applies three main steps and two rules.

3.1 First step: finding word senses

Most of words have more than one sense, and each sense of a given word might have their different hypernyms and troponyms. Therefore, to find semantic relations among verbs, our first step is to extract the definitions of each verb, by using web search. The input data used here composed 168 verbs which were extracted from *Sinica Corpus*. Although the 168 input verbs were randomly chosen by the authors, we firstly delimited out inputs labeled with syntactic categories VA, VC and VAC, respectively[3] for they contain most verbs that are commonly used. We then do search queries of each verb on Chinese Wordnet. If the result for a given word cannot be found here, then turn to the online version of the MOE Revised Chinese Dictionary[4] to find each sense of a given verb.

3.2 Second step: word segmentation

For later rule application, our next step is to do segmentation and POS (part-of-speech) tagging in each of the verbs' definitions via the online CKIP Chinese word segmentation system[5]. The following example shows the segmented result of one target verb and its definition:

買 (VAC) :	以 (P)	金錢 (Na)	購進 (VC)	物產 (Na)
mai :	yi	jingqien	gojin	wuchan
to buy :	with	money	purchase	products

to buy : to purchase products with money.

[3] According to both CWN and Sinica Corpus, verbs in Chinese are subdivided into 15different subcategories include VA VAC VC VB VCL VD VE VF VG VH VHC VI VJ VK VL.

[4] http://140.111.34.46/newDict/dict/index.html

[5] http://ckipsvr.iis.sinica.edu.tw/

By doing this, each word in this definition is segmented and POS tagged. After this step is done, our input data are established, which includes different entries[6] of each words' senses and all of the words are segmented and POS tagged.

3.3 Third step: Apply Rule 1 and Rule 2

After getting each of the verb's definitions, we propose two rules to find a given verb sense's hypernym.

3.3.1 Rule 1 Application

Rule 1: Definitions of verbs for particular sense often refer to certain or specific *manner* of their *hypernyms*. Hence, the definition may appear in the lexical syntactic pattern of '*以 (yi) / 用 (yong) . . .* V_j*. . . (by/with . . . . to* V_j*)*'.

Our first rule is, when a verb (V_i)'s definition contains the pattern *'以 (yi). . .* V_j *. . . ' or '用 (yong) . . .* V_j *. . . '* in a sentence, we could take this verb V_j or take all these verbs in this sentence out if there is more than one verb in this sentence. The verb(s) V_j could be labeled as a hypernym of V_i. For example, the definition of the verb 走 (zou, 'to walk') correspond to this pattern:

走: 以(P) 兩(Neu) 腿(Na) 交互(D) 向(P) 前(Ncd) 移動(VAC)
zou : yi lian tuei jiaohu xian qien yidong
to walk: by two feet mutually towards front to move
to walk: moving forwards by two feet.

Thus, 移動 (yidong, 'to move') could be labeled as the *hypernym* of the verbs 走 (zou, 'to walk').

3.3.2 Rule 2 Application

Rule 2: Deriving from the *is-a* relation of noun phrases, we may assume that, for verbs, a troponym is a certain way or a specific manner of its hypernym. Hence, the definition might appear in the pattern of '*一種 (yizhong) . . .* W_j *方式 (fangshi) (a way of* W_j*)*.

The second rule is, if a verb (V_i)'s definition contains the pattern *'一種 (yizhong) . . .* W_j *方式 (fangshi) (a way of* W_j*)*, then we could label this nominalized verb W_j as a hypernym of V_i. For example, the definitions of the verb 煎 (jian, 'to fry') is:

煎: 一(Neu) 種(Nf) 烹飪(VC) 方式(Na)
jian: yi zhong pengren fangshi
to fry: one kind to cook way
to fry: a way of cooking.

Since they follow the Rule 2 pattern, the verb 烹飪 (penren, 'to cook') is the hypernym of the verb 煎 (jian, 'to fry'), standing for a specific manner of cooking.

[6] Different entries in the data are separated by the step of tokenization.

3.4 Algorithmic Representation

The proposed algorithm outlined above could be summarized as the following:

```
Input: Verbs $V_1$, $V_2$, ... and $V_n$, web search for each definition of them.
Output: Predicted hyponyms between verbs
foreach definition of verbs $V_{df}$ do
    Word segmentation and POS tagging via CKIP;
    foreach definition of input verbs $V_i$ do
        check whether they contain the lexical syntactic pattern one;
        if matched then
            label the verb(s) $V_j$ as a hypernym of $V_i$;
        end
    end
end
while Unscheduled tasks remaining do
    foreach definition of verbs $V_{df}$ do
        check whether they contain the lexical syntactic pattern two;
        if matched then
            label the nominalized verb $W_j$ as a hypernym of $V_i$ ;
        end
    end
end
```

Algorithm 1: Algorithm to automatic labelling of troponomy

4 Experiment and Result

We implement our proposed method in Python (2.5.2). The module at first aims to extract entries containing our targets: '以 (yi) (P) / 用 (yong) (P) / 一種 (yi zhong) (Nf)' in the definition. Afterwards, all the verbs that occur after the targets will be extracted as the possible candidates. Figure 1 illustrates some of the results of a run of the labeling algorithm on Python, where the verbs occurred before the symbol '@' are the input verbs and other verbs occurred after '@' are their possible hypernyms.

```
IDLE 1.2.2      \verb==== No Subprocess \verb====

抖3(VAC)@
  重複(VC)
  晃動(VAC)
搖晃2(VAC)@
  搖動(VC)
震2(VAC)@
  震動(VAC)
  搖撼(VC)
...
```

Figure1: results run by the module

5 Evaluation and Discussion

To evaluate the system, we adopt the **substitution tests** [9] to examine whether a predicted pair of verbs has the relationship of troponymy. The specific substitution test that we implement is introduced in section 5.1. The acceptability of the sentences which contain the pairs of verbs we are examining is decided manually by 3 linguists. In section 5.2, we calculate the precision rate to properly evaluate the system. Some problems and further directions of our work are discussed in section 5.3.

5.1 The Substitution Test

Substitution test is commonly used in linguistic literature [9]. We apply the sentence pattern to the possible hypernyms found by our module: “如果他在 V_1 -ing, 那麼他便是在 V_2 -ing” (If he is V_1 -ing, then he is V_2 -ing). If the sentence “If he is V_1 -ing, then he is V_2 -ing” is always true but the other way around “If he is V_2 -ing, then he is V_1 -ing” is not, we say that V_2 is V_1’s hypernym and V_1, on the other hand, is V_2’s troponym. For example, we place the two verbs 走 (zou, ‘to walk’) and 移動 (yidong, ‘to move’) into this sentence. The result is that “If he is walking, then he is moving” is true, and that “If he is moving, then he is walking” is not necessarily true because if he is moving, he can also be running. We can therefore describe that 走 (zou, ‘to walk’) and 移動 (yidong, ‘to move’) are in a trponymy relation.

The reason for testing both these two sentences is to avoid the synonym pairs. We take two synonymous verb 立 (li, ‘to stand up’) and 站 (zhan, ‘to stand’) for example. “If he is standing up, then he is standing” is true, but (to stand up) is not (to stand)’s troponym. On the contrary, if we test by two sentences, this situation can be avoided. The second sentence “If he is standing, then he is standing up” is also true. Thus they are not in the relation of troponymy. By placing the verbs in the sentence pattern, we manually evaluated all the possible hypernyms found by our system and then we calculated the precision rate of the system.

5.2 Evaluation

The precision rate of our system is calculated as # `of correct answers given by system` / # `of answers given by system` = $\frac{93}{133} = 69.9\%$. There are totally 133 possible hypernyms returned by our system, and after manually filtered with the substitution test, there are 93 verbs left. The system shows that there are still some unwanted or incorrect hypernyms returned by our system. This will be further discussed in the following section. Although our system could have further improvement, but the high precision rate shows that our system has certain quality of performance and the design of our system is in the right direction.

5.3 Problems

There are two major problems in our system. First is the problem of synonym. In our input, the possible hypernyms returned by our system may be the synonym to the input verb. For example: for the input 揮 (huei, ‘to wave’), two possible hypernyms were returned by the system, including 揮 (VC) (huei, ‘to wave’), and 移動 (VAC) (yidong, ‘to move’).The underlined is an undesired synonym. Such instances will lower the precision rate. The second problem, and might be the most difficult one, is the ambiguity of the preposition 以 (yi, ‘by or woth’) and 用 (yong, ‘by or with’). In CWN, when 以 (yi, ‘by or with’) serves as a preposition (it could also serve as conjunction or modal), there are over 20 different polysemous meanings including '*because of*', '*according to'*, *'in order to'*, '*with'* etc. In this case, we have to determine which meaning the preposition '以 (yi) / 用 (yong)' belongs to and do the sense determination manually. This could remain a main problem for computational linguistics since so far, the disambiguation of these polysemy could not be fully solved using any system or algorithm.

Yet it is also very likely that hypernyms exists in inputs other than the patterns that we suggest. How to include those instances based on other methods, for example, by taking suffix-like forms as an indicator of troponymy (想 / 細想、回想 etc), is what we will consider in the future research.

6 Conclusion

In this study, we have suggested an automatic labeling method of troponymy for Chinese verbs. We identify the lexical-syntactic pattern '以 (yi) / 用 (yong)... V_j ... (by/with ....to V_j)' or '一種 (yizhong) ... W_j 方式 (fangshi) (a way of W_j)' that occur in the definition for V_i to see whether V_j and W_j indicate the troponymy relation of V_i. Though the range of the data is not exceedingly extensive, our approach has the advantages of low-cost and less-effort over other methods for automatic acquisition of lexical semantic relations from unrestricted text. Possible future works could be test the similar methods on the hypernymy/hyponymy relations among nouns. We beleive the comparison of the results will speed up the construction of Chinese lexical semantic relations knowledge base and ontology as well.

References

[1] C. Fellbaum. English verb as a semantic net. *International Journal of Lexicography*, 3:181–303, 1990.

[2] C. Fellbaum. *On the semantics of Troponymy*. Cognitive Science Laboratory, Princeton University, 2002.

[3] C. Fellbaum. On the semantics of troponymy. In Rebecca Green, A.Carol Bean, and H.M.Sung, editors, *The semantics of relationships: an interdisciplinary perspective*, pages 23–24, 2002.

[4] C. Fellbaum and G. Miller. Folk psychology or semantic entailment?– a reply to rips and conrad. *The Psychological Review*, 97:565–570, 1990.

[5] R. Girju, A. Badulescu, and D. Moldovan. Automatic discovery of part-whole relations. *Computational Linguistics.*, 31(1), 2006.

[6] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *proceedings of the Fourth International Conference on Computaional Linguistics (COLING)*, pages 539–545. Nantes, France, 1992.

[7] D.K. Lin, S.J. Zhao, L.J. Qin, and M. Zhou. Identifying synonyms among distributionally similar words. In *IJCAI-03*, 2003.

[8] J. Ramanand and Pushpak Bhattacharyya. Towards automatic evaluation of wordnet synsets. In *Global Wordnet Conference (GWC08)*, 2008.

[9] Bo-Sheng Tsai, Chu-Ren Huang, Shu-Chuan Tseng, Jen-Yi Lin, Keh-Jiann Chen, and Yuan-Shun Chuang. 中文詞義的定義與判定原則. ***中文信息學報** (Journal of Chinese Information Processing)*, 16.4:21–31, 2002.

電腦輔助中學程度漢英翻譯習作環境之建置

賴敏華　　　　劉昭麟
國立政治大學資訊科學系
{g9523, chaolin}@cs.nccu.edu.tw

摘要

電腦輔助教學系統主要在於幫助教師教學與學生自學。本研究提供能協助學生學習中翻英的優良環境，透過提供適當的參考資料，如相似文法與相似句型等，可增進學生對單字與文法的熟悉度，並提升學生學習英文的興趣及能力。本研究主要針對中學生程度設計，利用自然語言處理技術，藉由輸入中文句或中英文混合句，經文法、詞性及結構樹分析後，能提供相似的中英文句子，給予適當的英文翻譯建議，讓使用者有更多方向的參考。

關鍵詞：電腦輔助語文教學、句型搜尋、例句式教學、電腦輔助翻譯

1. 緒論

英文在現今社會上是非常普遍的語言，是地球村中不同母語背景的人用來相互溝通以及傳達訊息、信念不可或缺的工具。近年來，政府的教育政策以及家長對於小孩的期待，開始學習英文的年齡層下降，國小中高年級已排入英語課程、教育部國民教育司公佈國中常用兩千字字彙以及基礎一千字字彙[20]、各大專院校為學生英文程度把關，紛紛訂定畢業門檻等，由此可看出學習好英文、提升英文能力是未來生活重要的一環。

為了協助學生在課餘時間自學，各式各樣多元的輔助教學系統如雨後春筍出現，電腦輔助教學系統 (computer assisted tutoring system) 可以協助學生在課餘時間自我學習。目前許多學科，如：語言、數學、自然科學等教學，皆可以利用電腦技術與應用軟體設計出生動的互動教學軟體，也可以利用電腦輔助出題系統協助老師出題、批改。

本系統的中英文語料來源為從網路上收集，適合中學生或全民英檢中級與中高級程度的文件，包括教育部委託宜蘭縣建置語文學習領域國中教科書補充資料題庫[19]、旋元佑文法[16]、基礎英文 1200 句[17]、國民中學學習資源網[18] 的評量題庫及資源手冊等，經由人工擷取，建構中英文對照語料庫，以及利用現有的自然語言處理工具，中研院的中文斷詞系統將語料庫中的字彙標記其詞性以及中文句結構樹檢索系統建立中文句的結構樹，來建立標記化語料庫 (tagged corpus) 。這個標記化語料庫即是中英翻譯推薦句的來源。

本篇論文的組織架構如下：第二節為文獻探討；第三節為詳細描述本系統建立標記化語料庫的步驟；第四節則介紹本系統所提供的功能；第五節為系統評估，並於第六節提出簡單的結語。

2. 文獻探討

隨著電腦的普及化，電腦輔助教學系統在學生自我學習與教師教學上逐漸扮演著重要的角色。電腦輔助教學系統不但可以協助教師出題、閱卷、統計分數及對於學生學習效果作審慎的評估；亦可協助學生在沒有老師的陪同下自我學習，透過電腦輔助教學系統的回饋機制，系統可以依據學生所需，提供適合的教材，進而提升自我學習的效果。

有關電腦輔助出題系統，Mitkov和Ha[8]在2003年提出採用自然語言技術來自動產生閱讀測驗的試題。用來產生試題的句子皆具有「主詞+動詞+受詞」或「主詞+動詞」結構，而選擇為考題答案的通常是特定領域的詞彙。利用計算詞頻數以及WordNet[7]查詢詞的定義，來擷取英文句子中的關鍵詞作為考題答案。使用WordNet找尋語意上近似的觀念作關鍵詞的誘答選項，或從語料庫找尋語意不相似，但具有部分相同關鍵詞的片語或複合詞作為誘答選項。

Liu等學者[5]在2005年提出電腦輔助英文字彙出題系統之研究，利用詞性標記、字彙頻率的統計資料以及selectional preferences[6]的技術，配合搭配詞 (collocation) 概念與機器學習，從訓練資料歸納出法則，來輔助產生題目中的誘答選項。此研究只單純的利用搭配詞的訊息並沒有深入到語意層面的分析，故無法確保所產生的試題語義的完整性，仍須經由人為篩選作最後的確認。

陳佳吟等學者[21]在2005年提出電腦輔助英文文法出題系統，FAST (Free Assessment of Structural Tests) ，對於各式文法考題依照題型作分析，將考題分為九大類型，利用Tagger撰寫分類規則，使用Wikipedia網頁上收集有意義的句子為試題的主要來源。產生的題型為傳統四選一的單選文法試題。基於題型不同，會有不同出題方式的誘答選項，即不存在某種適用於全部題型的誘答選項產生方式，所以根據不同題型的誘答選項都是針對其特性去撰寫規則來產生合適的誘答選項。

林仁祥等學者[15]在2007年提出國小國語科測驗卷出題輔助系統，包含四聲辨識、中文克漏詞、改錯字、量詞等試題。其中中文克漏詞部分，使用HowNet[2]找尋相同義原的詞彙以及中研院現代漢語語料庫一詞泛讀[14]的學習工具將近義詞回傳，給予出題教師更多誘答選項的選擇。改錯字部分，提供同音字、相似字兩種功能，同音字利用新酷音輸入法的詞庫檔，給予相同發音的國字視為誘答選項參考；相似字利用倉頡碼建構構字式檔案，提供具有部分相同偏旁的中文字當作誘答選項參考。

除了電腦輔助出題系統外，另有電腦輔助自我學習的系統，其主要目的是在沒有老師的伴隨下可以輔助學生自我學習各個學科（如：語言、數學、科學等）；而語言學習上可分為聽、說、讀、寫四大部分，目前針對「讀」的方面，有較多相關的電腦輔助系

統；在「寫」的方面則較少。對於閱讀的輔助，多為網頁式 (Web-based) 即時的輔助翻譯，提供單字的解釋，進而為整個網頁的翻譯。

Weir和Lepouras[10]在2001年提出一個Web-based的自動註解英語的資訊為希臘文及中文。由於語言學習者對於新的單字或是不熟悉的字會有文字誤用、不同的定義、聯想錯誤或是對於上下文有所誤解，甚至於語言學習者並不熟悉語言上的特殊用法（如：專業術語、俚語或是慣用語）而導致閱讀上的困難。論文中將相關的對應文字在離線狀態 (off-line) 先行建立，再透過動態的連結來查詢英文單字的希臘文或中文的定義及解釋。由於希臘文與英文皆由羅馬字母構成，故中文翻譯比希臘文翻譯要來的複雜、效果也沒有希臘文來得好。

關於寫作方面，在第二語言的電腦輔助自我學習系統，其他語言也有相關研究。Kakegawa等學者[3]在2000年提出電腦輔助學習第二語言（日語），日文字主要可區分成 “用言 (Yougen) ” 以及 “体言(Taigen) ” 兩種，“用言” 為有語尾變化的詞性，如：動詞、形容詞；“体言” 則是沒有語尾變化，如：名詞、代名詞、指示詞 (demonstratives) 。論文中提出使用LTAG (Lexicalized Tree Adjoining Grammar) 來分析日文句子，句子結構樹採用bottom-up方式配合堆疊來建立，系統可以提供相似詞意的詞以及針對有語尾變化的日文字作偵錯，對於不恰當的用法可立即更正或是提供簡單的文字描述讓學生自行修正。

Knutsson等學者[4]在2003年提到有關第二語言（瑞典語）寫作學習環境文法檢查技術。面對非母語使用者在撰寫句子常常會有不可預期的文法型態產生，為了能提供學習者組織以及修正文章，需要足夠支持的語料資源。文章中，字詞的形態錯誤可由語言工具挑出；對於句法錯誤的部分僅僅提供錯誤訊息與建議，而非給予制式的解答。

Chang 和 Schaller [1]在 2005 年提到在學生撰寫英文作文時，透過互動式鍊結文法 (link grammar) ，可以作適時的英文文法確認，進而使學生提升書寫技巧；則教師在批改作文時，也可以專心致力於學生所想表達的意念，而不需專注於英文句文法的使用。提供使用者適時的文法確認的前提為，使用者必須對文法句型結構有相當足夠的認知，否則僅僅知道文法錯誤，卻不知道該如何修正，並無法達到良好學習的效果，若系統可以提供修正的句型參考則更能提升學習效果。

劉吉軒等學者[22]在 2007 年提出利用語言資訊檢索的方法，開發了名為 SAW (Sentence Assistance for Writing) 的雛形系統。此系統提供了完全比對與部分比對的檢索功能，為了允許不同語言程度的使用者能彈性的使用此系統，以正規表示法 (regular expression) 的概念為基礎，利用特殊符號來代表部分確定而部分設定範圍的查詢條件。對於查詢選取出的例句，利用多重序列排列 (Multiple Sequence Alignment) 的技術[9]進行查詢條件與選取例句之間相關程度評估與排序。

在語言寫作學習上，輔助系統提供寫作時，可為文句中的文法或是字詞的型態錯誤作偵錯並且給予修正的建議，但初學者可能對文法或句型完全沒概念，無法下筆將句子作適當翻譯或敘述想表達的意思，則偵錯的輔助系統也沒辦法發揮最大的效益。若輔助系統需要用較多符號去輔助來設定範圍查詢，使用者無法使用直覺去查詢、不夠人性化，

所以本研究的目標是設計並且實作出一個中英翻譯寫作輔助系統。學生透過查詢系統可以取得合適的英文翻譯以及類似句參考，藉由多組類似句可以學習到正確的單字用法以及文法知識。本節主要為介紹自然語言處理應用在電腦輔助教學上，但本系統技術上仍可參考 Example-Based Translation 的相關文章。

3. 建立標記化語料庫

本研究希望能提供一個簡單方便的系統介面，讓學習者可以輸入簡單的中英文關鍵字或句子，即可查詢到相關例句；並藉由本系統提供的中英翻譯推薦句，給予學習者在練習中英翻譯時有更多的參考資料，協助學習者在寫作時能達成既定的目標，達到學習的效果。

網路上提供程度適合國、高中生或全民英檢中級、中高級程度的中英文對照句資訊為數不多，大多數以學生學習單的樣式呈現，故採人工方式將中英文對照句從學習單中取出，此時語料庫內已經包含了對應好的中英文句子組合，每一組中的中英文句都是互為翻譯的句子，所以我們並不需要對語料作句子層次以上的處理。原始中英文對應的語料可經由圖 1 所列的流程，將人工擷取的中英文對照句的中英文句子，分別利用中研院的中文斷詞系統[12]及中文句結構樹系統檢索系統[11]進行詞性標記及結構樹的建立，將其結果回傳建立可用於查詢的標記化語料庫，以下為詳細描述各步驟的細節。

斷詞 (segmentation) ：中文與英文不同之處在於英文的詞與詞之間會以空白作為區隔，而中文的詞與詞是沒有空白，所以必須對中文作斷詞，本系統語料在詞性標記及中文結構樹建立時，會作中文句的斷詞，本系統是使用中央研究院所提供的中文斷詞系統[12]，中文斷詞系統會將中文句以常用的詞彙為基礎，將句子中的詞彙區隔出。如圖 2 所示，輸入的中文句為「我們都喜歡蝴蝶」，中研院中文斷詞系統回傳斷詞後的資訊為「我們(Nh)　都(D)　喜歡(VK)　蝴蝶(Na)」，其中 Nh 為代名詞，D 為副詞，VK 為狀態句賓動詞，NA 為普通名詞。

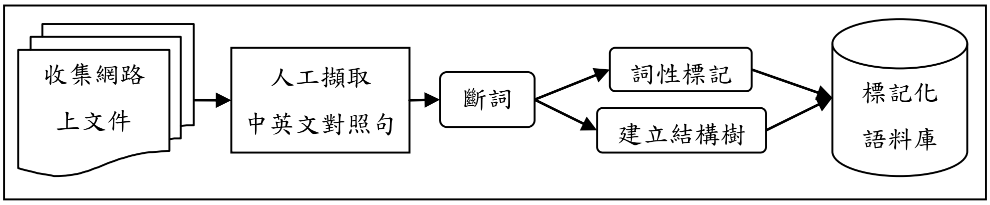


圖 1 建立標記化語料庫的流程圖

輸入：我們都喜歡蝴蝶

輸出：我們(Nh)　都(D)　喜歡(VK)　蝴蝶(Na)

詞性標記擷取： Nh D VK Na

圖 2 中研院中文斷詞系統輸出範例

詞性標記 (Part-Of-Speech tagging) ：將語料標記詞性，可以在搜尋功能中提供依照詞性來作為搜尋依據。因為擁有相同詞性順序的句子，句子相似度會越高，所以可依照詞性來給予使用者中英文類似句的推薦。本系統詞性標記是利用中央研究院所提供的中文斷詞系統[12]所回傳的斷詞結果會伴隨著詞性，將斷詞後的詞性擷取留下。如圖2所示，將中文斷詞系統回傳的結果中的詞性部分依序取出儲存，以“我們都喜歡蝴蝶”為例，擷取出的詞性順序為「Nh D VK Na」。

結構樹：透過結構樹可以知道中文句法、語意關係，兩個句子若結構樹的結構相似或是相同，可以猜測為語法相似。由結構樹的結構可以提供語法相似的例句，雖然使用的單字可能相差甚遠，但可以參考相同結構的句子，對於文法的搜尋有很大的幫助。本系統是利用中研院中文句結構樹資料庫檢索系統[11]作為取得結構樹的依據。如圖 3 所示，輸入的中文句為「她不覺得自己幸運」，接收中研院中文句結構樹系統回傳結構樹的資訊，並將結構樹依照分層取出。

我們將樹根定義為第 0 層，樹根的子樹為第 1 層，越往下層數字越大，故葉子節點為一個中文詞。在擷取各分層的結構樹時，樹根層（第 0 層）僅一個節點，不作記錄；依序將每一分層取出，假設某一中文句的結構樹深度為 i，而其一分支子樹最深深度為 j，則在第 j+1, j+2, …, i 層的詞性仍記錄第 j 層的詞性。將中研院中文句結構樹系統回傳結構樹的範例句資訊擷取各分層資料，如圖 3 所示，第一層結構樹為「NP Dc VK1 S」；第二層結構樹為「Nhaa Dc VK1 NP VH11」，第二層結構樹因為中文詞「不」與「覺得」的深度只有 1，所以在第二層結構樹詞性紀錄是記錄深度為 1 時的詞性「Dc」與「VK1」；第三層結構樹，中文詞「她」與「幸運」的深度只有 2，是記錄深度為 2 時的詞性「Nhaa」與「VH11」，中文詞「不」與「覺得」的深度為 1，詞性是記錄深度為 1 時的詞性「Dc」與「VK1」，故第三層結構樹為「Nhaa Dc VK1 Nhab VH11」。依照結構樹的分層取得各

輸入：她不覺得自己幸運

輸出：S(experiencer:NP(Head:Nhaa:她)|negation:Dc:不|Head:VK1:覺得
|goal:S(theme:NP(Head:Nhab:自己)|Head:VH11:幸運))

儲存結構：層數|第 n 層結構樹|……|第 1 層結構樹

範例：3 | Nhaa Dc VK1 Nhab VH11 | Nhaa Dc VK1 NP VH11 | NP Dc VK1 S

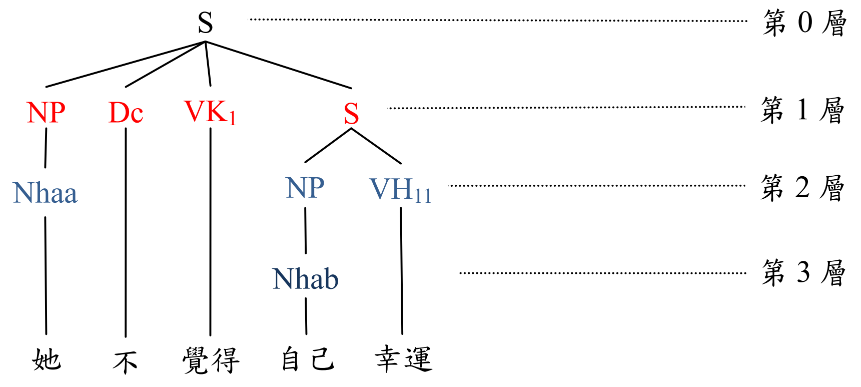
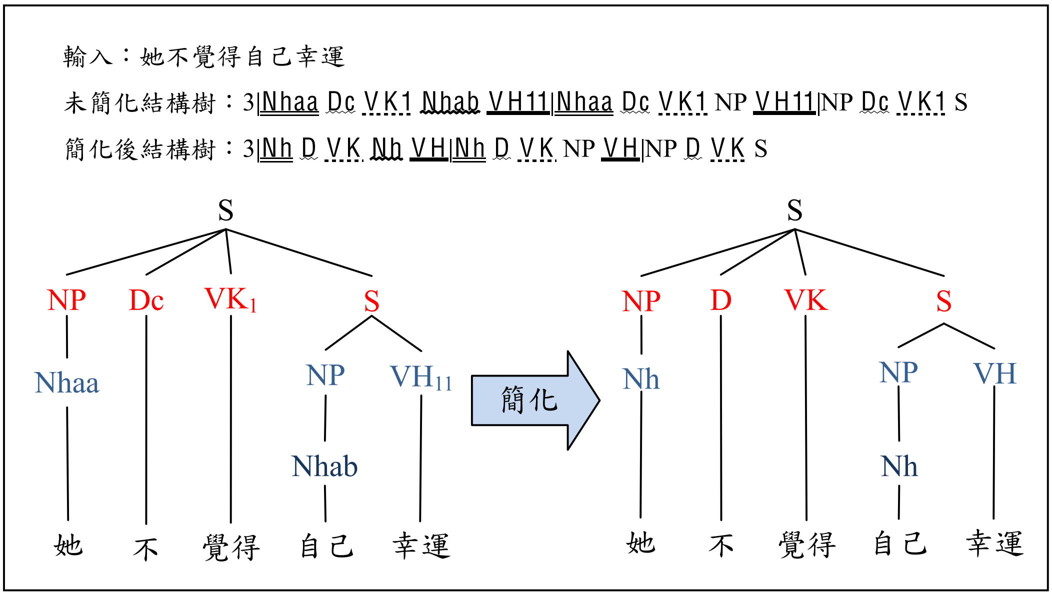


圖 3 中研院中文句結構樹輸出範例



輸入：她不覺得自己幸運

未簡化結構樹：3|Nhaa Dc VK1 Nhab VH11|Nhaa Dc VK1 NP VH11|NP Dc VK1 S

簡化後結構樹：3|Nh D VK Nh VH|Nh D VK NP VH|NP D VK S

圖 4 結構樹詞類簡化範例

表 1 平衡語料庫詞類標記集中四個簡化詞類對應表

簡化標記	對應的 CKIP 詞類標記	附註
Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	代名詞
D	Dab, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	副詞
VK	VK1,2	狀態句賓動詞
VH	VH11,12,13,14,15,17,VH21	狀態不及物動詞

分層的結構樹，儲存的結構為「層數|第 n 層結構樹|……|第 1 層結構樹」，而最後記錄在標記化語料庫資訊為「3|Nhaa Dc VK1 Nhab VH11|Nhaa Dc VK1 NP VH11|NP Dc VK1 S」。

由於結構樹回傳的詞性分類很細，共計有 115 種詞類，為了提升搜尋時提供更多的例句給予參考，我們根據中研院資訊科學所詞庫小組所編列的中研院平衡語料庫詞類標記集[13]，將標記化語料庫中的例句詞類由 115 種全面簡化成 46 種。以“她不覺得自己幸運”為例，如圖 4 所示，依據表 1 平衡語料庫的標記集（僅附例句中使用的詞類對應表）中得知，可將例句中的四個詞類作簡化，分別為代名詞、副詞、狀態賓動詞及狀態不及物動詞，圖 4 中將此四個詞類作相對應的底線粗體標示，並有簡化前後的樹狀結構表示。

4. 提供搜尋功能

在標記化語料庫建立完成後，可以透過自然語言處理的技術提供一些加值的服務。對於搜尋中文句的類似句，我們的系統利用中文詞、中文詞的詞性以及中文句結構樹的方法來分析處理並提供參考例句，本系統亦提供中英文混合的搜尋。

4.1 以中文詞為搜尋依據

當使用者輸入查詢句後，透過中文斷詞系統取得斷詞後的結果，利用斷詞結果在標記化語料庫搜尋；或是將斷詞的結果透過 HowNet 辭典[2]或是中研院現代漢語語料庫一詞泛讀[14]的學習工具取得與查詢句詞彙中相似的詞，來增加搜尋時中文詞彙的多樣性。

HowNet 辭典[2]是一個以中文和英文的詞所代表的概念為描述對象，以概念與概念之間以及概念所具有的屬性之間的關係作為基本內容的常識知識庫。在 HowNet 辭典中，每個詞彙多個欄位去記錄特殊資訊，如：中文詞條、中文詞性、英文詞條、英文詞性、義原關係等，其中一個欄位名稱為 DEF，描述著此詞彙的義原關係。在 HowNet 中，義原 (sememe) 是描述一個概念最小意義的單位，定義中英雙語知識詞典中的每個詞彙，並且建有描述各個義原之間關係的分類樹。例如：「讀書」一詞是由「從事」、「學」及「教育」三個義原定義而成，所以我們定義 S 為「讀書」義原的集合，S={從事，學，教育}，我們尋找 HowNet 中所有的詞彙，並將詞彙的義原與集合 S 作比對，與集合 S 有交集的詞彙都找出來，「讀書」一詞經過 HowNet 搜尋得到的結果為「攻，攻讀，苦讀，留，留美，念書，旁聽，求學，死記硬背，聽講，同窗，習作，修業，學到，學好，學習，學以致用，專攻，專修，自修，走讀」等，我們的系統會將這些詞彙視為與查詢句詞彙相似的詞，並加入搜尋時詞彙比對的依據。本研究中 HowNet 辭典是採用 1999 年版本。

中研院現代漢語語料庫一詞泛讀的學習工具[14]是利用電腦所收集的文本，針對一個詞彙，閱讀該詞彙出現的許多句子，記錄了各種該詞彙和其他詞彙共同出現的情形，作整理後所得的資料，可讓使用者藉由查詢更能掌握該詞彙的用法。我們的系統會將所要查詢的詞彙自動連到中研院一詞泛讀的網頁，網頁會回傳有關該詞彙的近義詞，本系統會將回傳的近義詞，視為搜尋時詞彙比對的依據。以「讀書」一詞經過一詞泛讀系統回傳所得到的近義詞詞彙為「學習，上，學，讀，念，修，讀書，就讀，念書，上學，入學，求學，攻，攻讀，就學，習，深造，修業，向學」。

如圖 5 所示，使用者輸入查詢句 T 後，經由中文斷詞系統回傳斷詞結果 $t_1, t_2, \ldots, t_n$，

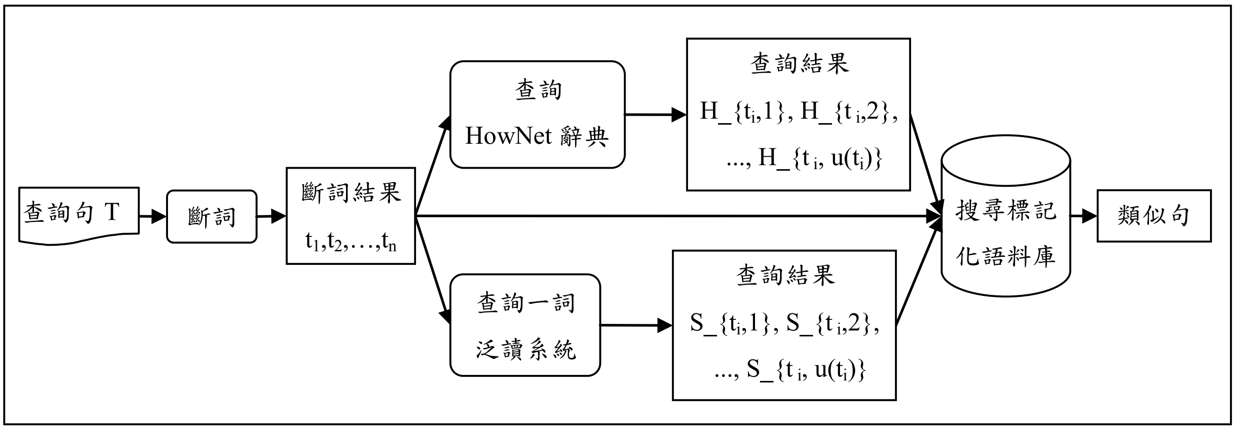

查詢句 T
斷詞
斷詞結果
$t_1,t_2,\ldots,t_n$
查詢
HowNet 辭典
查詢結果
$H\_\{t_i,1\}, H\_\{t_i,2\}, \ldots, H\_\{t_i, u(t_i)\}$
查詢一詞
泛讀系統
查詢結果
$S\_\{t_i,1\}, S\_\{t_i,2\}, \ldots, S\_\{t_i, u(t_i)\}$
搜尋標記
化語料庫
類似句

圖 5 使用中文斷詞結果搜尋

輸入：我有時上學遲到

中文斷詞輸出：我(N) 有時(ADV) 上學(Vi) 遲到(Vi)

擷取中文詞部分：我 有時 上學 遲到

利用一詞泛讀查詢近似詞彙：

我 身 個人 人家 本人 予 吾 余 咱 俺 儂 咱家 洒家 有時 學習 上 學 讀 念
修 讀書 就讀 念書 上學 入學 求學 攻 攻讀 就學 習 深造 修業 向學 遲到

輸出：

你上學常常遲到嗎?={Are you often late for school?}
我花了十分鐘走路上學。={It took me ten minutes to walk to school.}
我以前讀國中時，我都走路上學。={When I studied in junior high school, I used to walk to school.}
我每天騎腳踏車上學。={I go to school by bike every morning.}
前天他上學遲到了。={He went to school late the day before yesterday.}
上學讀書以前，他原本是個小頑童。={Before he was in school, he used to be a naughty child.}
我有時上學遲到。={I go to school late sometimes.}
我昨天和同學在圖書館讀書。={I studied with my classmates in the library yesterday.}

圖 6 使用中文詞一詞泛讀搜尋的範例

n 為詞彙個數，將 $t_1, t_2, \ldots, t_n$ 依序查詢 HowNet 辭典，取得查詢結果 H_{t_i,1}[†], H_{t_i,2}, ..., H_{t_i, u(t_i)}，其中 u(t_i)代表透過 HowNet 辭典所查到的第 i 個中文詞的近義詞的數量；或將 $t_1, t_2, \ldots, t_n$ 依序查詢中研院現代漢語語料庫一詞泛讀的學習工具，取得查詢結果 S_{t_i,1}, S_{t_i,2}, ..., S_{t_i, u(t_i)}，其中 v(t_i)代表透過一詞泛讀系統所查到的第 i 個中文詞的近義詞的數量。本系統會將查詢到的相似詞結果{H_{t_i,1}, H_{t_i,2}, ..., H_{t_i, u(t_i)}}或{S_{t_i,1}, S_{t_i,2}, ..., S_{t_i, u(t_i)}}與斷詞結果{$t_1, t_2, \ldots, t_n$ }作聯集，並將聯集結果於標記化語料庫作搜尋，提供類似句供使用者參考。

圖 6 為中文詞使用一詞泛讀搜尋的範例，我們輸入查詢句「我有時上學遲到」，經過中研院中文斷詞系統，得到斷詞後的結果為「我(N) 有時(ADV) 上學(Vi) 遲到(Vi)」，將所斷詞後的結果，擷取其中文詞，並利用一詞泛讀系統查詢得到每一個詞彙的近義詞後，與標記化語料庫中的中文句子作比對，若與語料庫中詞彙符合的句子，則會輸出視為類似句給予使用者作為參考例句。

4.2 以詞性為搜尋依據

在使用者輸入查詢的中文句後，本系統透過中文斷詞系統[12]取得斷詞後的結果，從斷詞的結果中把詞性依序擷取出來，接著透過搜尋標記化語料庫裡中文對照句的詞性，在僅考慮詞性出現順序，若相同即視為類似句。查詢句的詞性長度與標記化語料庫中英對照句所標記的詞性長度可能不相同，所以分為查詢句的詞性長度大於中英對照句詞性長度，以及查詢句的詞性長度小於或等於中英對照句詞性長度，兩種情形。

[†] X_Y 表示 Y 為 X 的下標，此底線為 LaTeX 語法

輸入：那個包包很好看
中文斷詞輸出：那(DET)　個(M)　包包(N)　很(ADV)　好看(Vi)
擷取詞性部分：DET M N ADV Vi

利用詞性搜尋：
我們這個週末要去露營了。={We are going camping this weekend.}
={N DET M N ADV ADV Vi ASP}
這個學生正在鐵路附近溜冰。={The student is roller-skating near the railroad.}
={DET M N ADV N N Vi}
那隻狗很凶惡。={The dog was mean.}={DET M N ADV Vi}
那個包包很好看。={That purse looks pretty.}={DET M N ADV Vi}
這個問題好像很容易。={The question seems easy.}={DET M N ADV ADV Vi}
我覺得這個包包很漂亮。={I find the purse pretty.}={N Vt DET M N ADV Vi}
那個島可以搭飛機或搭船去。={You can go to the island by plane or by boat.}
={DET M N ADV Vt N C Vi ADV}

圖 7 依照詞性搜尋範例

使用者輸入查詢句，經由中文斷詞系統回傳斷詞結果，並從中取得其相對應的詞性，透過與標記化語料庫中所記錄的中英文對照句的詞性標記資訊作比對，可以給予類似句當作參考依據。當查詢句的中文詞詞類個數小於或等於中英文對照句的詞類個數時，查詢句的中文詞詞類都依序出現在對照句的詞類序列裡面，則我們將此中文對照句視為查詢句的類似句，反之則不為查詢句的類似句；當查詢句的中文詞詞類個數大於中英文對照句的詞類個數時，對照句的詞類序列裡都依序出現在中文詞的詞類中，則我們將此中文對照句視為查詢句的類似句，反之則不為查詢句的類似句。如圖 7 所示，輸入的查詢句為「那個包包很好看」，輸出詞性標記順序為「DET M N ADV Vi」；搜尋後所得的部分結果如圖 7 所示，詞性比對成功部分以粗體底線標示之。

4.3 以結構樹為搜尋依據

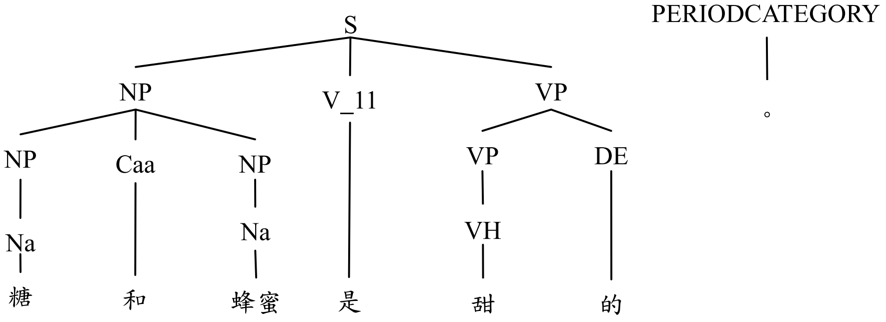
使用者輸入查詢句後，本系統可以透過中文結構樹取得查詢句的結構樹，為了查詢比對時能提供更多類似句供使用者參考，將結構樹中的詞性作簡化。依照結構樹階層分層取出，第 0 層（樹根層）不作記錄及搜尋，針對每一階層的結構樹，在語料庫中作結構樹比對搜尋，並將擁有相同結構樹的句子視為類似句給予使用者作參考。為了保有結構樹的原始結構，此功能只會針對每一階層去作完全比對以提供參考的類似句。若某一查詢句，經由中研院中文結構樹、詞性化簡以及擷取各分層的結構樹後，將各分層的結構樹與標記化語料庫中所記錄的結構樹作比對，其分層結構樹與標記化語料庫中的中英文對照句的結構樹相同時，則我們將此中文對照句視為查詢句的類似句，否則則不為查詢句的類似句。計算從最高層結構樹至第一層結構樹所查詢到的類似句，若較高層結構樹的類似句句數已達到 20 句，則不會再搜尋較低層結構樹的查詢句。

輸入：糖和蜂蜜是甜的。

中文結構樹輸出：S(theme:NP(DUMMY1:NP(Head:Naa:糖)|Head:Caa:|DUMMY2:NP(Head:Naa:蜂蜜))|Head:V_11:是|range:V(head:VP(Head:VH11:甜)|Head:DE:的)) %(PERIODCATEGORY:。)

簡化後結構樹：S(theme:NP(DUMMY1:NP(Head:Na:糖)|Head:Caa:和|DUMMY2:NP(Head:Na:蜂蜜))|Head:V_11:是|range:VP(head:VP(Head:VH:甜)|Head:DE:的)) %(PERIODCATEGORY:。)

結構樹：



分層結構樹： Na Caa Na V_11 VH DE PERIODCATEGORY
NP Caa NP V_11 VP DE PERIODCATEGORY
NP V_11 VP PERIODCATEGORY

利用結構樹搜尋：Na Caa Na V_11 VH DE PERIODCATEGORY

結果：糖和蜂蜜是甜的。={Sugar and honey are sweet.}
茶和咖啡是苦的。={Tea and coffee are bitter.}
柳橙和檸檬是不同的。={An orange and a lemon are different.}

利用結構樹搜尋：NP Caa NP V_11 VP DE PERIODCATEGORY

結果：糖和蜂蜜是甜的。={Sugar and honey are sweet.}
茶和咖啡是苦的。={Tea and coffee are bitter.}
柳橙和檸檬是不同的。={An orange and a lemon are different.}
布朗教授和太太是友善的。={Prof. and Mrs. Brown're friendly.}

利用結構樹搜尋：NP V_11 VP PERIODCATEGORY

結果：大衛是年輕而且強壯的。={David is young and strong.}
台灣是溫暖和潮濕的。={Taiwan is warm and humid.}
橋是方便的。={Bridges are convenient.}
誰是富有而且慷慨的？={Who's rich and generous?}
布朗教授和太太是友善的。={Prof. and Mrs. Brown're friendly.}
她的姐妹是害羞的。={Her sister is shy.}
台灣的天氣是溫暖和潮濕的。={The weather of Taiwan is warm and humid.}
她的洋裝的材料是進口的。={The material of her dress is imported.}

圖 8 依照結構樹搜尋範例

如圖 8 所示，輸入的查詢句為「糖和蜂蜜是甜的。」，透過中研院中文結構樹分析、詞性化簡以及擷取各分層的結構樹，得到「3|Na Caa Na V_11 VH DE PERIODCATEGORY|NP Caa NP V_11 VP DE PERIODCATEGORY|NP V_11 VP PERIODCATEGORY」的結果，利用每一階層結構樹去標記化語料庫中搜尋，取得與每一階層結構樹相同結構樹的句子。查詢句的第 3 層結構樹 (Na Caa Na V_11 VH DE PERIODCATEGORY) 比第 2 層結構樹 (NP Caa NP V_11 VP DE PERIODCATEGORY) 多搜尋到一句類似句，「布朗教授和太太是友善的。」，此類似句的分層結構樹為「3|Nb Na Caa Na V_11 VH DE PERIODCATEGORY|NP Caa NP V_11 VP DE PERIODCATEGORY|NP V_11 VP PERIODCATEGORY」，此類似句的各分層結構樹與查詢句的第 3 層結構樹不相符，所以在標記化語料庫中搜尋查詢句的第 3 層結構樹時，「布朗教授和太太是友善的。」並沒有被列為類似句給予推薦；而在標記化語料庫中搜尋查詢句的第 2 層分層結構樹時，「布朗教授和太太是友善的。」的分層結構樹中的第 2 層結構樹與查詢句的第 2 層結構樹相符，所以在此分層結構樹搜尋，「布朗教授和太太是友善的。」會被列為類似句推薦給使用者作參考例句。查詢句的第 1 層結構樹 (NP V_11 VP PERIODCATEGORY) ，因為搜尋到類似句句數較多，故僅附上部分結果。

4.4 中英文混合搜尋

本系統也提供中英文混合的搜尋，對於一道中翻英的題目，學生可能知道其中幾個中文詞的英文翻譯，而沒有英文文法的概念，或是可利用辭典查詢到各個中文詞彙的相對應英文字，卻因為不了解英文文法無法將查到的單字組織成一句完整的英文句子。透過輸入中英文混合的搜尋關鍵字，本系統可以幫忙尋找類似句，透過提供的類似句可以了解此句型結構。如：「宜蘭的空氣非常新鮮」，假設學生知道「空氣」與「非常」的英文字分別為「air」與「very」，則學生可以輸入「宜蘭的 air very 新鮮」，雖然輸入的查詢句

輸入：上學 bus

結果：我搭公車**上學**。={I go to school by bus.}

輸入：幾點 go school

結果：你**幾點**上學？={What time do you go to school?}

你**幾點**去上學？={What time do you go to school?}

輸入：You 很少 late

結果：你**很少**遲到。={You are seldom late.}

輸入：昨天 打 basketball

結果：我**昨天**跟麥克**打**籃球。={I played basketball with Mike yesterday.}

我**昨天**晚上沒有**打**籃球。={I did not play basketball last night.}

圖 9 中英文混合的搜尋範例

是不符合一般英文文法，但本系統可以透過搜尋，去找到類似句法的中英文對照句，進而協助學生練習中翻英。

對於使用者輸入的中英混合查詢句，本系統會依據使用者所輸入的中、英文在標記語料庫中分別搜尋中文句與英文句，若所輸入的詞彙在標記化語料庫中有被搜尋到，則本系統會將此中英文對照句子視為查詢句的類似句，輸出給予使用者作為參考。圖 9 為中英文混合的搜尋範例，可以輸入任意的中、英文，利用空白鍵作為分隔，本系統會根據所輸入的詞彙在語料庫中搜尋，輸出符合搜尋資訊的例句。假設學生不知道頻率副詞“很少”應該使用何字，則輸入「You 很少 late」來查詢，透過本系統查詢給予的建議例句可知道應使用「seldom」並且得知建議例句「你很少遲到。」與相對應的英文句「You are seldom late. 」；若輸入「昨天 打 basketball」來查詢，則可以得到兩句建議例句，分別為「我昨天跟麥克打籃球。={I played basketball with Mike yesterday.} 」以及「我昨天晚上沒有打籃球。={I did not play basketball last night.} 」。

5. 系統效率評估

本系統的中英文語料來源為從網路上收集，適合中學生程度的文件，包括教育部委託宜蘭縣建置語文學習領域國中教科書補充資料題庫[19]、旋元佑文法[16]、基礎英文 1200 句[17]、國民中學學習資源網[18] 的評量題庫及資源手冊等，標記化語料庫共計有七千多句中英文互為翻譯的語料。測試的資料類型為參考旋元佑文法[16]中，所提到的英文五大基本句型，分別為，句型一：主詞+動詞、句型二：主詞+動詞+受詞、句型三：主詞+動詞+補語、句型四：主詞+動詞+受詞+受詞，以及句型五：主詞+動詞+受詞+補語。

測試的資料來源為，賴世雄所編著的文法從頭學[23]、旋元佑文法[16]以及基礎 1200 句[17]中所挑選出來符合五大基本句型的例句，每一句型使用四句測試資料進行測試，透過本系統提供的以中文詞彙、詞性以及結構樹搜尋功能作查詢，並提供使用者類似句做為參考依據。

表 2 為利用五大句型例句使用本系統的搜尋功能所得到的類似句句數。以中文詞為搜尋依據，所得到的類似句以句型一（主詞＋動詞）句數較多，因為搜尋句本身字數較少，相對與標記化語料庫裡中文句詞彙數完全符合的機會較高，故所得到的類似句就會較多。句型三（主詞+動詞+補語）中，其中三句例句類似句句數也高達百句，其推測原因為在中文詞彙的近義詞詞數較多，所以在搜尋標記化語料庫中的中英文對照句，會有為數較多的類似句提供參考。

以詞性搜尋依據，搜尋句的詞性若與標記化語料庫中，中英文對照句的詞性順序相符，則會視為類似句給予推薦。句型較為簡單的句型一（主詞＋動詞）會得到較多的類似句，甚至高達上千句；而較複雜的句型的類似句，則會較少。平均來說以詞性為搜尋依據所得到的類似句大於以中文詞與結構樹為搜尋依據的句數，推測其原因為詞性比對始採用較寬鬆的詞性順序只要依序出現即視為類似句，而不是採用較嚴格的完全比對，故使用詞性比對搜尋會得到較多的參考類似句。

表 2 五大句型例句使用搜尋功能查詢所得到的類似句句數

	中文例句	以中文詞為搜尋依據所得類似句句數	以詞性為搜尋依據所得類似句句數	以結構樹為搜尋依據所得類似句句數
句型一： 主詞+動詞	我走路。	97	2655	6/17
	有事發生了。	232	64	0
	他過世了。	1189	1045	13/29
	你跑。	32	2655	6/17
句型二： 主詞+動詞+受詞	我愛她。	13	3877	1/4
	你是健康的。	38	658	18/21/141
	貓抓了一隻老鼠。	6	88	1/28
	他寫了一本書。	87	88	3/28
句型三： 主詞+動詞+補語	這個問題好像很容易。	1	69	4/23
	他繼續保持單身。	307	249	1/1/7
	他是個大英雄。	297	130	0/0/0/312
	他看起來很慈祥。	379	443	1/23
句型四： 主詞+動詞+受詞+受詞	老闆覺得你的提議很刺激。	1	31	2/2/2/33
	他餵貓吃罐頭。	42	611	1/1/39
	他給我一本書。	37	146	2/8
	大衛寫給蘇珊一封信。	1	146	1/8
句型五： 主詞+動詞+受詞+補語	他們覺得新房子很舒服。	1	27	1/2/33
	他把鑰匙留在那裡。	108	35	1/1/6
	他叫瑪麗擦窗戶。	44	611	0/3
	我聽到門被關了起來。	7	0	42

以結構樹為搜尋依據，所得到的類似句句數，不一定只有一個數值。因為結構樹會利用各分層結構樹在標記化語料庫中搜尋，所以表格中記錄了各分層結構樹搜尋標記化語料庫所得到的句數，記錄格式為「第 n 層結構樹查詢標記化語料庫所得類似句句數/……/第 1 層結構樹查詢標記化語料庫所得類似句句數」，若搜尋句僅只有一層結構樹，則只會有一個數值。由於結構樹層數越小，結構樹的結構較為簡單，所得到的類似句會越多；相反的，層數越大，所得到的類似句雖較少，但是類似句的句法會與查詢句較為類似。

針對不同的搜尋依據給予使用者類似句作為參考，會因為句型較為簡單而得到過多的類似句，或因為查詢句與標記化語料庫中的句型相似度不高，所以無法查詢到使用者期望的例句。我們期望能利用更多的語料或是針對查詢後提供過多的參考類似句，給予較嚴格的搜尋限制，將類似句句數降低，並期望本系統能達到輔助使用者學習英文、提升使用者對於英文文法的認知及熟悉度。

6. 結語

我們利用人工收集中學生英語學習單以及網路文件中，符合中學生程度的中英文對照句，並將語料作斷詞、詞性擷取以及結構樹建立等前處理，建立標記化語料庫。本系統提供以詞彙為單位、詞性、結構樹來搜尋相似推薦句，亦可透過輸入中英文混合字查詢推薦句，可以協助學生經由簡易的關鍵字搜尋到相似的中英文對照句給予建議，並且期望學生透過本系統所提供的多句相似文法或類似句型的中英文對照句，得以學習到正確的文法以及字彙的用法，並增進學生學習外語的興趣與能力。

目前本系統評估僅使用英文參考書中例句來測試本系統可提供的類似句句數，後續會設計使用者介面並請受測者來實際使用本系統，並且評估本系統效能以及是否達到輔助的效果。本系統可讓使用者利用中文輸入或中英文混合輸入，搜尋相關英文例句，適時給予英語學習者，在寫作時提供參考例句；我們期望若依照相同概念去分析並建構英文句結構樹，亦能讓外國人利用英文輸入或中英文混合輸入法，查詢到相關的中文建議例句，讓外國人也能利用本系統學習到中文句型或是中文文法概念。

致謝

本研究承蒙國科會研究計畫 NSC-95-2221-E-004-013-MY2 的部分補助謹此致謝。我們感謝匿名評審對於本文初稿的各項指正與指導。雖然我們已經在從事相關的部分研究議題，不過限於篇幅因此不能在本文中全面交代相關細節。

參考文獻

[1] Y.-F. Chang and D. L. Schallert, The Design for a Collaborative System of English as Foreign Language Composition Writing of Senior High School Students in Taiwan. *Proceedings of the Fifth IEEE International Conference on Advance Learning Technologies*, 774-775, 2005.

[2] Z. Dong and Q. Dong, HowNet, 2000. http://www.keenage.com [Accessed: Jun. 26, 2008]

[3] J. Kakegawa, H. Kanda, E. Fujioka, M. Itami and K. Itoh, Diagnostic Processing of Japanese for Computer-Assisted Second Language Learning. *Proceedings of the Thirty Eighth Annual Meeting on Association for Computational Linguistics*, 537-546, 2000.

[4] O. Knutsson, T. C. Pargman and K. S. Eklundh, Transforming Grammar Checking Technology into a Learning Environment for Second Language Writing. *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing*, Volume 2, 38-45, 2003.

[5] C.-L. Liu, C.-H. Wang, and Z.-M. Gao, Using Lexical Constraints to Enhance the Quality of Computer-Generated Multiple-Choice Cloze Items. *International Journal of Computational Linguistics and Chinese Language Processing*, Volume 10, Number 3, 303-328, 2005.

[6] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, the MIT Press, 1999.

[7] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller, Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, Volume 3, Number 4, 235-244, 1990. http://wordnet.princeton.edu/doc/ [Accessed: Jun. 26, 2008]

[8] R. Mitkov and L. A. Ha, Computer-Aided Generation of Multiple-Choice Tests. *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing*, Volume2, 17-22, 2003.

[9] S. B. Needleman and C. D. Wunsh, A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, *Journal of Molecular Biology*, Volume 48, Number 3, 443-453, 1970.

[10] G.R.S. Weir and G. Lepouras, English Assistant: A Support Strategy for On-Line Second Language Learning. *Proceedings of the Second IEEE International Conference on Advance Learning Technologies*, 125-126, 2001.

[11] 中研院中文句結構樹資料庫檢索系統，http://turing.iis.sinica.edu.tw/treesearch/ [Accessed: Jun. 24, 2008]

[12] 中研院中文斷詞系統，http://ckipsvr.iis.sinica.edu.tw/ [Accessed: Jun. 24, 2008]

[13] 中研院平衡語料庫詞類標記集，http://ckipsvr.iis.sinica.edu.tw/category_list.doc [Accessed: Jun. 24, 2008]

[14] 中研院現代漢語語料庫一詞泛讀，http://140.109.150.65/cwordframe.html [Accessed: Jun. 24, 2008]

[15] 林仁祥及劉昭麟。國小國語科測驗卷出題輔助系統，*2007 台灣網際網路研討會論文集*，論文光碟。台灣，台北，2007。

[16] 旋元佑文法，http://tw.myblog.yahoo.com/jw!GFGhGimWHxN4wRWXG1UDIL_XSA--/ [Accessed: Jun. 24, 2008]

[17] 基礎英文 1200 句，http://hk.geocities.com/cnlyhhp/eng.htm [Accessed: Jun. 24, 2008]

[18] 國民中學學習資源網，http://140.111.34.172/teacool/new_page_2.htm [Accessed: Jun. 24, 2008]

[19] 教育部委託宜蘭縣發展九年一貫課程建置語文學習領域（英語）國中教科書補充資料暨題庫建置計畫，http://140.111.66.37/english/ [Accessed: Jun. 24, 2008]

[20] 教育部國民教育司，http://www.edu.tw/EJE [Accessed: Jun. 24, 2008]

[21] 陳佳吟、柯明憲、吳紫葦及張俊盛，電腦輔助英文文法出題系統，*第十七屆自然語言與語音處理研討會論文集*。台灣，台南，2005。

[22] 劉吉軒、洪培鈞及李金瑛，以英語寫作輔助為目的之語料庫語句檢索方法，*第十九屆自然語言與語音處理研討會論文集*，5-19。台灣，台北，2007。

[23] 賴世雄，*文法從頭學*，長春藤有聲出版有限公司。2007。

以範例爲基礎之英漢 TIMSS 試題輔助翻譯

張智傑　　　　劉昭麟
國立政治大學 資訊科學系
{ g9512 ,chaolin }@cs.nccu.edu.tw

摘要

本論文應用以範例爲基礎的機器翻譯技術，應用英漢雙語對應的結構輔助英漢單句語料的翻譯。翻譯範例是運用一種特殊的結構，此結構包含來源句的剖析樹、目標句的字串、以及目標句和來源句詞彙對應關係。將翻譯範例建立資料庫，以提供來源句作詞序交換的依據，最後透過字典翻譯，以及利用統計式中英詞彙對列和語言模型來選詞，產生建議的翻譯。我們是以 2003 年國際數學與科學教育成就趨勢調查測驗試題爲主要翻譯的對象，以期提升翻譯的一致性和效率。以 NIST 和 BLEU 的評比方式，來評估和比較線上翻譯系統和本系統所達成的翻譯品質。

關鍵詞：自然語言處理，試題翻譯，機器翻譯，TIMSS

1. 緒論

國際教育學習成就調查委員會(The International Association for the Evaluation of Education Achievement, 以下簡稱 IEA)[20]主要目的在於了解各國學生數學及科學(含物理、化學、生物、及地球科學)方面學習成就、教育環境等，影響學生的因素，找出關聯性，並在國際間相互作比較。自 1970 年起開始第一次國際數學與科學教育成就調查後，世界各國逐漸對國際數學與科學教育成就研究感到興趣，IEA 便在 1995 年開始每四年辦理國際數學與科學教育成就研究一次，稱爲國際數學與科學教育成就趨勢調查(Trends in International Mathematics and Science Study，以下簡稱 TIMSS)，至今已辦理過 1995、1999、2003 和 2007 共四屆，共有 38 個國家參加。

我國於 1999 年開始加入 TIMSS 後，由國科會委託國立台灣師範大學科學教育中心(以下簡稱師大科教中心)負責試題翻譯及測驗工作。1999 年的調查對象只有國中二年級學生， 2003 年的調查對象包括四年級及八年級學生。翻譯試題主要的流程包含：從 IEA 取得試題內容，由師大科教中心決議進行翻譯工作分配、中文試題交換審稿校正及翻譯問題討論，最後將中文翻譯試題定稿。至目前爲止，師大科教中心已將 1999 和 2003 年試題內容和評量結果，公布於台灣 TIMSS 官方網站[21]，以提供研究之參考。在 TIMSS 的試題內容上，主要的題型種類有選擇題和問答題，試題句型大多爲直述句和問句結構所組成，選擇題則多了誘答選項。

以往使用人工翻譯雖然可以達到很高的翻譯品質，但是需要耗費相當多的人力資源和時間，而且在翻譯過程中不同的翻譯者會有不同的翻譯標準(例如：相同的句子，翻譯後的結果不同)；相同的翻譯者也可能在文章前後翻譯方式不一致而產生語意上的混淆。因此間接影響試題難易程度。若直接將英文詞彙透過英漢字典翻譯成相對的中文詞

彙，翻譯的結果可能會不符合一般人的用詞順序。另外中文的自由度較高，很容易造成翻譯上用詞順序的不同。例如："下圖顯示某一個國家所種穀物的分布圖"，也可翻譯爲"某一個國家所種穀物的分布圖，如下圖顯示"。可能會影響到受測者的思緒，使作答時粗心的情形會增加。因此，若能利用機器翻譯(machine translation)的技術來輔助翻譯以及調整詞序，以期提高翻譯的品質和效率。

在人工智慧領域，機器翻譯是一個很困難的問題。機器翻譯是指將一種自然語言經過電腦運算翻譯成另一種語言，困難程度也跟來源句和目標句有關，像是英文和葡萄牙文語言的特性較相近，較容易翻譯。而中文跟英文詞序差異很大，且中文比較沒有特定的語法，寫法較自由，對翻譯來說較爲困難。機器翻譯發展至今已經超過 50 年。Dorr 等學者[9]將現在機器翻譯依據系統處理的方式來分類，分成以語言學爲基翻譯(linguistic-based paradigms)，例如基於知識(knowledge-based)和基於規則(rule-based)等；以及非語言學爲基翻譯(non-linguistic-based paradigms) ，例如基於統計(statistical-based)和基於範例(example-based)等。

以知識爲基礎的機器翻譯(knowledge-based machine translation)系統是運用字典、文法規則或是語言學家的知識來幫助翻譯。Knight 等學者[11]結合 Longman 字典、WordNet 和 Collins 雙語字典建立一個知識庫，運用在西班牙文翻譯成英文。這種利用字典來幫助翻譯的系統，會有一字多義的情形發生，一個詞彙在字典中通常有一個以上的翻譯。以英翻中爲例"current"這個字在字典裡就有十多種不同的翻譯，即使專家也無法找出一個統一的規則，在何種情況下要用何種翻譯，所以在翻譯的品質和正確性上很難滿足使用者。因此，翻譯系統通常都會限定領域來減少一字多義，例如 current 在電子電機類的文章中出現，最常被翻譯爲電流，在文學類的文章中，最常被翻譯爲現代。

統計式機器翻譯(statistical machine translation，以下簡稱 SMT)是將語料在翻譯之前就已經過計算轉換成統計數據，不需要在翻譯過程中作龐大的數學運算，能有較高的效能。Brown 等學者[6]於 1990 年以英文及法文的雙語語料爲來源，提出統計式雙語翻譯架構。假設目標語言爲 T 及來源語言爲 S，P(T)爲目標語言 T 在語料庫中出現的機率，稱爲語言模型(language model)，P(S|T)爲目標語言 T 翻譯成來源語言 S 的機率，稱爲翻譯模型(translation model)。SMT 系統需要大量的語料庫輔助，大多都需要具備雙語對應的語料庫(parallel corpora 或稱 bilingual corpora)，再透過機率公式計算出機率模型。其中 SMT 困難的地方在於需要收集大量可用的雙語語料，當語料越多建立模型所花費的時間越多。Oct 等學者[16]提出單字式(word-based)翻譯模型運用在詞彙對準(word alignment)，並且發展出 GIZA++這套系統。Koehn 等學者[12]進一步將單字式轉變成片語式(phrase-based)翻譯模型，運用片語式翻譯模型翻譯的結果會比單字式翻譯的結果要正確。

以範例爲基礎的機器翻譯(example-based machine translation，以下簡稱爲 EBMT)的相關研究已有相當多年歷史，在 1990 年日本學者 Sato 和 Nagao[19]所提出的 EBMT 是將翻譯過程分爲分解(decomposition)、轉換(transfer)和合成(composition)三步驟。分解階段是將來源句到範例庫中搜尋，並將所搜尋到 word-dependency tree 當作來源句的 word-dependency tree，並且形成來源句的表示式；轉換階段將來源句的表示式轉換成目

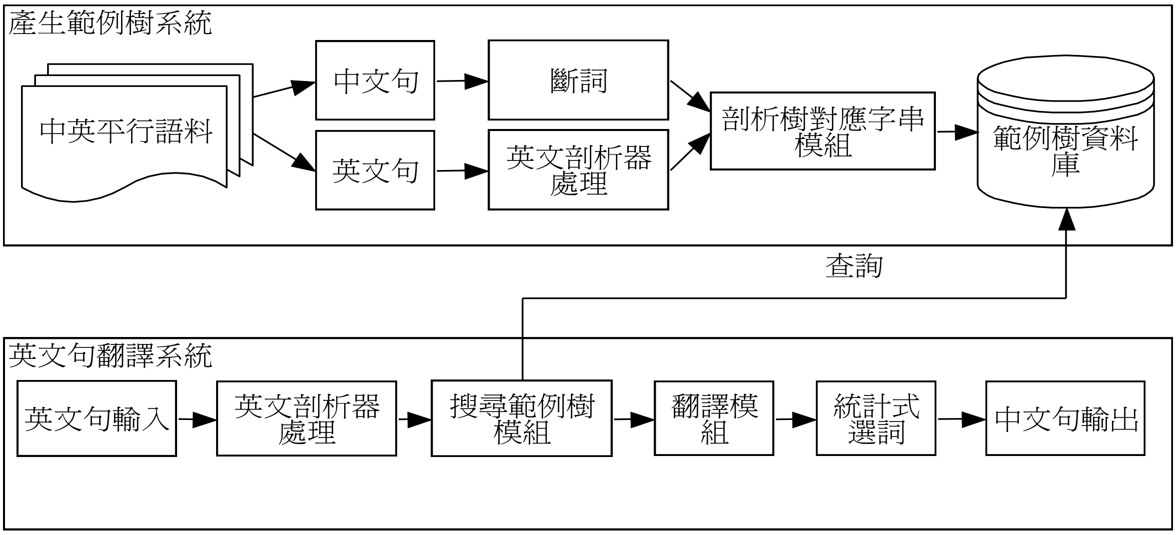
標句的表示式；合成階段將目標句的表示式展開爲目標句的 word-dependency tree，並且輸出翻譯結果。Al-Adhaileh 等學者[5]將 structured string tree correspondence(SSTC) [7]運用在英文翻譯成馬來西亞文上，SSTC 是一種能將英文對應馬來西亞文的結構，但此結構並沒有解決詞序交換的問題。目前較完整的 EBMT 系統有 Liu 等學者所提出 tree-string correspondence (TSC)結構和統計式模型所組成的EBMT系統[13]，在比對TSC結構的機制是計算來源句剖析樹和 TSC 比對的分數，產生翻譯的是由來源詞彙翻譯成目標詞彙的機率和目標句的語言模型所組成。

黃輝等學者所提出的 translation corresponding tree (TCT) [24]，TCT 是針對英文翻譯成葡萄牙文的系統，在 TCT 結構上可以記錄來源句詞彙和目標句詞彙對應的關係、來源句詞彙和目標句詞彙對應的翻譯結果和詞序，但是 TCT 是二元的剖析樹，也就是每個節點都只有兩顆子樹，在 TCT 上詞序只用布林值(boolean value)來記錄，所以 TCT 只能運用在二元剖析樹上。但是有些剖析器所產生剖析樹是多元樹，因此我們提出雙語樹對應字串的結構(bilingual structured string tree correspondence，簡稱爲 BSSTC)可以運用在多元剖析樹上，並且 BSSTC 可在翻譯過程中當作詞序交換的參考，根據我們實驗結果，我們能有效的調動詞序，以提升翻譯的品質。完成詞序交換後，再透過字典翻譯成中文，最後運用統計式選詞模型，產生了初步翻譯結果，但本系統尚屬於半自動翻譯系統，故需要人工加以修飾編輯。

除了本節簡單介紹本研究以外，我們將在第二節描述整個系統的架構，第三節說明本篇論文所運用的技術，第四節則呈現出我們的實驗結果，第五節則是結論。

2. 系統架構

由於我們的目的在於利用中英互爲翻譯的句子找出詞序關係，並且將英文句和中文句詞序的資訊儲存在電腦中，儲存的格式是將中英文句的詞序關係記錄在英文剖析樹的結構中，此結構將成爲之後英文句的結構調整爲適合中文的結構的參考。最後再將英文詞彙翻譯成中文詞彙，並利用統計式選詞選出最有可能翻譯成的中文詞彙，讓翻譯的結果更符合一般人的用詞和順序。



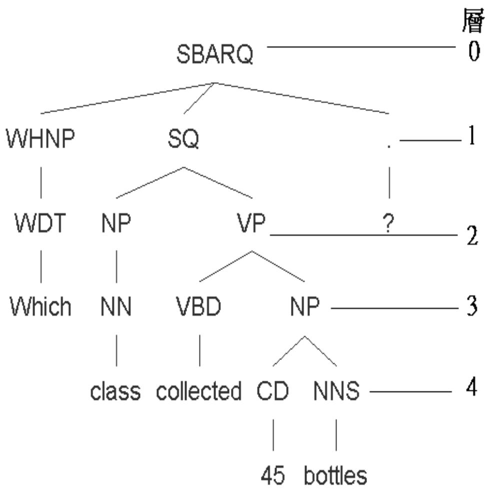
圖一、系統架構圖

本系統的架構如圖一所示。我們針對範例樹產生系統和英文句翻譯系統這兩部份分別簡介如下。

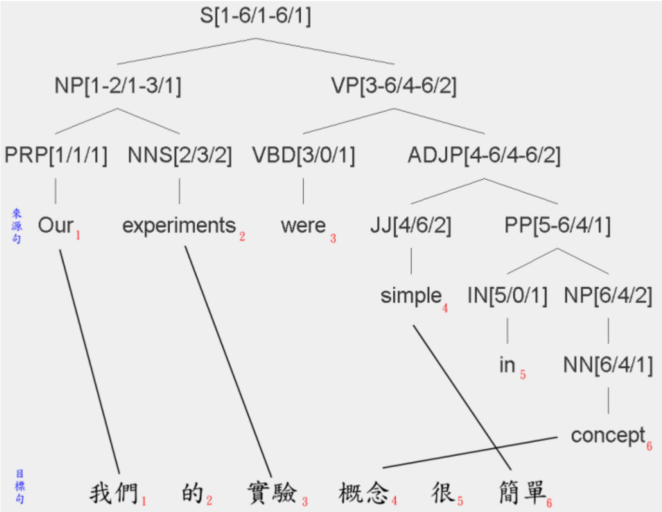
- **範例樹產生系統：** 這個系統利用中英平行語料，這裡的中英平行語料必需要一句英文句對應一句中文句，且每一組中英文句都要是互為翻譯的句子。中文句經過斷詞處理後，被斷成數個中文詞彙，以空白隔開；英文句經過英文剖析器建成英文剖析樹。將斷詞後的結果和英文剖析樹經過剖析樹對應字串模組處理，建成英文剖析樹對應字串的結構樹，此結構樹稱為範例樹。再將每個範例樹取出子樹，並且判斷是否有詞序交換，將需要詞序交換的範例樹全部存入範例樹資料庫中方便搜尋。
- **英文句翻譯系統：** 當輸入英文句後，先將句子透過英文剖析器，建成英文剖析樹。有了英文剖析樹就可以透過搜尋範例樹模組，標記英文剖析樹上需要調動詞序的結構，並依照所標記的詞序作調整。詞序調整完成後再將英文結構樹中的英文單字或片語透過翻譯模組做翻譯。其中翻譯模組包含了大小寫轉換、斷詞處理、stop word filtering及stemming，之後將處理過的詞彙透過字典檔做翻譯[3]。每個英文單字或片語都可能有一個以上的中文翻譯，因此需要選詞的機制來產生初步翻譯結果，此翻譯結果尚需要人工作後續的編修。

3. 系統相關技術

根據上一節系統架構的描述分為範例樹產生系統和英文句翻譯系統兩大系統。範例產生樹系統的執行流程為先將中文句斷詞和剖析英文句，再將斷詞和剖析後的結果輸入至剖析樹對應字串模組，並將處理後的範例樹存入資料庫中。英文句翻譯系統的執行流程區分為三大部分，第一部分是由搜尋範例樹模組，將英文剖析樹跟範例樹資料庫作比對，並且將未比對到的子樹做修剪；第二部分將修剪後的剖析樹輸入到翻譯模組翻成中文；第三部分以中英詞彙對列工具及 bi-gram 語言模型，計算出中英詞彙間最有可能之翻譯組合。



圖二、英文剖析樹



圖三、BSSTC 結構的表示法

3.1 雙語樹對應字串的結構(BSSTC)

在建立 BSSTC 結構之前，我們必須將中英平行語料中的中英文句先作前處理，我們將英文句透過 StanfordLexParser-1.6[17]建成剖析樹，剖析樹的每個葉子節點爲一個英文單字，並以英文單字爲單位由 1 開始標號。這裡我們將樹根定義爲第 0 層，樹根的子樹是第 1 層，越往下層數越大，故葉子節點必定是英文單字，且不屬於任何一層，如圖二所示。而中文句是使用中研院 CKIP 斷詞系統[1]作斷詞，並以斷詞後的單位由 1 開始標號。這裡的中文句代表來源句；英文句代表目標句。本結構是假設在中英文對應都是在詞彙的對應或連續字串的對應。

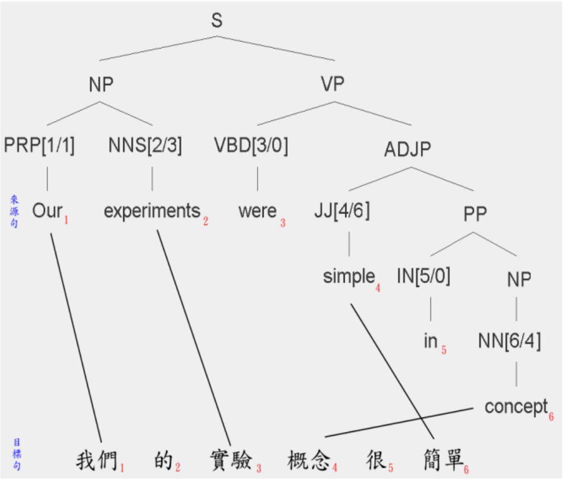
假設剖析樹的節點集合 $N=\{N_1, N_2, \ldots, N_m\}$，$m$ 爲剖析樹上節點個數，對任一節點 $n \in N$，n 有三個參數分別是 n[STREE//]、n[/STC/]和 n[//ORDER]；我們以 n[STREE/STC/ORDER]來表示。爲了方便說明，若節點 n 只有 n[STREE//]和 n[/STC/]，則以 n[STREE/STC/]表示。再假設 $n_{C(n)}$爲節點 n 有 1 到 C(n)個子節點。n[STREE//]爲節點 n 所涵蓋來源句的範圍，層數最大節點的 n[STREE//]必定對應到一個來源句單字，此參數的功用爲當作每個節點的鍵值(primary key)，故在同一棵剖析樹中 n[STREE//]不會重複。n[/STC/]表示以 n 爲樹根的子樹，所涵蓋來源句字串的範圍對應到目標句字串的範圍；n[/STC/]也可以是一個數字，表示此子樹包含的目標句字串爲目標句字串中的一個字；n[/STC/]也可能是 0，代表來源句無法對應到目標句。n[//ORDER]是由 n[/STC/]計算出來，n[//ORDER]是用來表示來源句跟目標句詞序對應的關係，若來源句跟目標句有詞序不同的情形，就可由 n 與所有兄弟節點的 n[//ORDER]來判斷。ORDER 的範圍由 1 到 C(n)，當 ORDER 越小，代表 n 所對應目標句範圍，比其他兄弟節點的目標句範圍更靠近句子的前段。

圖三是一個 BSSTC 結構的例子，來源句爲英文：“Our experiments were simple in concept”；目標句爲中文：“我們的實驗概念很簡單”。首先英文句必須先建成剖析樹，每個葉子節點爲一個英文單字，並以英文單字爲單位做標號，例如：“Our(1)”, “experiments(2)”, “were(3)”, “simple(4)”, “in(5)”,“concept(6)”。另外中文句經過斷詞的處理後，以斷詞後的單位做標號，例如：“我們(1)”, “的(2)”, “實驗(3)”, “概念(4)”, “很(5)”, “簡單(6)”。中英對應句都標號後，以標號爲單位開始做詞彙對準(word alignment)，並標記在剖析樹的節點上。剖析樹是用文法結構來分層，不同層節點能對應到不同的範圍的目標句字串。n[STREE/STC/]若爲 VP[3-6/4-6/]，則 STREE 代表節點 VP 對應來源句第三到第六個字 “were simple in concept“；STC 代表“were simple in concept”對應目標句的第四到第六個字“概念很簡單”。$n_{C(n)}$[STREE/STC/ORDER]的兄弟節點(sibling node)若爲 JJ[4/6/2]和 PP[5-6/4/1]，我們可以觀察到 JJ 的 ORDER 大於 PP 的 ORDER，故 PP[5-6/4/1] 的中文對應「概念」在 JJ[4/6/2] 的中文對應「簡單」之前。

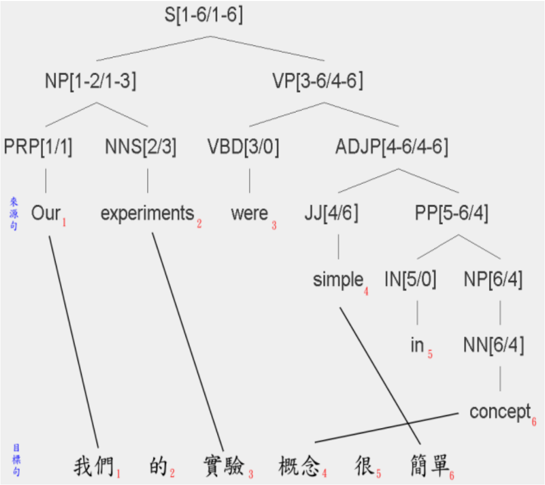
3.2 建立 BSSTC 結構和產生範例樹

建立 BSSTC 結構必需要有英文跟中文互爲翻譯的句子，建構的順序是從最底層也就是層數最大的開始標記，再一層一層往上建置到第 0 層爲止，標記參數順序是先將所有節點的 n[STREE//]和 n[/STC/]標記完後，再標記 n[//ORDER]。首先，標記最底層 n[STREE//]

的方法，是將最底層的節點 n 所對應葉子節點的編號標記在 n[STREE//]。如圖三節點 NNS 所對應來源句的“experiments”的編號爲 2，故 NNS[STREE//]中的 STREE 標記爲 2。接著標記最底層 n[/STC/]的方法是尋找中英對應句中互爲翻譯的中文詞彙和英文詞彙，也就是詞彙對準。詞彙對準若採用人工方式，則相當耗時費力，其本身也是一項困難的研究。因此，我們在此用一個簡單的方法，首先先將中文句經過斷詞處理，這裡我們使用中研院 CKIP 斷詞系統[1]；將英文句每個英文字查尋字典檔，查尋後可能會有超過一個的中文翻譯，將這些中文翻譯跟斷詞後的中文詞彙一個一個作比對，如有比對到則認定互爲翻譯，並且標記 n[/STC/]在剖析樹上。如圖三來源句的“experiments”在字典中的翻譯有“實驗”、“經驗”和“試驗”，將這三個中文翻譯到目標句去比對，此例子將會比對到目標句第三個詞彙“實驗”，接著將目標句“實驗”的編號標記在 NNS[2/STC/]中的 STC 上。最後將比對到的個數除以英文句單字的個數，稱爲對應率。最佳情況下是每個英文單字都有相對應的中文翻譯，對應率爲 1；最差的情況下每個英文單字都沒有相對應的中文翻譯，對應率爲 0，所以對應率會落在 0 到 1 之間，值越大代表對應率越高。我們需要夠大的對應率，才能認定爲範例樹。因此，需要定一個門檻值來篩選，根據實驗結果當門檻值越高留下來的範例樹越少，而門檻值越低會使翻譯的品質下降。



圖四、僅標記最底層



圖五、僅標記 STREE 及 STC

目前範例樹只將最底層的 n[STREE/STC/]標記完成，如圖四，現在要逐層將未標記 n[STREE/STC/]的節點標記上去。n[STREE//]標記的方式，是將 n_1 到 $n_{C(n)}$的 STREE 都加入 ES 中。ES 爲用來儲存 $n_{C(n)}$ [STREE//]中 STREE 的集合。當層數越小，則 n[STREE//]將會涵蓋 1 個以上的來源句的詞彙。若 $n_{C(n)}$ [STREE//]爲一個範圍，則將此範圍最大和最小的值加入 ES，最後 ES 內可能爲一個數字或兩個以上的數字這兩種情況，如只有一個數字則 n[STREE//]只標記該數字，如有兩個以上的數字則 ES 中最小和最大的數值標記在 n[STREE//]上，格式爲 n[最小-最大//] ；n[/STC/]標記的方式，是將 n_1 到 $n_{C(n)}$的 STC 都加入 CS 中。CS 爲用來儲存 $n_{C(n)}$ [/STC/]中 STC 的集合。當層數越小，則 n[/STC/]將會涵蓋 1 個以上目標句的詞彙。如 $n_{C(n)}$ [/STC/]爲一個範圍，則將此範圍最大和最小的值加入 CS。若 $n_{C(n)}$ [/STC/]出現 0 則不加入 CS。最後 CS 可能爲空、一個數字或兩個以上的數字這三種情況，如爲空則將 n[/STC/]標記爲 0，若只有一個數字則 n[/STC/]只

標記該數字，假如有兩個以上的數字將 CS 中最小和最大的 STC 標記在 η[/STC/]上，格式爲 η[/最小-最大/]。

假如我們現在要標記圖五第一層的節點 VP，則必需將節點 VP 的子節點 VBD 和 ADJP 的 VBD[3//]及 ADJP[4-6//]中的 STREE 加入 ES 中，因此 ES 包含了 3、4 和 6 三個數字，所以 VP[STREE//]中的 STREE 標記爲 3-6。接著標記 STC，將節點 VP 的子節點 VBD 和 ADJP 的 VBD[3/0/]及 ADJP[4-6/4-6/]中的 STC 加入 CS 中，因爲 0 不會被加入 CS 中，因此 CS 只有 4 和 6 兩個數字，所以 VP[3-6/STC/]中的 STC 標記爲 4-6。

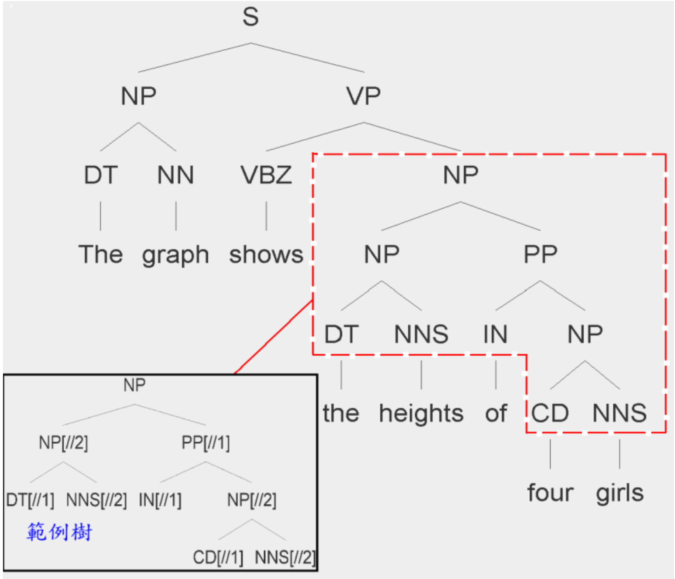
最後，整棵剖析樹的 STREE 跟 STC 都已經標記完成，如圖五，只剩下 ORDER 還沒標記上。ORDER 處理方式分爲兩部分，第一部份：STC 爲 0 之兄弟節點按照由左至右的順序編號；第二部分：比較 n 與 STC 非 0 之兄弟節點的大小，並接在第一部份的編號後，由小到大繼續標記編號。例如圖五若要標記 JJ[4/6/ORDER]和 PP[5-6/4/ORDER]的 ORDER，則將 JJ[4/STC/]中的 STC=6 和 PP[5-6/STC/]中的 STC=4 由小排到大，所以 PP[5-6/4/ORDER]中的 ORDER 標記爲 1，JJ[4/6/ ORDER] 中的 ORDER 標記爲 2。

利用上述的方法得到範例樹，如圖三。如直接用整個句子的範例樹到資料庫中作搜尋，將很難搜尋到相同的範例樹，因爲句子越長句子的結構會越複雜，所以相同結構的句子重複出現的可能很低。因此，我們將範例樹的所有子樹分別取出來，每一個子樹所包含的範圍的都是英文句的子句，在不同的句子裡可能會有相同結構的子句，不但可以增加比對到的機率，也能增加範例樹的數量。最後記錄在範例樹資料庫的內容，只有範例樹和 ORDER 參數。STREE 和 STC 不需記錄的原因是每一個句子的每個詞彙都在不同的位置上，則在資料庫中不需要記錄 STREE 和 STC。

範例樹的結構有可能相同，而詞序不同。例如“NP(NP(NN fork))(PP(IN of)(NP(DT the)(NN road)))”，中文翻譯爲“岔路”，而“NP(NP(NN leader))(PP(IN of)(NP(DT a)(NN company)))”， 中文翻譯爲“一間公司的領導者”。很明顯後者中英文用詞順序不同。這裡我們採用多數決，將出現過相同範例樹結構的每種詞序作統計，在範例樹資料庫中記錄出現最多次詞序的結構。如出現最多次的次數相同，則以隨機方式選擇一種記錄在範例樹資料庫中。最後再將範例樹資料庫中沒有詞序交換的範例樹刪除，只保留有詞序交換的範例樹，可以減少搜尋相同範例樹的時間。

3.3 搜尋相同範例樹

範例樹資料庫裡，每一筆資料都包含範例樹和範例樹的 ORDER，而範例樹就是用來當作調整詞序的參考。將輸入的英文句，先透過 StanfordLexParser-1.6[17]建立剖析樹，再將剖析樹中去掉葉子節點的結構，到範例樹資料庫去搜尋是否有相同結構的範例樹，這裡我們將所搜尋到相同的範例樹稱爲匹配子樹。如圖六所示，紅色虛線框是一棵子樹其結構爲“(NP(NP(DT)(NNS))(PP(IN)(NP(CD)(NNS))))”，方形框爲範例樹資料庫中其中一棵範例樹結構爲“(NP(NP[//2](DT[//1]) (NNS[//2])) (PP[//1](IN[//1]) (NP[//2](CD[//1]) (NNS[//2]))))”，我們可以發現範例樹去除 ORDER 後的結構，會跟子樹的結構完全相同，故將此範例樹認定爲匹配子樹。



圖六、剖析樹與範例樹的對應關係

根據搜尋範例樹演算法的流程，如圖七。首先將來源句的剖析樹加到佇列(queue)裡，從佇列裡面取出一棵剖析樹到範例樹資料庫中，搜尋是否有相同結構的範例樹；如爲否，則將此棵樹的下一層的子樹加入佇列，加入佇列的順序爲左子樹到右子樹；如爲是，則將該樹的 ORDER 標記在來源句的剖析樹上，繼續取出佇列內的剖析樹，直到佇列裡沒有剖析樹爲止。所以來源句的剖析樹是由一個以上的匹配子樹所組成。

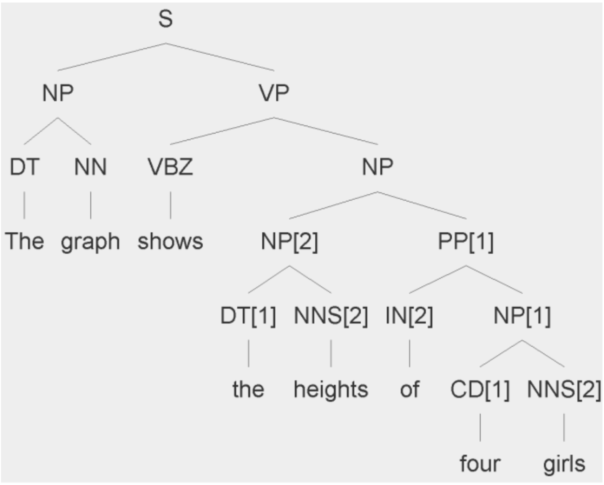
```
輸入：來源句剖析樹 S
      範例樹資料庫 D={$D_1$, $D_2$, ..., $D_m$}，$D_i \in D$
      $D_i$包含 $T_i$與 $O_i$ ，$i$爲 1 到 $m$
      $T_i$是第 $i$ 棵範例樹，$O_i$是 $T_i$所標記的詞序
開始
      設佇列 Q 用來儲存剖析樹，初始爲 NULL
      S 加入 Q
      當 Q≠NULL
          從 Q 中 pop 一棵範例樹
          如果在 D 中搜尋到相同的範例樹 $T_i$
            則將 $O_i$標記在 S 上
          否則將下一層子樹加入 Q
結束
輸出：標記好 ORDER 的剖析樹
```

圖七、搜尋範例樹演算法

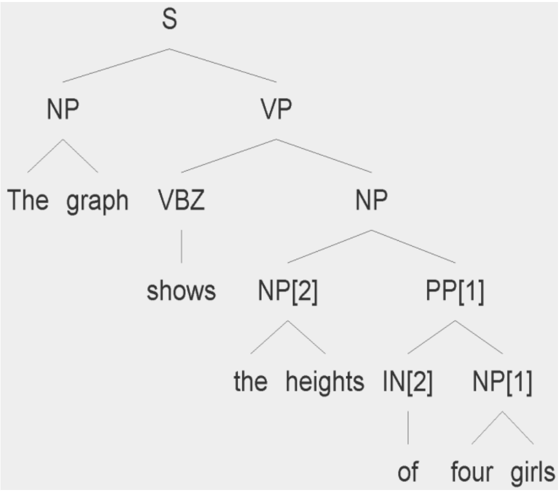
圖六爲剖析樹搜尋範例樹的情形。來源句：“The graph shows the heights of four girls”，剖析樹爲 “(S(NP(DT The)(NN graph))(VP(VBZ shows)(NP(NP(DT the)(NNS heights))(PP(IN of)(NP(CD four)(NNS girls)))))”。透過搜尋範例樹演算法找出匹配子樹，首先以節點 S 爲樹根的剖析樹到資料庫作搜尋，搜尋時不包含葉子節點，此例子沒搜尋到匹配子樹，則將節點 S 的子樹 NP 和 VP 加入佇列中。接下來將從佇列中取出的子樹爲 NP，到範例樹資料庫搜尋匹配子樹，但資料庫中沒有相同的範例樹，此時 NP 的子樹皆爲葉子節點，所以並無子樹在加入佇列中。依照先進先出的原則下一個從佇列取出的是 S 的右子樹 VP，在範例樹資料庫中還是搜尋不到，因此要將 VP 的子樹 VBZ 和 NP 加入佇列中，但 VBZ 爲葉子節點，故只有 NP 加入佇列中。接下來是子樹 NP 從佇列中被取出來，子樹 NP 在資料庫中搜尋到相同的範例樹，如圖六的範例樹就是所搜尋到的匹配子樹，因此將範例樹的 ORDER 標記上去，標記後的剖析樹將如圖八所示。此時佇列中已經爲空，搜尋範例樹的流程到此爲止。

標記完 ORDER 之後，將沒有標記的子樹作修剪，也就是將不用作詞序交換的子樹修剪到最小層樹。如圖八節點 S 的右子樹、NP[2]和 NP[1]的子樹皆不需要作詞序交換，因此修剪的結果爲“(S(NP The graph)(VP(VBZ shows)(NP(NP[2] the heights)(PP[1](IN[2] of)(NP[1] four girls))))) ”，如圖九所示。最後從層數最大的每個兄弟節點開始逐層往上依照優先權順序調整剖析樹的結構；調整後的結果將會輸入到翻譯模組產生翻譯。若我們直接取來源句剖析樹的葉子節點作翻譯，將會成爲單字式的翻譯，我們將無法對詞組或片語作翻譯。翻譯的部分會在下一節會作詳細說明。

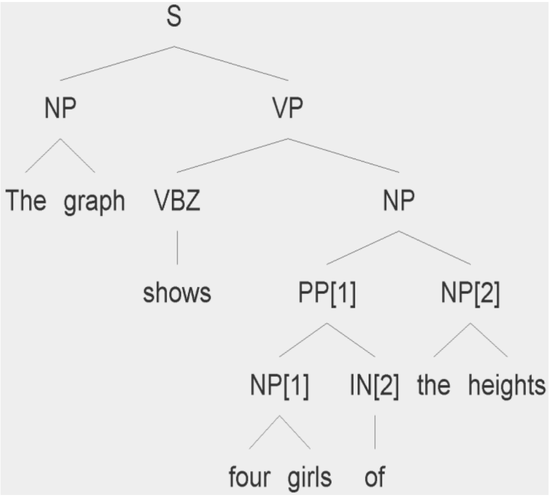
圖九的剖析樹有四層，首先將第四層的兄弟節點“(IN[2] of)(NP[1] four girls)”，依照 ORDER 的順序調整後的順序爲“(NP[1] four girls) (IN[2] of) ”，接下來第三層的兄弟節點“(NP [2] the heights)(PP[1] (NP[1] four girls)(IN[2] of))”交換後的順序爲“(PP[1] (NP[1] four girls)(IN[2] of)) (NP [2] the heights)，此例子接下來詞序沒有再調動，如圖十所示；最後輸入翻譯模組的順序爲“The graph”、“shows”、“four girls”、“of”、“the heights”，由此順序分別作翻譯處理。



圖八、完成 ORDER 標記



圖九、剖析樹修剪後的結果



圖十、調整詞序後的結果

3.4 翻譯處理

經過上一節處理最後得到修剪樹，修剪樹的葉子節點可能爲英文單字(word)、詞組(term)。詞組即爲數個單字結合的字串，不一定爲完整的句子，如“would be left on the floor”或片語(phrase，如名詞片語、動詞片語、形容詞片語等) ，如“in order to”。在翻譯處理上會遇到英文單字或詞組，在英文單字的部分，直接查尋字典檔作翻譯；詞組的部分利用規則詞典檔的片語，和詞組進行字串比對，以找出符合的片語及中文翻譯。以下爲字典檔及規則詞典檔分項說明。

字典檔：字典檔部分我們使用 Concise Oxford English Dictionary[8](牛津現代英漢雙解詞典，收錄 39429 個詞彙)，將前處理過後的英文單字或片語做翻譯對等字搜尋的動作，找出所有和該英文單字的中文詞組，作爲翻譯的候選名單。如無法在字典檔中搜尋到對應的中文翻譯。如姓名和專有名詞，則直接輸出該英文字。

規則詞典檔：爲常用的名詞片語、動詞片語、形容詞片語等詞組，以及試題翻譯小組所決議之統一翻譯詞組以人工的方式建立的中英翻譯對照檔，如 in order to(爲了)。

分成單字和詞組翻譯是因爲若在規則詞典檔比對不到，則用空白來做一般字和字之間的斷詞，也就變成單字的翻譯，因爲詞組較能完整表現出動作或敘述。如只用單字作翻譯，會造成翻譯上的錯誤。惟須注意的是比對的句型若有相似結構但不同長度的字串樣式，則取長度最長的爲結果。如一英文句子爲“…as shown in diagram…”，同時滿足規則詞典檔內的“as shown in diagram”和“in diagram”片語句型，則我們會選擇長度較長的“as shown in diagram”而不是選擇“in diagram”加上“as show”作爲斷詞的結果。

在英文翻譯成中文的過程中，有些英文單字不需要翻譯或是無意義的情形，所以我們將這些單字過濾不翻譯，這些單字稱爲 stop word。例如：冠詞 the 直接去除。介系詞 for、to、of 等，若前一單字爲 what、how、who、when、why 等疑問詞，則允以刪除，另外，to 出現在句首直接刪除。助動詞 do、does 等，判斷方式與介系詞相同。

在翻譯過程中還可能出現詞幹變化(如~ing、~ed 等)和詞性變化(如動詞 break，其過去式爲 broke，被動式爲 broken，以及名詞單複數型態)。詞幹變化的部份，我們利用 Porter[22]演算法還原各詞性(名詞、動詞、形容詞、副詞)；詞性變化的部分，有些是不規則的變化，較難用演算法處理。因此，我們透過 MXPOST[14]詞性標記工具將單字加入標記，再利用 WordNet[23]依照詞性做字典檔搜尋找到原始的型態。

3.5 統計式模組選詞

本系統將英文詞彙利用上一節介紹的翻譯方式，查詢詞典找出所有可能適合英文詞彙的翻譯結果，再利用統計式模組找出最有可能的中文詞彙，此部分已經有呂明欣等學者從事這一項研究工作[3]。以下爲我們修改後的機率模型。

$$\underset{C_{1,n}}{\operatorname{argmax}} \Pr(C_{1,n}|E_{1,n}) = \underset{C_{1,n}}{\operatorname{argmax}} \prod_{i=1}^{n} [\Pr(E_i|C_i)\Pr(C_i|C_{i-1})] \tag{1}$$

公式(1)中定義 C 爲中文翻譯詞彙，E 爲英文詞彙，$E_{1,n}$爲英文句有 1 到 n 個英文詞彙，中文翻譯詞彙也會有 1 到 n 個，即 $C_{1,n}$。從公式中可發現中文詞彙翻譯成英文詞彙的機率，稱爲中英詞彙對列，即$\Pr(E_i|C_i)$；以及利用前一個中文翻譯選詞的結果 C_{i-1}，找出目前中文翻譯詞彙 C_i共同出現的機率，稱爲 bi-gram 語言模型，即$\Pr(C_i|C_{i-1})$，將兩者相乘取計算後最大的機率值，以近似$\Pr(C_{1,n}|E_{1,n})$的機率值，作爲所選擇的中文翻譯詞彙。在選詞的過程中，$\Pr(E_i|C_i)$與$\Pr(C_i|C_{i-1})$的機率值皆有可能爲 0，我們將乘 0 換成乘上一個極小數(我們預設爲 10^{-6}) ，爲了避免機率值爲 0 的情形，會影響選詞的結果。以下將針對中英詞彙對列和 bi-gram 模型詳細介紹。

中英詞彙對列：將中英語料雙語語料，經過人工的中英語句對列(sentence alignment)技術，接著將中文語料利用中研院 CKIP 斷詞系統[1]加以斷詞；英文語料則是經過大小寫轉換及利用字和字之間空白斷詞，最後輸入至 GIZA++[16]及 mkcls[15]等工具，產生中英詞彙對列結果以及中英詞彙對照機率表。

bi-gram 語言模型：將中文語料統計各中文詞彙和下一個中文詞彙出現的次數，計算其出現機率。我們是利用 SRI Speech Technology and Research Laboratory 所開發的自然語言工具 SRILM[18]來建立 bi-gram 語言模型。

4. 系統翻譯效果評估

本節主要介紹利用本系統翻譯國際數學與科學教育成就趨勢調查 2003 年考題，簡稱 TIMSS2003，並將試題依照年齡別和科目別，分別比較翻譯的品質。最後將與線上翻譯以及呂明欣等學者研發的翻譯系統作比較。評估方式爲利用 BLEU 及 NIST 指標。

4.1 實驗來源

我們主要用來翻譯的來源爲 TIMSS2003 試題，區分數學及科學類別，並且以四年級及八年級爲考試對象，共有四種試題分別爲四年級數學領域 31 題；四年級科學領域 70 題；八年級數學領域 41 題；八年級科學領域 38 題。所有試題都有英文原文試題和師大科教中心所翻譯的中文試題。

所有實驗語料句對數、中英詞彙數、中英總詞彙個數及平均句長，皆如表一所示。用來建立範例樹的來源有教育部委託宜蘭縣建置語文學習領域國中教科書補充資料題庫[4](以下簡稱國中補充資料題庫)及科學人雜誌。國中補充資料題庫以人工方式完成中英語句對列(sentence alignment)，再經過範例樹的篩選門檻值為 0.6 的情況下有 565 句。

用來訓練選詞機率模型的來源有自由時報中英對照讀新聞及科學人雜誌。自由時報中英對照讀新聞從 2005 年 2 月 14 日至 2007 年 10 月 31 日，而自由時報中英對照讀新聞本身就已經作好中英語句對列。科學人雜誌是從 2002 年 3 月創刊號至 2006 年 12 月共 110 篇為語料來源。

表一、實驗語料來源統計

語料	語言	句對數	辭彙數	總詞彙個數(tokens)	平均句長
國中補充資料題庫	中文	2059 句	2333	12460	6.1
	英文		2887	13170	6.4
科學人	中文	4247 句	9279	70411	16.6
	英文		10504	68434	16.1
自由時報中英對照讀新聞	中文	4248 句	19188	145336	34.2
	英文		25782	133123	31.3

4.2 實驗設計

首先，將 TIMSS2003 試題問句以逗號、問號或驚嘆號做為斷句的單位，每個誘答選項做為斷句的單位，若一道題目為一句試題問句及四項誘答選項所組成，則一道題目可斷出五句。經過人工斷句處理 TIMSS2003 試題，四年級數學領域有 165 句；四年級科學領域有 262 句；八年級數學領域有 439 句；八年級科學領域有 236 句，並整理為文字檔。翻譯時中文試題所運用的中文斷詞為中研院 CKIP 斷詞系統[1]，英文試題所運用的剖析器為 StanfordLexParser-1.6[17]，建立範例樹資料庫所使用的語料為國中補充資料題庫，訓練機率模型所使用的語料自由時報中英對照讀新聞加上科學人雜誌，其中訓練語言模型得到的 bi-gram 共有 134435 個；GIZA++產生中英詞彙對列結果有 128551 組。

表二、TIMSS 試題實驗組別表[†]

八年級 2003 M組	八年級 2003 S 組	四年級 2003 M組	四年級 2003 S組	八年級 2003 MS組	四年級 2003 MS組
TIMSS2003 國中數學領域試題	TIMSS2003 國中科學領域試題	TIMSS2003 國小數學領域試題	TIMSS2003 國小科學領域試題	TIMSS2003 國中數學及科學領域試題	TIMSS200 國小數學及科學領域試題

我們評估所使用的工具為依照 BLEU 及 NIST 標準的 mteval-10，並且我們將參考的中文標準翻譯和系統建議翻譯，每個中文字跟中文字之間用空白作分隔，計算出各別 n-gram 及累加各個 n-gram 的 BLEU 及 NIST 值。主要評估的對象有 Google 線上翻譯、Yahoo!線上翻譯、呂明欣學者的系統(Lu)及本系統互相做比較，並且評估翻譯系統在不同年級的試題內容上，翻譯品質是否會按照越低年級其翻譯品質越好的趨勢。因此，我們將實驗組別分為八年級和四年級；數學領域以 M 為代號；科學領域以 S 為代號，當

[†] 本篇論文 TIMSS 試題實驗組，僅包含 2003 年試題，與呂明欣學者的實驗組並不相同。

作實驗組別的名稱。可以 TIMSS2003 分為八年級 2003 M 組、八年級 2003 S 組、四年級 2003 M 組及以四年級 2003 S 組四組；在加上 TIMSS 2003 數學及科學領域之八年級試題，和 TIMSS 2003 數學及科學領域之四年級試題，分別為八年級 2003 MS 組及四年級 2003 MS 組，總共六組，如表二所示。

4.3 實驗結果

依照上一節的實驗設計，我們針對 TIMSS2003 試題驗證本系統、Lu 系統及線上翻譯系統在 BLEU 和 NIST 比較數據。從表三是以 cumulative n-gram scoring 之 4-gram 為平均值，整理之各組 NIST 及 BLEU 值之比較表。NIST 跟 BLEU 最大的不同在於，NIST 將各 n-gram 詞彙中共現（co-occurrence）的次數比的累加值，當作各 n-gram 平均資訊量的大小，而 BLEU 針對各 n-gram 匹配正確率及相似度進行計分。由此可知當參考翻譯句子和系統翻譯句子用的詞彙相同時，NIST 分數會比較高；當參考翻譯句子和系統翻譯句子用的詞彙順序較相近時，BLEU 分數會比較高。

表三、本系統、Lu 系統及線上翻譯系統之 NIST 及 BLEU 值比較表

組別	八年級 2003 M 組		八年級 2003 S 組		四年級 2003 M 組	
指標	NIST	BLEU	NIST	BLEU	NIST	BLEU
本系統	4.7002	0.1440	4.4089	0.1254	3.9819	0.1304
Lu	3.6185	0.1007	3.5831	0.0890	3.3319	0.0983
Google	4.5268	0.1467	4.8587	0.1848	3.7573	0.1016
Yahoo!	4.8793	0.1455	4.6136	0.1396	4.0457	0.1419
組別	四年級 2003 S 組		八年級 2003 MS 組		四年級 2003 MS 組	
指標	NIST	BLEU	NIST	BLEU	NIST	BLEU
本系統	4.2228	0.1018	4.8613	0.1309	4.4400	0.1138
Lu	3.2495	0.0682	3.8031	0.0966	3.4970	0.0803
Google	4.4445	0.1527	4.9343	0.1611	4.4720	0.1344
Yahoo!	4.4361	0.1442	5.0755	0.1435	4.6070	0.1436

從表三可觀察到，八年級 2003 M 組 NIST 分數以 Yahoo!最高分，但 BLEU 分數與本系統差不多，可知 Yahoo!對八年級 2003 M 組所翻譯的詞彙跟參考翻譯較相同，但 Yahoo!和本系統翻譯後詞序的正確性是差不多的。四年級 2003 M 組試題中有較多特殊符號，例如○和●等，Yahoo!及 Google 線上翻譯系統會將這些特殊符號處理成亂碼，但本系統可以將特殊符號保留下來，故四年級和八年級 2003 M 組與最高分系統的差距較小。先前我們假設翻譯品質是否會按照越低年級其翻譯品質越好的趨勢，觀察八年級 2003MS 組及小四 MS 組，可發現與假設相反，各系統在八年級 2003 MS 組的表現都比四年級 2003 MS 組要好。可推測出本系統其中一種語料為國中補充資料題庫較符合 TIMSS 八年級 2003 的試題。

我們將八年級 2003M 組和八年級 2003S 組作比較，四年級 2003 M 組和四年級 2003 S 組作比較，可以發現各系統除了 Google 之外，在 M 組上表現都比 S 組好，因為 M 組的試題內容包含較多的數字，對於翻譯系統較容易處理，而 S 組則包含較多專有名詞，對於翻譯系統較為困難。接著將本系統與 Lu 系統作比較，Lu 系統和本系統的差別為沒有作詞序的交換。經過詞序交換後，得到正確的中文詞序，因此選詞的正確性相對會提升，所以本系統在各組的表現都比 Lu 系統要好，顯示詞序交換後會得到品質較好的中文翻譯。

5. 結論

本論文提出 BSSTC 結構，此結構能夠記錄來源句詞彙的位置、目標句詞彙的位置及來源句與目標句詞彙對應的關係；並且將 BSSTC 結構運用在我們實作的翻譯系統上。本系統是利用 BSSTC 結構建立範例樹，將來源句經過搜尋範例樹演算法，來達到修正詞序的目的。最後，在依據修正後的詞序進行翻譯，翻譯時再利用中英詞彙對列工具及 bi-gram 語言模型，選出最適合的中文翻譯，產生建議的翻譯，此翻譯還需要人工編修。

TIMSS 的試題爲數學及科學類，應該要用大量數學及科學類的語料，但實際上我們並無法找到夠多的數學及科學類語料，尤其以中英對應的語料最少，所以我們選用新聞及國中補充資料題庫來擬補語料的不足。不過訓練量還算是不足夠，在選詞上會有許多機率爲 0 的情況，造成選詞錯誤。未來將盡量找尋相關領域的語料，來建立範例樹和訓練語言模型，就能針對不同領域的來客製化翻譯，使翻譯的結果更爲精確。

訓練語料中的斷詞是使用中研院 CKIP 系統，而我們翻譯使用的字典爲牛津字典，兩者所使用的字典並不相同，會使斷詞後的詞彙可能無法在牛津字典中找到，造成選詞錯誤。未來可將翻譯後的詞彙，找出同義詞來擴充詞彙數，便能增加被找到的可能性。

英文的語言特性上並沒有量詞，而中文句中運用了很多的量詞，如缺少量詞也會使中文的流暢度下將。本系統的翻譯結果也缺少中文的量詞。未來若能將翻譯結果填補上缺少的量詞，便可達到更好的品質。

致謝

本研究承蒙國科會研究計畫 NSC-95-2221-E-004-013-MY2 的部分補助謹此致謝。我們感謝匿名評審對於本文初稿的各項指正與指導，雖然我們已經在從事相關的部分研究議題，不過限於篇幅因此不能在本文中全面交代相關細節。

參考文獻

[1] 中研院中文剖析器檢索系統, http://parser.iis.sinica.edu.tw/ [Accessed: Jun. 30, 2008].

[2] 自由時報中英對照讀新聞, http://www.libertytimes.com.tw/2008/new/jan/15/english.htm [Accessed: Jun. 30, 2008].

[3] 呂明欣，*電腦輔助試題翻譯：以國際數學與科學教育成就趨勢調查爲例*，國立政治大學資訊科學學所，碩士論文， 2007。

[4] 教育部委託宜蘭縣發展九年一貫課程件至語文學習領域（英語）國中教科書補充資料暨題庫建置計畫, http://140.111.66.37/english/ [Accessed: Jun. 30, 2008].

[5] M. H. Al-Adhaileh, T. E. Kong and Y. Zaharin, “A synchronization structure of SSTC and its applications in machine translation”, *Proceedings of the International Conference on Computational Linguistics -2002 Post-Conference Workshop on Machine Translation in Asia*, 1–8, 2002.

[6] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer and P. S. Roossin, “A Statistical Approach to Machine Translation”, *Computational Linguistics*, 79-85, 1990.

[7] C. Boitet and Y. Zaharin, “Representation trees and string-tree correspondences”, *Pro-*

ceedings of the Twelfth International Conference on Computational Linguistics, 59–64, 1998.

[8] Concise Oxford English Dictionary, http://stardict.sourceforge.net/Dictionaries_zh_TW.php [Accessed: Jun. 30, 2008].

[9] B. J. Dorr, P. W. Jordan and J. W. Benoit, "A Survey of Current Paradigms in Machine Translation" *Advances in Computers*, London: Academic Press, 1-68, 1999.

[10] Google Translate http://www.google.com/translate_t [Accessed: Jun. 30, 2008].

[11] K. Knight and S. K. Luk, "Building a large-scale knowledge base for machine translation", *Proceedings of the Twelfth National Conference on Artificial intelligence*, 773-778, 1994.

[12] P. Koehn, F. J. Och and D. Marcu, "Statistical phrase-based translation", *Proceedings of the Human Language Technology Conference,* 127–133, 2003.

[13] Z. Liu, H. Wang and H. Wu, "Example-based Machine Translation Based on TSC and Statistical Generation", *Proceedings of the Tenth Machine Translation Summit*, 25–32, 2005.

[14] MXPOST, http://www.inf.ed.ac.uk/resources/nlp/local_doc/MXPOST.html [Accessed: Jun. 30, 2008].

[15] F. J. Och, "An Efficient Method for Determining Bilingual Word Classes", *Proceedings of European Chapter of the Association for Computational Linguistics*, 71–76, 1999.

[16] F. J. Och and H. Ney, "Improved Statistical Alignment Models", *Proceedings of the Thirty-eighth Annual Meeting of the Association for Computational Linguistics*, 440–447, 2000.

[17] The Stanford Parser: A statistical parser, http://nlp.stanford.edu/software/ [Accessed: Jun. 30, 2008].

[18] A. Stolcke. SRILM – an extensible language modeling toolkit. *Proceedings of the intelligence Conference on Spoken Language Processing*, 901–904, 2002. http://www.speech.sri.com/projects/srilm/ [Accessed: Jun. 30, 2008].

[19] S. Sato and M. Nagao, Toward Memory-Based Translation", *Proceedings of International Conference on Computational Linguistics*, 247–252, 1990.

[20] The International Association for the Evaluation of Education Achievement, http://www.iea.nl/ [Accessed: Jun. 30, 2008].

[21] TIMSS 中文版官方網頁, http://timss.sec.ntnu.edu.tw/timss2007/news.asp [Accessed: Jun. 30, 2008].

[22] The Porter Stemming Algorithm, http://www.tartarus.org/martin/PorterStemmer/ [Accessed: Jun. 30, 2008].

[23] WordNet API, http://nlp.stanford.edu/nlp/javadoc/wn/ [Accessed: Jun. 30, 2008].

[24] F. Wong, M. Dong and D. Hu, Machine Translation Based on Translation Corresponding Tree Structure, *Tsinghua Science & Technology*, 25–31, 2006.

[25] YAHOO! 雅虎線上翻譯, http://tw.search.yahoo.com/language/ [Accessed: Jun. 30, 2008].

電腦輔助推薦學術會議論文評審委員之初探

陳禹勳　　　　劉昭麟
國立政治大學　資訊科學系
{g9418,chaolin}@cs.nccu.edu.tw

摘要

會議論文評審委員由會議議程主席指派，目的在分配適當且數量平均的論文給評審委員，以求審核論文的公正性與正確性。本研究以系統化的方法讓機器輔助人工，達到避免個人的主觀因素及節省人力的目標，並利用文件分類技術以及 Google 學術搜尋提供的資訊，建構協助議程主席指派論文的環境。我們依照一般學術會議論文的小節結構，將論文切成數個區段，藉由整合論文不同區段的特性，期望得到一個較佳的指派結果。

關鍵詞：文件分類、向量空間模型，社群網路

1. 緒論

投稿學術研討會的論文審核時程，以國內會議人工智慧與應用研討會[1](Taiwan Association for Artificial Intelligence)爲例，由 2004、2006 及 2007 這三年的研討會網頁得知，從截稿日期至通知接受日期，大約需要一個月以上的時間。主要考量在於議程主席指派待審論文給評審委員，以及評審委員研讀待審論文所花的時間。指派論文給評審委員，需要知道評審委員的研究領域與待審論文的研究領域是否相近。由於論文評審委員的領域有所不同，甚至有跨領域的研究，因此待審論文對評審委員的分配不容易決定。通常議程主席對於各教授的領域只有大略的了解，指派評審是從該教授的著作來決定，因此在面對不熟領域的教授著作時，常需要花費大量的時間與精力。加上各領域教授人數眾多，在眾多的議程委員中選取論文評審委員變得窒礙難行。

Peterson[15]的研究指出，由於閱讀論文相當費時，因此研究生及學者閱讀論文時通常不是看全篇論文，而是挑出摘要、簡介、結論及參考文獻區段來看。摘要區段透露比較多論文主題及應用技術的訊息；簡介區段則是大致說明此論文的研究動機、研究背景以及架構流程；結論區段敘述此研究的研究成果，由實驗結果印證研究方法並提出相關研究方向；參考文獻區段提供一個相關領域的查詢。因此本研究認爲對論文的指派，可以細分成各區段的相似度比對，再將其結果整合，使得建議評審委員的正確性較高。

引用共同的參考文獻強烈暗示著領域相近。各種不同領域的論文，不容易引用到同一篇論文。參考文獻區段的相似度比對上，我們採取參考文獻標題以及參考文獻作者比對作爲相似度的考量。參考文獻的部分含有許多的資訊，包含作者、論文名稱、出處及年份。我們應用 Google 學術搜尋及正規表示式取出參考文獻的標題以及作者，藉由找出待審論文及各評審委員著作的共同引用參考文獻數或作者數，作爲參考文獻區段建議評審委員的根據。對於摘要、簡介及結論區段，本研究採用向量空間模型來做相似度比對。向量空間模型[17](Vector Space Model)是文件分類的重要技術，我們希望應用文件分類的技術，來輔助議程主席指派論文。

文件分類是根據文件內容或主題給定類別的工作，以往文件分類的研究，都是對整篇文件去取出特徵，接著藉由某些分類方法去作分類。文件分類的特徵大多是找出關鍵詞，也就是這篇文章中具有鑒別度的詞。顧皓光等[8]在 1997 年提出網路文件自動分類的方法，採用向量空間模型去對 Yahoo 內部資料庫的網頁進行分類。由於網路文件資料量相當的大，在大量資料的情形下，向量空間模型可以分出相當不錯的結果。

然而在資訊不夠充足的情形下，向量空間模型分類的效果會變的非常的差，錢炳全等[7]在 2002 年提出中文試題自動分類方法，試圖對簡短的試題作分類。在系統自動學習試題的情形下，資訊量越來越多，而分類效果也隨之改善。駱思安等[6]則是在 2006 年提出一個以機率爲主的中文網站分類系統，此系統可自動學習詞彙來改善分類效果。Dow 等[10]在 2007 年利用 DSpace[11]建立了一個論文的查詢網站，不但可從查詢的關鍵詞推薦相關的論文，並提供與查詢的關鍵詞相關的關鍵詞、領域及相關的教授，同時也提供各教授論文領域的分布，使用者可以更容易找出要查詢的資料。

專利文件的自動分類是向量空間模型在文件分類上的另一種應用。爲了避免侵犯智慧財產權，專利文件寫法上較爲格式化且嚴謹，也因此專利文件的篇幅通常相當巨大。專利文件通常分成數個段落，分別是標題、摘要、專利權利範圍、專利技術描述以及總結。Larkey 等[18]建立一個專利文獻的查詢與分類系統，藉由抽取出不同段落及計算詞彙的重要性來分類專利文件。李駿翔等[4]則是嘗試著將標題跟不同段落的分類結果整合，發現標題結合總結與標題結合專利技術描述的分類效果最好。林蘭綺等[5]則是應用標題加上總結段落部分，利用詞彙的不同權重來提高分類效果。

本研究介紹順序如下: 第二節描述系統架構、第三節說明研究方法、第四節爲實驗結果以及第五節爲結論。

2. 系統架構

此節介紹本研究的整體流程以及所需要的資料及來源出處。

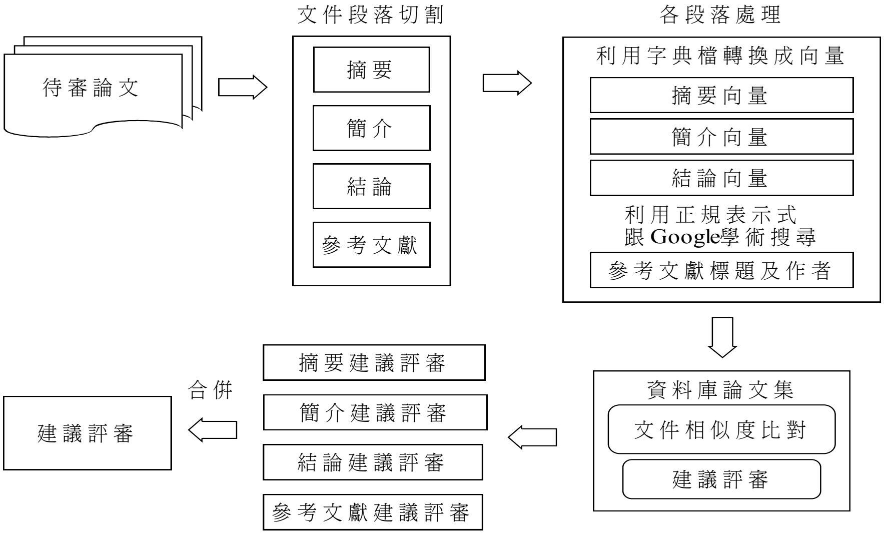


圖 1、系統架構流程圖

2.1 系統流程

本研究推薦中文論文評審委員，研究流程如圖 1 所示。將待審論文切成各個文件區段，使用向量空間模型等方法，進行待審論文各區段與資料庫論文集各區段的相似度比對。再藉由最相似論文來找出該區段的評審委員，最後整合各區段建議評審委員結果，得出待審論文的建議論文評審委員。

2.2 資料來源

本研究爲了處理的方便，論文一律從 PDF 檔轉爲文字檔，檔案轉換工具是使用 Acrobat Professional 版裡的批次處理功能來進行轉換，從 PDF 檔轉爲文字檔的成功率約略爲 74.85%。這些論文的資料來源，除了從網路上下載，還有選自於人工智慧與應用研討會 2002 年、2003 年、2004 年及 2005 年論文集的論文以及全國計算機會議 (National Computer Symposium) 2001 年、2003 年及 2005 年論文集論文共 1089 篇。測試資料則是選自 2007 年的人工智慧與應用研討會論文集，共 71 篇論文。

中文文件分詞的部分，本研究採取使用字典檔分詞的作法，以 HowNet[13]辭典作爲基礎來處理中文分詞。由於 HowNet 辭典是收納一般生活常用的中文詞彙，未必能對論文作精確的分詞，因此我們從九二八電腦股份有限公司[2]的網站，收集了兩岸三地較常見的電腦詞彙字庫，刪除重複詞，分別加入到現有詞庫中。詞庫共有總數量五萬一千多個詞，我們發現五萬一千多個詞中，只有八千多個詞彙出現在訓練資料論文過。因此，我們將沒出現過的四萬多個詞彙刪除，對剩下這八千多個詞彙依照詞的長度作分類，分成二字詞、三字詞與四字詞等，建立出一個較精簡的字典檔作爲中文分詞的依據。

3. 研究方法

本節描述處理論文區段的方法及流程。一般來說論文可分成數個區段，分別是摘要、簡介、研究方法、實驗結果及結論等等。研究方法與實驗結果區段描述研究過程，用詞以解釋清楚爲目的，站在文件分類關鍵詞爲特徵的角度來看，文件關鍵詞應具有代表性而非只是詞頻高，而這兩個區段的詞彙多爲描述研究過程，作爲關鍵詞較爲不適當。摘要、簡介及結論等區段常精簡的描述研究，很有機會出現重要的關鍵字。因此不同於一般文件分類研究以一篇文章作爲分類的基本單位，本研究把論文的各個區段切出，分別是摘要、簡介、結論和參考文獻區段，藉由整合各區段的相似度比對結果來改善分類效果。

3.1 取出論文區段的方法

一篇論文的各個區段往往都有特別的詞作爲開頭，因此本研究利用每段區段的開頭詞來做分區段的依據。我們採取一列列讀取每篇論文文件的做法，以便找出各區段的開頭詞。由於從 PDF 檔轉爲文字檔的成功率約略爲七成多，轉檔時可能會有文字的錯誤，因此會有區段取出不完整的情形。不同的論文會有不同的區段開頭詞敘述法，因此我們建立一個區段開頭詞的相關用語表，如表 1 所示。

摘要區段通常位於文章前段，以「摘要相關用語」爲摘要區段的開頭詞，而摘要區段後面通常是接關鍵字段落，因此取「關鍵字相關用語」爲摘要區段結尾。本研究取以「摘要相關用語」作爲開頭的一行到以「關鍵字相關用語」作爲開頭的一行之間這段文字作爲摘要區段。

表 1、各區段開頭詞相關用語表

摘要	摘要
關鍵字	關鍵字、關鍵詞
簡介	緒論、概論、簡介與相關研究、前言及研究背景、前言、背景動機、序論、簡介、研究背景與動機、研究動機與目的、研究動機、引言、背景與理論基礎、研究背景、介紹、導論、背景、緒言、緣由與目的
結論	結論、結語、討論、啓示、建議、未來發展方向、未來發展、未來研究方向、未來研究、未來工作、未來展望、未來後續工作、後續研究建議、後續研究、研究成果
參考文獻	參考文獻、參考資料
系統架構	系統架構、系統運作流程、設計架構、系統架構與方法、系統架構與規劃
相關研究	相關研究、相關文獻、文獻探討、理論背景與文獻探討、研究目的、背景與相關研究、相關文獻探討、背景知識與相關研究、相關研究背景說明、相關工作、相關文獻研究
研究方法	研究方法

簡介區段多位於摘要和關鍵字區段後面，簡介區段便以「簡介相關用語」作爲開頭，簡介的結尾卻是難以認定，我們觀察數篇論文的簡介開頭詞，發現一般論文中簡介開頭詞的寫作方式可大致分爲兩類：

- 用數字、英文字或羅馬符號對開頭詞標號
- 無任何標號

對於有標號的簡介開頭詞，我們建立一個對應表去對應標號跟數字間的關係，如此可得知簡介區段是標號在第幾段落，再推出簡介區段的下一區段是標號在第幾段落，便可找到簡介區段的結尾詞，進而切出簡介區段。表 2 是標號跟數字的對應表。

表 2、標號數字對應表

阿拉伯數字標號	1	2	3	4	5	6	7	8	9
中文數字標號	一	二	三	四	五	六	七	八	九
中文國字大寫標號	壹	貳	叁	肆	伍	陸	柒	捌	玖
羅馬標號	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ	Ⅶ	Ⅷ	Ⅸ

若今天簡介開頭詞標號是Ⅰ，那麼下一段落的開頭詞標號就會是Ⅱ，可由此開頭詞標號切出簡介段落。其他開頭詞標號作法亦同。若簡介開頭詞無標號，本研究觀察簡介區段的下一區段通常是系統架構、相關研究或研究方法區段，因此藉由這三個區段的開頭相關用語來找出簡介區段的結尾，進而取出簡介區段。

結論區段多位於論文的後段，後面通常是接參考文獻區段，因此取以「結論相關用語」爲開頭詞的一行到以「參考文獻相關用語」爲開頭詞的一行之間的段落作爲結論區段。參考文獻區段則是取以「參考文獻相關用語」爲開頭詞的一行到文章結尾的段落。

3.2 摘要、簡介及結論區段處理

在論文切出的區段之中，由於參考文獻區段可細分出參考文獻作者及參考文獻標題，因此參考文獻區段我們額外處理，其他區段則一致使用向量空間模型做相似度比對。

要將文章轉成向量，首先要將所有的文章去做分詞的動作，我們便可以得知各個詞彙在每篇文章中出現的次數，再利用資料檢索的 tf-idf(term frequency - inverted document frequency)[16]技術，計算出每個詞彙的 tf-idf 值。tf-idf 的計算法是資訊檢索以及文件探勘等相關領域中相當重要的公式，是由每個詞彙的 tf 值（term frequency）和 idf 值（inverse document frequency）所相乘所得出的一個常數。其中 tf 爲詞彙在單一文件中的出現頻率，可視爲在該文件內部的分布特性；idf 則是用來量測詞彙在所有文件中的重要程度，可視爲全域資料的分布特性。

$$idf_i = \log(N / n_i) \tag{1}$$

其中 N 爲論文訓練資料的總篇數，i 代表詞彙，n_i 則是包含詞彙 i 的論文總數，由公式(1)得知當一個詞彙 idf 值越小時，表示該詞彙在絕大部分的文件都有出現，因此鑑別度就會很低。一個詞彙 i 的 tf-idf 算法如公式(2)所示。

$$tfidf_i = tf_i \times idf_i \tag{2}$$

我們將每篇文章分詞之後的詞彙分別去做各自的 tf-idf，將詞彙當作向量的一個屬性，tf-idf 的值則作爲屬性裡面的值，文件就可以向量的方式去表示。

3.3 參考文獻區段處理

本研究特別針對人工智慧與應用研討會，及全國計算機會議的論文集來處理參考文獻區段。發現這些論文的參考文獻格式，大多是用數字條列式標示各筆參考文獻，因此可以將參考文獻區段細分成一筆筆的參考文獻。一筆參考文獻，我們取出它的作者及論文標題。至於論文出處，由於期刊及會議數量眾多且規模大小不一；出版年份無法直接反映論文的領域，兩者皆不容易定義分類相關性，因此本研究目前先不去對其做處理。

3.3.1 論文標題的擷取

以人工智慧與應用研討會跟全國計算機會議論文集來說，一篇論文的參考文獻，論文名稱的寫法大致分爲兩種，一種是以引號框住論文標題，用來強調論文標題；另一種則是沒有引號標示出論文標題。第一種寫法的論文標題較易取出，只需取出每篇參考文獻裡引號內的文字。由於第二種寫法無法判定論文標題，我們利用了 Google 學術搜尋網頁[14]作爲找出論文標題的工具。

Google 學術搜尋網頁是 Google 的一個學術文獻資源搜尋引擎，Google 學術搜尋網頁會依關連性排序搜尋結果，也就是考量文章的內文、作者、文章所在出版物以及內容片段出現在其他學術文獻出現的頻率。雖然我們並不知道 Google 學術搜尋網頁如何去排序這些資料，但是以 Google 的強大功能，我們認爲他的技術可信度很高的。使用 Google 學術搜尋網頁搜尋論文時，有一個特別的地方，就是當作者跟論文標題一起查詢時，查詢結果會用論文標題回傳連結。我們撰寫了一個程式，能夠送字串給 Google 學術搜尋網頁，並抓回搜尋的結果，用連結來驗證是否是參考文獻的論文標題。一筆參考文獻的寫法，是由左至右依序先寫作者，再寫論文標題，因此我們將參考文獻切成數個參考文獻片段，用圖 2 所示的演算法得出參考文獻標題。

範例：利用 Google 學術搜尋網頁找出標題

J. Setubal and J. Meidanis, Introduction to Computational Molecular Biology, PWS, Boston, MA, 1997.

如圖 2 演算法所示，這一筆參考文獻會被切成數個參考文獻片段

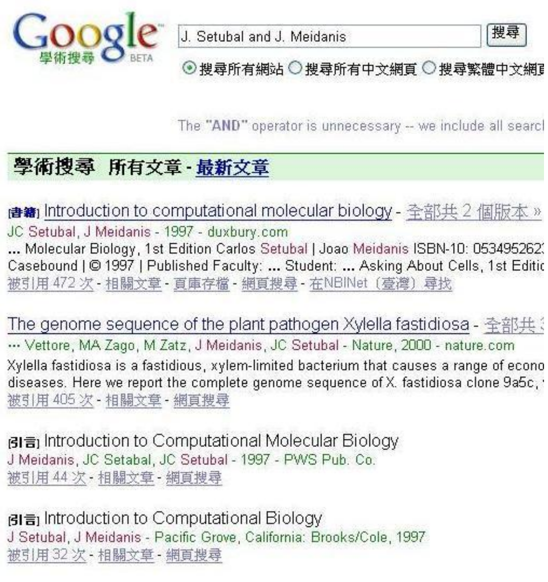
J. Setubal and J. Meidanis
Introduction to Computational Molecular Biology
PWS
Boston
MA
1997

我們將第一個參考文獻片段送給 Google 學術搜尋網頁，其結果如圖 3。

```
輸入  ：一筆參考文獻 R
輸出  ：參考文獻的論文標題 T

步驟 1：將 R 照逗號切開，成爲由數個參考文獻片段 $k_j$ 組成的集合 K，K={$k_1$,$k_2$,…$k_m$} 爲參
        考文獻片段集合，m 爲參考文獻片段個數
步驟 2：令 F 爲丟字串到 Google 學術搜尋網頁的程式，L 爲 Google 學術搜尋的前十名結果，
        L=F($k_j$)，將切開後的參考文獻片段依序丟到 Google 學術搜尋網頁，並回傳搜尋到的
        前十筆資料 $\forall k_j \in K, L_j = F(k_j)$
步驟 3：比對搜尋的結果
    步驟 3.1：如果 $L_j$ 是 $k_j$ 字串的一部分，表示 $L_j$ 爲論文標題，回傳 T=$L_j$
    步驟 3.2：如果沒有相同比對的字串
              若 j 不等於 m，則設字串 $k_{j+1}$ 爲字串 $k_j$ 跟字串 $k_{j+1}$ 相連的結果，回到步驟 2
              若 j 等於 m，表示找不到標題，結束
```

圖 2、利用 Google 學術搜尋網頁找出標題的演算法



J. Setubal and J. Meidanis

◉ 搜尋所有網站 ○ 搜尋所有中文網頁 ○ 搜尋繁體中文網頁

The "AND" operator is unnecessary -- we include all searcl

學術搜尋 所有文章 - 最新文章

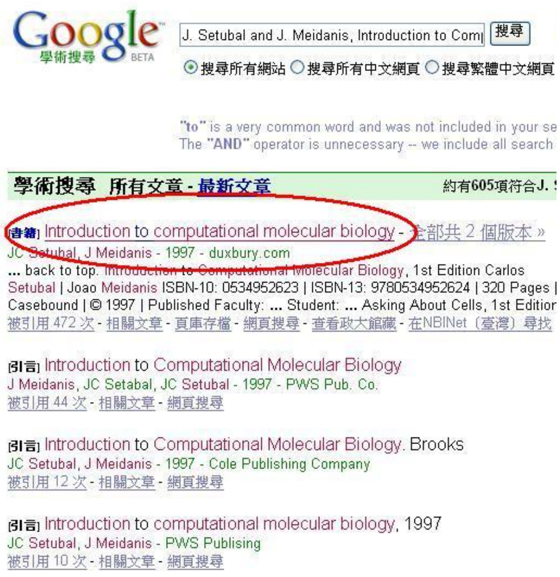
[書籍] Introduction to computational molecular biology - 全部共 2 個版本 »
JC Setubal, J Meidanis - 1997 - duxbury.com
... Molecular Biology, 1st Edition Carlos Setubal | Joao Meidanis ISBN-10: 053495262
Casebound | © 1997 | Published Faculty: ... Student: ... Asking About Cells, 1st Editic
被引用 472 次 - 相關文章 - 頁庫存檔 - 網頁搜尋 - 在NBINet（臺灣）尋找

The genome sequence of the plant pathogen Xylella fastidiosa - 全部共
... Vettore, MA Zago, M Zatz, J Meidanis, JC Setubal - Nature, 2000 - nature.com
Xylella fastidiosa is a fastidious, xylem-limited bacterium that causes a range of econo
diseases. Here we report the complete genome sequence of X. fastidiosa clone 9a5c,
被引用 405 次 - 相關文章 - 網頁搜尋

[引言] Introduction to Computational Molecular Biology
J Meidanis, JC Setabal, JC Setubal - 1997 - PWS Pub. Co.
被引用 44 次 - 相關文章 - 網頁搜尋

[引言] Introduction to Computational Biology
J Setubal, J Meidanis - Pacific Grove, California: Brooks/Cole, 1997
被引用 32 次 - 相關文章 - 網頁搜尋

圖 3、Google 學術搜尋網頁－字串搜尋結果 I



J. Setubal and J. Meidanis, Introduction to Com

◉ 搜尋所有網站 ○ 搜尋所有中文網頁 ○ 搜尋繁體中文網頁

"to" is a very common word and was not included in your se
The "AND" operator is unnecessary -- we include all search

學術搜尋 所有文章 - 最新文章 約有605項符合J.

[書籍] Introduction to computational molecular biology - 全部共 2 個版本 »
JC Setubal, J Meidanis - 1997 - duxbury.com
... back to top. Introduction to Computational Molecular Biology, 1st Edition Carlos
Setubal | Joao Meidanis ISBN-10: 0534952623 | ISBN-13: 9780534952624 | 320 Pages |
Casebound | © 1997 | Published Faculty: ... Student: ... Asking About Cells, 1st Edition
被引用 472 次 - 相關文章 - 頁庫存檔 - 網頁搜尋 - 查看政大館藏 - 在NBINet（臺灣）尋找

[引言] Introduction to Computational Molecular Biology
J Meidanis, JC Setabal, JC Setubal - 1997 - PWS Pub. Co.
被引用 44 次 - 相關文章 - 網頁搜尋

[引言] Introduction to Computational Molecular Biology. Brooks
JC Setubal, J Meidanis - 1997 - Cole Publishing Company
被引用 12 次 - 相關文章 - 網頁搜尋

[引言] Introduction to computational molecular biology, 1997
JC Setubal, J Meidanis - PWS Publising
被引用 10 次 - 相關文章 - 網頁搜尋

圖 4、Google 學術搜尋網頁－字串搜尋結果 II

搜尋結果沒有與 *J. Setubal and J. Meidanis* 相符的字串，於是將 *J. Setubal and J. Meidanis* 跟 *Introduction to Computational Molecular Biology* 兩字串相連，送給 Google 學術搜尋網頁，其結果如圖 4。發現搜尋結果其中之一爲 *J. Setubal and J. Meidanis,* *Introduction to Computational Molecular Biology* 的子字串，因此回傳此搜尋結果爲參考文獻的標題，或是著作的書名。

3.3.2 論文作者的擷取

以一筆參考文獻來說，已經成功取出論文標題之後，在論文標題之前的部分就是作者群，之後的部份就是出處及年份。因此我們把參考文獻在論文標題之前的段落取出，作爲取出作者的根據。參考文獻的作者姓名可分中文和英文兩種。中文姓名的部分，李振昌[3]提出一套有效的人名識別規則，除了以中文姓氏來辨識外，還加入了性別常用字以及前後文變異性等來斷定是否爲人名。由於參考文獻爲簡短的文字段落，從文字能獲得的資訊很少，因此本研究只使用百家姓姓氏比對的方式去找出人名。演算法如圖 5。

```
輸入：一筆參考文獻 R
輸出：作者名字集合 $A=\{A_1,A_2,\ldots A_\gamma\}$，γ 爲一筆參考文獻的中文作者數

步驟 1：由前一節演算法找出的論文標題, 將參考文獻在論文標題 T 之前的文字段落 S 取出，
        S=R-T-(T 之後的段落)
步驟 2：將此文字段落 S 依標點符號切開，得出 $B=\{B_1,B_2,\ldots B_o\}$，B 爲可能是作者的人名集合，
        o 爲此參考文獻切成段落的段落總數
步驟 3：$\forall B_j \in B$, 檢查 $B_j$ 的第一個字元是否是百家姓
    步驟 3.1: 若是, 表示可能爲作者名稱，先檢查是否有「和」、「與」等中文連接詞，如果有則
              將 $B_j$ 依該連接詞切開，並將 $B_j$ 切開後的兩段落加入 A 中，若無連接詞，則直接加
              入 $B_j$ 到 A 中
    步驟 3.2: 若否, 則可能抓錯，結束
```

圖 5、參考文獻找出中文作者的演算法

英文姓名部分，由於參考文獻的作者群寫法有一定的格式，單獨大寫字母往往表示英文名字縮寫，本研究觀察國內論文集常出現的參考文獻作者寫法，用正規表示法(Regular Expression)定義作者姓名的樣式(patterns)。表 3 是我們觀察幾個常見參考文獻作者的寫法。

表 3、常見英文姓名的寫法範例

T. Nishita 或 C. C. Liu	由至少一組一個英文字母加一個英文句號與一個空白的字串，加上英文字母字串組成。
Beckmann, N.或 Robinson, J. T	由英文字母字串，一個逗號加空白，一個英文字母，一個英文句號加空白，加上零個以上的英文字母組成。
John H. Holland 或 David Andre	由英文字母字串，空白，至少一組一個英文字母加一個英文句號的字串，加上英文字母字串組成。
C.-T. King	由一個大寫英文字母加英文句號，英文破折號加上一個大寫字母及英文句號，空白加上英文大寫加上至少一個英文字母組成

```
letter      -> A|B|C…|Z|a|b|c…..|z
letters      -> letter
dot         -> .
comma        -> ,
space       ->
Name         -> (letter dot)$^+$ letters | letters comma letter dot letter$^*$ |letters space (letter dot)$^*$ letters
```

圖 6、英文姓名正規表示式

因此我們定義幾個非終止符號(non-terminal)以及單詞(token)，並建立幾個英文姓名的樣式。圖 6 是英文姓名樣式的正規表示式。這些樣式之中，樣式二有逗號存在，而逗號常用在分開作者姓名，因此優先權上樣式二要最後處理，樣式一跟樣式三並不衝突，因此先後做的順序並無差別。得知姓名的樣式之後，就可以取出一筆參考文獻作者或作者群。圖 7 是取出參考文獻作者的演算法。

```
輸入: 一筆參考文獻 R
輸出: 作者名字集合 A=$\{A_1,A_2,\ldots A_\Theta\}$，Θ 爲一筆參考文獻的英文作者數

步驟 1: 由前一節演算法找出的論文名稱，將參考文獻在論文名稱 T 之前的文字段落 S 取出，
        S=R-T-(T 之後的段落)
步驟 2: 令 W 爲符合表 3 樣式的字串，作者名字集合 A={}，先檢查 S 是否有對等連接詞 and，
        以查證 S 是否有一個以上的作者名字
  步驟 2.1: 如果有，檢查對等連接詞前後是否有空白，避免誤認人名中間的字母爲連接詞
            若有空白，表示確實爲對等連接詞，將 S 依照連接詞切開並將切開後的後面段落 C
                      加入 A 中，並將段落 C 從 S 中移除，S=S-C
            若無空白，則表示誤認，跳至步驟 3
  步驟 2.2: 如果沒有對等連接詞，跳至步驟 3
步驟 3: 若 S 爲空，表示已經處理完畢，結束
        若 S 有符合樣式的字串 W，則加入 W 到 A 中，並將字串 W 從 S 中移除，S=S-W
        若 S 沒有符合樣式的字串 W，結束
```

圖 7、取出參考文獻英文作者的演算法

3.3.3 參考文獻的直接與間接引用

當兩篇論文引用到同一篇論文時，我們稱這兩篇論文直接引用同一篇論文。因此我們將論文的參考文獻區段細分成一筆筆的參考文獻，再找出兩篇論文的參考文獻中取出標題，最後對兩篇論文所取出的參考文獻標題進行字串比對，如果比對有相同的參考文獻標題，我們就認定兩篇論文有參考文獻的直接引用。

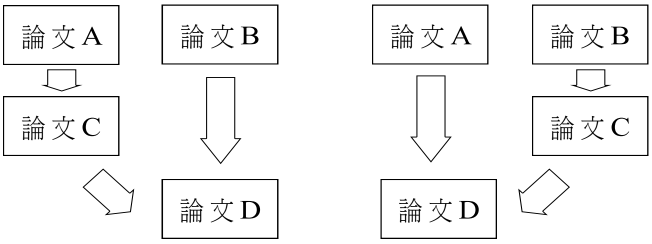


圖 8、論文間接引用的兩種情形

有時候論文會有間接引用的情形，也就是兩篇論文並不是直接引用同一篇，而是在引用到參考文獻的原文中引用到共同的論文，使得兩篇表面上沒有引用到共同參考文獻的論文卻有著極高的關連性。間接引用又可分成兩種情形，如圖 8 所示，圖 8 左邊論文A 跟論文 B 並沒有直接引用到同一篇論文，但是論文 A 引用的論文 C 卻跟論文 B 共同引用論文 D；同樣的圖 8 右邊論文 B 跟論文 A 並沒有直接引用到同一篇論文，但論文B 引用的論文 C 卻跟論文 A 共同引用了論文 D，我們稱論文 A 跟論文 B 有著間接引用。因此在處理參考文獻的引用上，就必須處理間接引用，其演算法如圖 9。

```
輸入: 兩筆參考文獻 P 和 Q
輸出: 是否間接引用

步驟 1 : 到資料庫找出 P 和 Q 的論文本文 $U_p$ 和 $U_q$，分別取出 $U_p$ 和 $U_q$ 的參考文獻段落，集合
         $R_p=\{p_1,p_2\cdots,p_g\}$，$R_q=\{q_1,q_2\cdots,q_h\}$
步驟 2 : $\forall p_j \in R_p$，取出論文的名稱，並檢查是否跟 Q 有直接引用
         $\forall q_j \in R_q$，取出論文的名稱，並檢查是否跟 P 有直接引用
步驟 3 : 若有，回傳是。反之回傳否。
```

圖 9、檢查兩篇參考文獻是否間接引用的演算法

3.3.4 摘要、簡介與結論區段相似度計算

摘要、簡介與結論區段部分，由於採用向量空間模型的方法處理，本研究用餘弦函數[9]來計算評審教授發表的論文與待審論文相似度。在幾何學中，兩個向量若是越相近，所夾的夾角 θ 也會越小；而利用餘弦函數的特性，θ 夾角越小所得出的餘弦函數值也會越大。因此可以得到一個結論：兩個純文字向量的餘弦函數值越大，代表兩篇文章的文字向量所形成的夾角越小，則此兩篇文章內容越相似；反之，就代表兩篇文章越不相關。

3.3.5 參考文獻相似度計算

由參考文獻標題的直接引用、間接引用以及參考文獻共同作者，可計算出兩篇論文的近似程度，進而由最相似論文來建議評審委員。在論文間接引用處理上，由於並不是直接引用到同一篇論文，因此其關連性弱於直接引用。如果直接引用一筆參考文獻，我們定義配分是 1 分，若是像圖 8 那樣間接引用一筆參考文獻則定義 0.7 分。本研究中參考文獻共同引用作者也是給予 0.7 分的權重，一篇參考文獻可能引用同一作者的數篇論文，在相關性上也較直接引用為弱。公式(3)是兩篇論文的參考文獻區段相似度計算公式。

$$Simularity_{reference} = Title_{direct} \times 1 + Title_{indirect} \times 0.7 + Authors \times 0.7 \qquad (3)$$

其中 Title$_{direct}$ 表示參考文獻標題直接引用的數量，Title$_{indirect}$ 表示參考文獻標題間接引用的數量，Authors 表示參考文獻共同引用作者數量。

4. 實驗結果

本研究從 2007 年人工智慧與應用研討會的評審委員名單中，挑選九十八位作為評審委員的候選人，並收集這些評審委員的相關論文，平均一位收集十篇到十五篇左右，總共約收集一千零八十九篇來做為訓練資料。訓練資料來源除了從網路上下載，還包含了 2002 年、2003 年、2004 年及 2005 年人工智慧與應用研討會論文集的論文以及 2001 年、2003 年及 2005 年全國計算機會議 (National Computer Symposium)論文集的論文。測試資料則是選自 2007 年人工智慧與應用研討會共 74 篇論文。

4.1 取出區段的精確度

計算取出區段的精確度，我們以人工方式取出原始論文的區段文字作為標準答案，分別對系統取出的區段及標準答案區段進行中文分詞處理。中文分詞處理我們是透過對詞庫

裡詞彙的比對搜尋，依照「長詞優先」法則來對中文語句作分割，分詞處理上能得出較正確的詞彙。對論文做完分詞工作後，可以得出論文出現了哪些詞彙，以及詞彙在論文中出現的頻率，因此可計算系統取出區段的總詞彙數，標準答案區段詞彙數以及系統取出區段詞彙與標準答案區段詞彙的交集詞彙數，我們以這些資訊計算出摘要、簡介與結論區段的 precision 與 recall。公式(4)和(5)中 precision$_{part}$ 表示區段的 precision 值，recall$_{part}$ 表示區段的 recall 值，y 代表字典檔的詞彙總數，i 代表詞彙，V_i 表示詞彙 i 在取出區段的出現次數，W_i 表示詞彙 i 在標準答案區段出現的次數，TP_i 表示詞彙 i 在取出區段跟標準答案區段共同出現次數。

$$precision_{part} = \frac{\sum_{i=0}^{y} TP_i}{\sum_{i=0}^{m} V_i} \qquad (4) \qquad\qquad recall_{part} = \frac{\sum_{i=0}^{y} TP_i}{\sum_{i=0}^{m} W_i} \qquad (5)$$

本研究從測試資料中選出 25 篇論文來評估取區段的精確度，將 25 篇論文取各區段的 precision 與 recall 平均值，可得出取出各區段的平均 precision 與 recall。參考文獻區段由於要取出作者跟論文名稱，因此不同於其他段落要進行分詞，取出參考文獻精確率的算法，我們細分成參考文獻作者取出精確率，以及參考文獻標題取出精確率。公式(6)和(7)分別是取出參考文獻作者的 precision 與 recall。

$$precision_{Author} = \frac{NAC}{NAR} \qquad (6) \qquad\qquad recall_{Author} = \frac{NAC}{NAA} \qquad (7)$$

其中 NAR 表示系統取出的參考文獻作者數，NAC 表示系統取出且合乎標準答案的作者數，NAA 表示標準答案的作者數。參考文獻標題取出精確度算法與參考文獻作者取出精確率算法相同，只需將 NAR 代換成系統取出的參考文獻標題數，NAC 表示系統取出且合乎標準答案的參考文獻標題數，NAA 表示標準答案的參考文獻標題數。我們以人工去對 25 篇論文取出一筆筆參考文獻的標題，以及參考文獻作者作爲標準答案，計算取出參考文獻標題以及作者的 precision 跟 recall。表 4 是取出摘要區段、簡介區段、結論區段、參考文獻標題及參考文獻作者的 precision 與 recall 對應表。

表 4、取出摘要、簡介及結論區段、參考文獻標題與作者的 precision 和 recall

	摘要	簡介	結論	參考文獻作者	參考文獻標題
precision	94.32%	99.67%	91.47%	44.23%	60.74%
recall	94.50%	77.83%	72.93%	52.11%	61.68%

4.2 推薦單一評審委員

由之前所介紹的方法，一篇論文可轉換成數個由不同區段而成的向量以及數條參考文獻。經由相似度的計算，可得出一篇論文各區段的各自最相似論文，再由找出論文作者得出建議評審委員的名字。因此一篇待審論文會有各個區段的建議評審委員，本研究採取各區段權重相同的做法，也就是以區段評審委員名字出現最多次的作爲該篇論文的建議評審委員。爲了顯示的方便，本節實驗列表僅列出測試資料中選出六篇論文的實驗結果，建議評審委員則是從 2007 人工智慧與應用研討會的評審委員名單中選出。

藉由相似度的計算，我們可以找出論文區段的建議評審委員。以摘要區段爲例，取一篇待審論文的摘要區段，對所有訓練資料論文的摘要區段做餘弦函數值，同時記錄訓練資料論文的作者名稱。再找出最大的餘弦函數值的論文作者，即是建議的評審委員。同樣的，簡介及結論區段也可找出個別的建議評審委員。參考文獻區段則是利用公式(3)，計算待審論文跟所有訓練資料論文參考文獻區段的相似度，進而找出建議評審委員。圖 10 是由參考文獻找出建議評審委員的演算法。

```
輸入：一篇論文 E 與資料庫 F
輸出：建議的評審委員

Step1：取出論文 E 的參考文獻集合 $R=\{r_1,r_2\ldots r_\Delta\}$，Δ 爲論文 E 的參考文獻總數量
Step2：$\forall r_j \in R$，取出論文的名稱，找出資料庫 F 的論文是否有共同引用此論文，並找出參考文獻
       跟資料庫 F 內論文的共同作者
Step3：計算 E 與 F 內論文其參考文獻標題直接引用、間接引用以及參考文獻共同作者的數量，並依
       照公式(3)算出總值。
Step4：回傳最大總值論文作者作爲建議評審委員。若總值爲零，則回傳"找不到建議的委員"
```

圖 10、由參考文獻建議評審委員的演算法

表 5、參考文獻標題與作者共同引用數量範例表

檔名	建議評審委員	標題直接引用篇數	標題間接引用篇數	共同引用作者數
基於 SVM 與 LDA 演算法之人臉辨識	曾守正	1	0	2
模組化線性鑑別式分析應用於人臉辨識	劉吉軒	0	0	1
基於紋理特性之移動物體偵測法則	劉吉軒	0	0	1
應用於 BDI Agent 之案例式推理系統開發工具	李宗南	0	0	1
使用小腦模型類神經網路控制冷氣空調機馬達	找不到建議的委員	0	0	0
可拓基因演算法	吳志宏	1	0	0

表 5 是參考文獻建議委員的列表，欄位從左到右依序是檔名、建議評審委員、標題直接引用篇數、標題間接引用篇數以及共同引用作者數。由表 5 可看出參考文獻的共同引用機會其實很低，以「使用小腦模型類神經網路控制冷氣空調機馬達」這篇論文來說，沒有任何直接引用、間接引用及共同作者，這樣的情況下我們會在建議評審委員欄位填上「找不到建議的委員」。

表 6、各區段建議評審委員

	摘要	簡介	結論	參考文獻	評審委員
基於 SVM 與 LDA 演算法之人臉辨識	黃有評	張嘉惠	張嘉惠	曾守正	張嘉惠
模組化線性鑑別式分析應用於人臉辨識	蔡正發	方國定	張智星	劉吉軒	找不到建議的委員
基於紋理特性之移動物體偵測法則	范欽雄	范欽雄	范欽雄	劉吉軒	范欽雄
應用於 BDI Agent 之案例式推理系統開發工具	林豐澤	許永真	劉吉軒	李宗南	找不到建議的委員
使用小腦模型類神經網路控制冷氣空調機馬達	古鴻炎	王學亮	楊正宏	找不到建議的委員	找不到建議的委員
可拓基因演算法	許永真	張嘉惠	陳慶瀚	吳志宏	找不到建議的委員

藉由整合一篇論文各區段的建議評審委員，我們可以找出該篇論文的建議評審委員。由表 6 所示，第一行欄位爲論文檔案名稱，第二行欄位爲這些論文摘要區段的建議評審委員，第三行欄位爲這些論文簡介區段的建議評審委員。同理，第四行、第五行欄位分別代表這些論文結論與參考文獻區段的建議評審委員。最後一行評審委員欄位則是各區段投票後的結果，以最多區段建議的評審委員作爲該篇論文的建議評審委員。

以「基於紋理特性之移動物體偵測法則」這篇論文爲例，該篇的摘要、簡介及結論區段建議評審委員皆是「范欽雄」，而參考文獻區段則是「劉吉軒」，因此投票數最高的評審委員爲「范欽雄」，投票數爲三票。由於有些論文有跨領域的情形，各區段建議評審委員可能會是不同的人而導致沒有共同的建議評審，這時就無法建議評審委員。以「模組化線性鑑別式分析應用於人臉辨識」這篇論文爲例，該篇各區段的建議評審委員皆是不同的人，這樣的情況我們就無法建議評審委員，因此在評審委員欄位填上「找不到建議的委員」。像這樣區段無法建議評審委員的情形我們視同無投票能力，通常參考文獻區段欄位會有這樣的情形。對一般論文來說，由於領域眾多因此論文數量相當可觀，只有領域很相近的論文才有可能引用到同篇參考文獻。兩篇論文引用到同一篇論文的機率可說是微乎其微，因此參考文獻推薦評審教授的論文數量很少是可以預見的。

4.3 推薦多重評審委員

現實的會議論文指派情形，一篇論文不會只指派給一位評審委員，而是會分給三位評審委員左右，議程主席再整合評審委員們的意見決定該論文是否被會議接受。因此我們考慮取出前十名近似的論文，再以這些論文的作者選出三位來作爲建議的評審委員。在建議評審委員的選擇上，論文最近似無疑是最佳選擇，但是如果有多篇近似的論文則也暗示著作者對該領域有興趣，可以作爲建議評審委員的考量。

在各區段的建議評審委員上，不再是以取最相似的論文作者而是改以取前十名。本研究將前十名的十篇論文作者由近似度高到低排序，並分別給予分數第一名 10 分到第十名給予 1 分，接著觀察這些前十名的作者是否有重複出現，若有，則表示該作者有數篇近似的論文，於是將該作者出現在這前十名的這幾篇論文分數加總，作爲待審論文建議給該作者的分數。若無，則以該作者現有分數作爲分數。再由分數最高的作者依序排序選出最高的前三名作爲評審委員。本節列表只僅列測試資料中選出六篇論文的實驗結果，建議評審委員則是從九十八位 2007 人工智慧與應用研討會的評審委員中選出。

表 7、摘要區段前十名近似論文的作者

基於 SVM 與 LDA 演算法之人臉辨識	黃有評	林豐澤	林豐澤	呂永和	陳士杰	方國定	廖純中	陳慶瀚	曾憲雄	呂永和
模組化線性鑑別式分析應用於人臉辨識	蔡正發	李昇暾	陳慶瀚	陳士杰	方國定	廖純中	孫光天	林豐澤	呂永和	曾憲雄
基於紋理特性之移動物體偵測法則	范欽雄	林正堅	曾新穆	呂永和	廖純中	方國定	陳士杰	陳慶瀚	曾憲雄	林豐澤
應用於 BDI Agent 之案例式推理系統開發工具	林豐澤	方國定	陳士杰	廖純中	呂永和	楊東麟	黃有評	曾守正	吳志宏	曾憲雄
使用小腦模型類神經網路控制冷氣空調機馬達	古鴻炎	楊正宏	陳慶瀚	林豐澤	方國定	陳士杰	廖純中	呂永和	曾憲雄	孫光天
可拓基因演算法	許永真	呂永和	陳士杰	廖純中	方國定	呂永和	林豐澤	曾新穆	曾憲雄	林豐澤

以摘要區段爲例，表 7 是跟待審論文最近似的前十名論文作者表，在這裡取六篇論文來觀察，最左邊的欄位是論文檔案名稱，接著欄位由左到右是最近似論文的作者到較不近似論文的作者，觀察「基於 SVM 與 LDA 演算法之人臉辨識」這篇論文，前三名近似部分有兩名是林豐澤，表示這位作者有近似且數量不少的著作，在建議委員的選擇上可能更勝於最近似的黃有評；「可拓基因演算法」這篇論文也是類似的情況，「可拓基因演算法」這篇論文呂永和佔了兩篇名額，一篇較爲近似另一篇則大約處於中等近似的程度，但由於篇數加成的關係在選擇上優先權要高過第一名的許永真。表 8 是這六篇論文的摘要區段前三名建議評審委員的列表。

表 8、摘要區段前三名建議評審委員

基於 SVM 與 LDA 演算法之人臉辨識	林豐澤(17 分)	黃有評(10 分)	呂永和(8 分)
模組化線性鑑別式分析應用於人臉辨識	蔡正發(10 分)	李昇暾(9 分)	陳慶瀚(8 分)
基於紋理特性之移動物體偵測法則	范欽雄(10 分)	林正堅(9 分)	曾新穆(8 分)
應用於 BDI Agent 之案例式推理系統開發工具	林豐澤(10 分)	方國定(9 分)	陳士杰(8 分)
使用小腦模型類神經網路控制冷氣空調機馬達	古鴻炎(10 分)	楊正宏(9 分)	陳慶瀚(8 分)
可拓基因演算法	呂永和(14 分)	許永真(10 分)	陳士杰(8 分)

表 9、利用多重委員建議評審結果表

	摘要	簡介	結論	參考文獻	評審委員

基於 SVM 與 LDA 演算法之人臉辨識	林豐澤(17 分)
	黃有評(10 分)
	呂永和(8 分)
		張嘉惠(10 分)
		林正堅(9 分)
		陳慶瀚(8 分)
			張嘉惠(10 分)
			陳慶瀚(9 分)
			陳士杰(8 分)
				曾守正(10 分)	張嘉惠
					陳慶瀚
					林豐澤

模組化線性鑑別式分析應用於人臉辨識	蔡正發(10 分)
	李昇暾(9 分)
	陳慶瀚(8 分)
		陳士杰(16 分)
		方國定(10 分)
		廖純中(8 分)
			張智星(10 分)
			林正堅(9 分)
			陳慶瀚(8 分)
				劉吉軒(10 分)	陳士杰
					陳慶瀚
					蔡正發

基於紋理特性之移動物體偵測法則	范欽雄(10 分)
	林正堅(9 分)
	曾新穆(8 分)
		范欽雄(10 分)
		林正堅(9 分)
		陳慶瀚(8 分)
			范欽雄(10 分)
			鄭炳強(9 分)
			林正堅(8 分)
				劉吉軒(10 分)	范欽雄
					林正堅
					劉吉軒

應用於 BDI Agent 之案例式推理系統開發工具	林豐澤(10 分)
	方國定(9 分)
	陳士杰(8 分)
		林豐澤(17 分)
		許永真(10 分)
		廖純中(7 分)
			劉吉軒(10 分)
			曾憲雄(9 分)
			廖純中(8 分)
				李宗南(10 分)	林豐澤
					廖純中
					許永真

使用小腦模型類神經網路控制冷氣空調機馬達	古鴻炎(10 分)
	楊正宏(9 分)
	陳慶瀚(8 分)
		王學亮(10 分)
		林正堅(9 分)
		曾守正(8 分)
			楊正宏(10 分)
			林豐澤(9 分)
			陳士杰(8 分)
				找不到建議的委員(0 分)	楊正宏
					古鴻炎
					王學亮

可拓基因演算法	呂永和(14 分)
	許永真(10 分)
	陳士杰(8 分)
		呂永和(11 分)
		張嘉惠(10 分)
		林正堅(9 分)
			呂永和(11 分)
			陳慶瀚(10 分)
			陳士杰(9 分)
				吳志宏(10 分)	呂永和
					吳志宏
					許永真

簡介區段與結論區段作法跟摘要區段相同。如果有建議評審同分的情況，我們保留兩位教授都可作爲評審委員的候選人，將來指派教授當其中一位教授被指派到過多的論文時，就由另一位教授補上。參考文獻的建議評審委員方面，由於一般論文不容易有共同引用的現象，取前三名建議評審似乎沒有實質的幫助，因此在選擇上仍然是以取一名委員來做處理。參考文獻的建議評審委員如表 5。由於參考文獻的重要性，我們給予參考文獻段落建議委員 10 分。表 9 是利用多重委員建議評審結果的表。我們將各段落前三名的作者及其分數加總，選出總分最高的前三名作者作爲建議評審委員。

得出系統建議評審委員之後，我們評估建議評審的準確率，評估的標準包含 precision，recall 及 F-measure。本研究用人工建立建議評審委員的答案表，由「本研究的指導老師」協助建立，以計算出系統推薦論文評審委員的準確率。在本研究前面的段落，我們展示了整合各區段推薦委員的方法，並推薦出三位評審委員，同樣的可將推薦委員的人數增加，不一定只推薦三位。表 10 是多重委員推薦三位評審及五位評審的準確率對應表。系統推薦五名評審委員的 precision 稍高過兩成五，也就是說系統推薦五名的評審委員，至少有一名是人工推薦在答案表上的。另外由表 10 可看出系統的 recall 相當的低，可能是因爲是我們在建立建議評審委員的答案表時，並未限定一篇論文的評審委員人數，使得每篇論文可以指派給多位評審委員，因而造成 recall 值不高。此外，評審委員的著作論文篇數不同，也可能造成論文分派到著作多而非最適合的評審。

表 10、多重委員建議三位評審及五位評審的對應答案準確率

	precision	recall	F-measure
三名評審委員	28.16%	5.72%	9.13%
五名評審委員	26.20%	8.56%	12.11%

5. 結論

藉由計算待審論文區段與資料庫論文區段之間的相似度，我們得出各區段的建議評審委員。本研究整合不同區段的建議評審，來找出待審論文的建議評審委員。目前本研究是採取不同區段相等比重的方法，然而一篇論文中各個區段重要性不一定相同，因此在整合區段評審委員時，各區段應有著各自的權重。未來本研究可能會應用機器學習的技術，調整各區段最適當的權重，使得指派效果得以提昇。同時，如果待審論文是新的領域或技術，也會造成找不到適當的建議評審。也許可以採取設立門檻值的作法，找出跟每位評審相似度都不高的論文，並回報讓議程主席得知這些論文不容易被分配到建議評審。

由於論文領域的不同，論文評審的建議變得困難，也因此論文關鍵詞擷取相對來的重要。本研究目前採用以 HowNet 爲主的字典檔，未必能包含論文的重要關鍵詞。將來可能會找尋含有更多資訊技術關鍵字的字典檔，使得論文的建議評審結果更加準確。同時不同詞彙也存在不同程度的關連性，單純的使用 tf-idf 無法完全反映評審委員對該論文主題專長程度。另外，本研究並未進行適當篇數論文指派給評審，以及論文指派給適當數量評審的處理[12]，這些都是需要加強改善的項目。

致謝

本研究承蒙國科會研究計畫 NSC-95-2221-E-004-013-MY2 的部分補助謹此致謝。我們感謝匿名評審對於本文初稿的各項指正與指導。

參考文獻

[1] 人工智慧與應用研討會，http://www.taai.org.tw/ [Access: Jun. 28, 2008]

[2] 九二八電腦股份有限公司，http://www.928n.com.tw/928index.asp [Access: Jun. 28, 2008]

[3] 李振昌，李御璽，陳信希，“中文文本人名辨識問題之研究”，*第七屆自然語言與*

語音處理研討會，1994

[4] 李駿翔，*應用資料探勘分類技術於專利分析之研究*，碩士論文，中原大學，台灣，桃園，2003。

[5] 林蘭綺，*專利文件之自動分類研究*，碩士論文，國立交通大學，台灣，新竹，2006。

[6] 駱思安、李中彥及徐俊傑，"以MMB 演算法改良中文網站自動分類系統的效能"，*全國計算機會議論文集*，論文光碟，2005。

[7] 錢炳全及廖雙德，"中文試題自動分類方法"，*第七屆人工智慧與應用研討會論文集*，論文光碟，2002。

[8] 顧皓光及莊裕澤，"網路文件自動分類"，*全國計算機會議論文集*，論文光碟，1997。

[9] Amit Bagga and Breck Baldwin, "Entity-Based Cross-Document Coreferencing Using the Vector Space Model", *Proceedings of the 36th Annual Meeting on Association for Computational Linguistics*, 1998, Volume 1, Pages 79–85.

[10] Chyi-Ren Dow, Khong-Ho Chen, Shu-Yi Lin, Yan-Ling Liu, Chih-Chieh Peng, and Sheng-Jie Guan, "Design and Implementation of a DSPACE-based Recommender System for Digital Literature Retrieval", *Proceedings of the 12th Conference on Artificial Intelligence and Applications*, CD-ROM, 2007.

[11] DSpace, http://www.dspace.org/ [Access: Jun. 28, 2008]

[12] David Hartvigsen, Jerry C. Wei and Richard Czuchlewski, "The Conference Paper-Reviewer Assignment Problem", *Decision Sciences*, 1999, Volume 30, Issue 3, Page 865-876.

[13] HowNet, http://www.keenage.com. [Access: Jun. 28, 2008]

[14] Google Scholar, http://scholar.google.com [Access: Jun. 28, 2008]

[15] Ann P. Bishop, "Document Structure and Digital Libraries: How Researchers Mobilize Information in Journal Articles", *Information Processing and Management*, 1999, Pages 255-279.

[16] Gerard Salton and Michael J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1986.

[17] Gerard Salton, A. Wong and C. S. Yang, "A Vector Space Model for Automatic Indexing", *Communications of the ACM*, 1975, Volume 18, Issue 11, Pages 613–620.

[18] Leah S. Larkey, "A Patent Search and Classification System", *Proceedings of the 4th ACM Conference on Digital Libraries*, 1999, Pages 179–187.

多領域文件集之詞彙概念擴展與知識架構之建立

Conceptual Expansion and Ontological Mapping of Multi-domain Documents

陳永祥　Yong-Xiang Chen
中央研究院語言學研究所 臺灣大學資訊工程研究所
yxchen@gate.sinica.edu.tw

柯綉玲　Xiu-Ling Ke
中央研究院語言學研究所
vitake@gate.sinica.edu.tw

陳克健　Keh-Jiann Chen
中央研究院資訊科學研究所
kchen@iis.sinica.edu.tw

黃居仁　Chu-Ren Huang
中央研究院語言學研究所
churen@gate.sinica.edu.tw

摘要

傳統資訊檢索過程經常透過查詢擴展技術來增加檢索結果的數量。在大型數位博物館中，考量存放的典藏品項具有領域特殊性，以及不同品項由各自適用的文字描述所造成之差異性，因此傳統查詢擴展方式不盡合用。本研究提出概念擴展之構想，首先由數位博物館典藏品項的標題中抽取出成分詞集，再透過中英雙語知識本體（BOW）將成分詞對應至建議上層知識本體（SUMO），藉此可將成份詞轉換成為概念並對應至具有結構的知識本體節點上，再由集群演算法計算相近節點並將分散之對應節點聚合成具有代表性的群集，最後以構詞分析所得之規則進行群集縮減，決定出符合構詞分析規則的群集用以進行概念擴展。研究成果除提出知識概念擴展流程外，亦以數位典藏國家型科技計畫典藏項目為例，歸納出保存多領域典藏品之數位博物館中文標題構詞樣式，進行分析探討，研究成果可作為機器自動處理之基礎。

關鍵詞：概念擴展，構詞樣式，知識本體，群集縮減

一、前言

傳統資訊檢索過程經常透過查詢擴展技術來增加檢索結果的數量。在大型數位博物館中，考量存放的典藏品項具有領域特殊性，以及不同品項由各自適用的文字描述方式所造成的差異性，因此傳統查詢擴展方式不盡合用。概念擴展的構想是以詞彙所代表的知識概念在知識本體中結構中找到相似的概念進行擴展，以使原本只能對應到少數概念的詞彙透過概念延伸得以對應至較多的詞彙，進而從知識概念層次的擴展提升數位博物館典藏品資料的資訊檢索效能。

「數位典藏國家型科技計畫」自 2002 年開始推動，旨在將珍貴的重要文物典藏加以數位化，建立國家數位典藏，以保存文化資產、建構公共資訊系統，促使精緻文化普及、資訊科技與人文融合，並推動產業與經濟發展。因規模龐大，因此目前開發整合型的成果查詢介面提供各界使用者查詢應用，分別為聯合目錄及公共展示系統。

透過知識本體的結構系統，可以比較嚴謹的將知識結構系統建立起來，本研究即以數位典藏國家型科技計畫所提供之 39,765 個典藏品標題（2~5 字詞）為實驗資料，透過中文詞彙網路（Chinese Wordnet; CWN）及中英雙語知識本體詞網（Sinica BOW）所建立的中文詞義分析與知識本體架構，將標題詞彙對應至 SUMO ontology 節點上進行概念擴展與群集，再以中文構詞樣式為標準進行群集縮減，以詞彙語義分析方式提供特殊領域典藏資料庫中資訊檢索可行之方案。

二、相關研究

(一) 語義相似度與查詢擴展

距離導向的相似度方法是從大的文字語料庫中去學出分布的相似度來建立模型，Leacock & Chodorow [1]，Resnik [2]，Lin [3]所提出的是三種在自然語言處理應用上很標準的方法。這些公式都是定義用來測量概念(Concept)上的相似度, 而非詞彙(word)上的, 但在轉換上可以用一對詞與詞之間，多組概念對概念相似度中最高的那組來作為語義相似度的代表。因此可以簡單轉換成詞與詞的相似度計算。

資訊檢索一般可針對檢索的資料類型區分為兩種，第一種是針對網際網路上所有的資料內容所進行的檢索，由於檢索的範圍太過廣泛，因此必須透過許多不同的策略來針對查詢關鍵字進行擴展以求找出使用者有興趣的內容，大多數的網際網路搜尋引擎網站所提供之服務皆屬於此類。第二種則是針對特定範圍資料所進行的資訊檢索，此類資訊檢索的使用者是在資料內容固定的情況下進行查詢，例如新聞媒體網站或是數位博物館網站。

一般資訊的檢索是用詞彙來代表概念。但概念與詞彙的關係並非都是一對一的，如同義詞（Synonym）用來表示多個詞彙都具有相同的概念，即一個概念可對應到多個詞彙。因此，在檢索時若能建構概念與詞彙間的明確關係，將有效提升檢索效益。在一般搜尋引擎的設計上，最常見的策略是利用詞形比對的方式作為資訊檢索的基本方法，再輔助以各種的查詢詞彙擴展或是相關統計運算結果來找出使用者感興趣的資料。

除了利用關鍵詞進行全文檢索（Full-Text Search）外，有些資訊檢索系統尚針對文件的內容進行分析，給予文件資料檢索標識（如主題詞彙或分類號），並使用索引詞彙來表示文件內容，資訊使用者與資訊檢索系統之間藉由索引詞彙與檢索詞彙之間的對應來達到擷取與過濾資訊的目的。查詢問句的擴展通常以使用者提供的檢索詞彙為基礎，當原始查詢問句的檢索效益不好時，則可以追加更多的詞彙來改善。關於查詢問句的擴展，尚有相關研究提出利用相關回饋（Relevance Feedback）或是使用知識架構（Ontology）的元知識（Atom Knowledge）來進行[4]。Stiles 是最早提出利用相關詞彙來改進檢索效益理論的學者之一[5]。

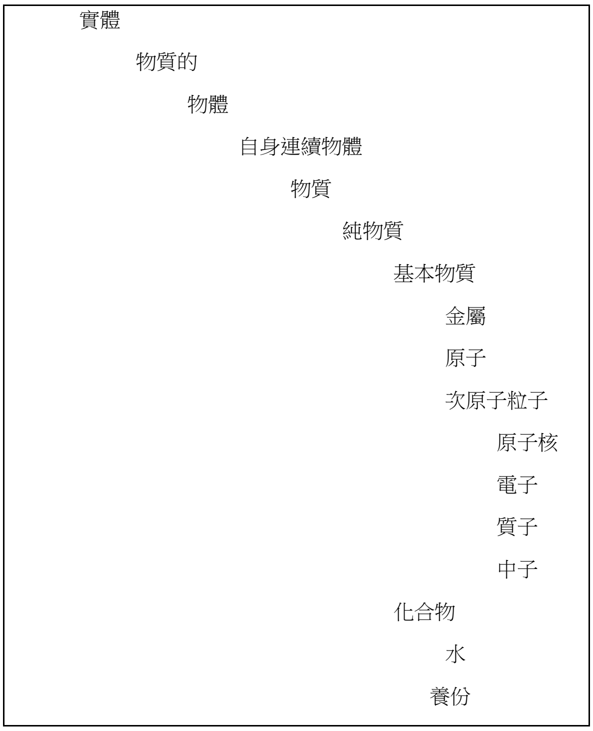
(二) 建議上層共用知識本體（Suggested Upper Merged Ontology；SUMO）

SUMO（Suggested Upper Merged Ontology，建議上層共用知識本體）[6]是由 IEEE 標準上層知識本體工作小組所提出的知識本體架構，目的是發展成標準的上層知識本體，這將促進資料互通性、資訊搜尋和檢索、自動推理和自然語言處理。知識本體（ontology）類似於一組字典或術語表，但能夠使電腦處理更多內容的細節和其結構。透過知識本體可將人們有興趣的領域正規化為一套概念、關係和定理（axiom）。上層的知識本體被限制在 meta 的概念、一般、抽象或者哲學，因此足夠一般提出（在一定水準上）一個涵蓋廣闊範圍的領域區域[7]。特殊領域具體的概念不被包括在上層知識本體中，但是這樣的知識本體可提供特殊領域（例如：藥、財政、專案…等等）的知識本體結構的建立。SUMO 藉由最高層次的知識本體，鼓勵其他特殊領域知識本體以其為基礎衍生出其他特殊領域的知識本體，並為一般多用途的術語提供定義。目前 SUMO 已經和英語詞彙網路 WordNet1.6 版本作初步的連結。SUMO 中的節點以階層樹方式連結，如圖一所示。

(三) 中央研究院中英雙語知識本體詞網（Sinica BOW）

中英雙語知識本體詞網（Sinica BOW）[8]是一結合詞網（WordNet）知識本體與領域標記的詞彙知識庫，由中央研究院語言所中文詞彙網路小組與資訊所中文詞知識庫小組合作建置，從語言工程的角度，以台灣地區的語言使用為經驗基礎，提供語言和語言、語言和概念以及語言和領域的資訊，甚至是跨語言間的訊息。Sinica BOW以建立一完整精確的中英對譯資料庫及檢索介面為目的，作為數位典藏知識國際化的基礎；並持續建立各領域之雙語領域辭典，以作為各領域／典藏之雙語控制詞彙參考標準。中英雙語知識本體詞網同時提供具領域判斷能力之資訊檢索應用。此外，建立附加領域標記之雙語辭典及檢索介面使中英雙語知識本體詞網成為一知識加值雙語電子辭典。

Sinica BOW 主要使用的資源包含 WordNet、ECTED（English- Chinese Translation Equivalents Database）以及 SUMO（Suggested Upper Merged Ontology，建議上層共用知識本體）。其中 WordNet[9]是 1985 年普林斯頓大學認知科學實驗室以現代心理語言學理論所述的人類詞彙記憶為啟發所開發出的語意式電子字典，以每個同義詞集表達一種詞彙概念，將同義詞集區分為四種英文詞類：名詞、動詞、形容詞、副詞，並以二十幾種詞義關係組織同義詞集。由中研院資訊所與語言所合作建構的 ECTED 以 WordNet 為基礎，經由現有英中或中英電子辭典的詞形對應，爲每個同義詞集詞義找出可能相對應的中譯詞組，再經由人工檢驗。尋找對譯盡可能的以詞彙而非描述性短語表達，目的在於讓每個同義詞集都有最適當的一至三個左右的中文對譯。[10]

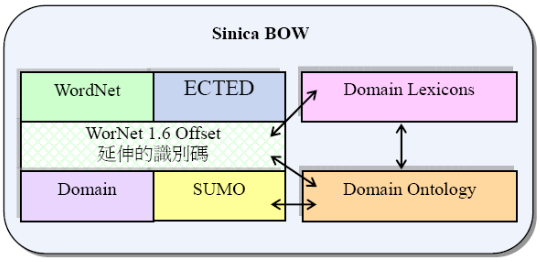


圖一、 SUMO階層節點示例

依據SUMO 2002年版資料，黃居仁等人[11]將系統介面以及概念節點進行中文化納入Sinica BOW之中並進行對應連結，其涵蓋11大類的概念，每大類又區分為二至五個類別，總共囊括3,912個概念。SUMO已經與WordNet1.6版本結合，且以同義（synonymy）、上位（hypernym）、體例（instantiation）這三種類別顯示同義詞集和SUMO概念間的對應關係。除此，更以「中國圖書分類法」為基準，並參考各知識分類與實際研究經驗，提出：包含九大類的知識分類（Knowledge Content），涵蓋427個領域。另外，並因應語言資源特性加入下列語言使用（Language Usage）的各類訊息：專名（說明文字符號的指涉）（Proper Name）、語體（說明文字符號的使用）（Genre/Strata）、各種語言／詞源（Language/Etymology）、各國地名（Country Name）。領域階層的建立在於替不同詞義中的詞彙項目區別其使用的領域。加註領域信息可降低詞彙歧異性，增加資料交換時的互通性，輔助領域詞彙庫之建構。Sinica BOW透過WordNet1.6 offset延伸所產生的識別碼作為媒介，進行串連，將每個資源以及各類訊息連結。

因WordNet1.6 offset延伸的識別碼可獲得原本WordNet存在的詞類、解釋、英文例句、同義詞集、各同義詞集間的詞義關係及其所屬詞彙。而SUMO概念與WordNet的連結，使得可透過該識別碼獲取詞義與概念搭配的訊息。以WordNet為基礎所建置的ECTED與針對WordNet同義詞集的各詞彙項目所給予的領域值，也是透過該識別碼獲取。而特殊領域詞彙庫，加上相對應的Sinica BOW識別碼，也可保留原始資源的資料庫格式和WordNet連結。

因領域知識本體則是在SUMO某些概念下進行延伸發展。每個特殊領域詞彙庫中的詞彙一樣具有所屬的概念，其所屬概念可能是SUMO或特殊領域知識本體的某一概念，特殊領域詞彙庫和領域知識本體的結合，使得透過該識別碼又串起所有的訊息。Sinica BOW的資源和架構如圖二所示。由於透過WordNet可以和同是以WordNet為基礎架構所建置的其他語系WordNet資源加以連結，例如：EuroWordNet[7]，因此以此基礎架構可編製成多語的詞彙網路，成為多語環境中所需之語言知識結構的基礎資料。



圖二、 SINICA BOW 架構圖

三、知識本體對應與構詞分析

(一) 研究資料

對於儲存多樣領域知識的數位博物館而言，將典藏品項涵蓋的知識概念對應至知識結構可作為許多延伸應用與研究之基礎，例如查詢擴展與跨語言知識交換。而典藏品項之標題名稱可用來作為典藏品特性的具體描述。本研究採用中央研究院中文詞知識庫小組所開發的斷詞系統對數位典藏國家型計畫中的典藏品標題進行斷詞[12][13]，取其中二至五字詞所構成的典藏品標題作為研究資料，共 39,765 筆。這些標題分屬於生物、考古、地質、人類學、檔案、拓片、器物、書畫、地圖與遙測、善本古籍、新聞、漢籍全文、影音、建築等十四個主題領域，其中 96%的標題詞為詞典中未收錄之項目，如馬銀花、嘉義中學與瑪瑙雙耳杯等。

對這些詞典中未收錄的標題詞而言，由於缺乏足夠之上下文資訊，不易以計算詞彙共現度的方式來進行詞彙擴展。因此，本研究在詞彙層次上對這些未知詞進行分析與處理。其中，以詞彙所包含的詞義概念進行概念擴展以找出構成詞彙的概念在知識本體上的位置並延伸是本文研究的主要方向。以標題詞所包含的概念作為基本單位則可將之與儲存知識概念的知識本體進行對應連結。而在進行概念分析上，因詞彙與概念的對應過程可能產生歧義，所以需要納入處理歧義的機制。

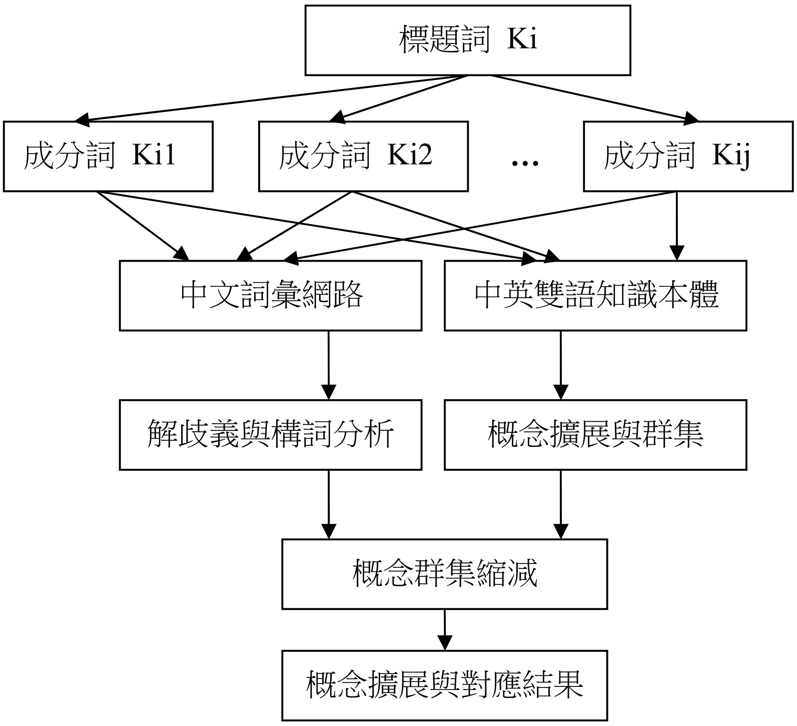
(二) 知識本體對應策略

現有已開發之知識本體可以用來作為知識交換與知識儲存之基礎，如 SUMO 與 MILO（Mid-Level Ontology）。但一般數位博物館典藏項目均具有其特殊性，因此無法順利在泛用的知識本體上找到對應節點，例如數位典藏國家型計畫中便有超過 96%的項目無法在 SUMO 中直接找到對應節點。

由於絕大部分典藏品標題無法直接對應至詞典紀錄，因此在建構整個典藏內容的知識系統時，可以選擇的方向主要有二：（1）自行建立領域特用之知識本體或（2）透過對應策略將典藏品項之標題對應至現有知識本體節點上。

依據研究目的，本研究提出一對應策略，透過中文構詞分析，先將無法直接對應之標題詞切分為較小的成分詞，再以詞形比對的方式將成分詞與中文詞彙網路及中英雙語知識本體中現有節點產生比對連結，得到一個對應結果。藉此方式可於不花費高額成本的情況下，在概念連結層次將標題詞彙擴展並對應至知識本體中。詳細步驟如圖三所示。

在成分詞的定義上，本研究將 N 字標題詞切分成二字成分、三字成分、…、(N-1)字成分，將所有的可能成分定義為成分詞集合，再透過這些成分詞與 149,751 筆 BOW 概念進行詞形比對與連結。



圖三、知識概念擴展流程

(三) 構詞分析

將標題詞切分為成分詞的對應方式具有增加對應結果數量之效果，然而亦有納入太多歧義資料的副作用。因此，為了在知識本體對應過程過濾過多的非目標結果，本研究透過 CWN 中的詳細詞義資料進行中文構詞分析解決歧義問題，將成分詞構成標題詞的組合方式以人工方式進行語言分析，歸納出兩種主要的樣式：事件驅動與實體驅動。

一、事件驅動 event driven：(專有名詞(人名) +) 動詞 (+受詞)

1. 專有名詞(人名) + 動詞

例：

羅漢松／哭了

小學生／跳舞

2. 專有名詞(人名) + 動詞 +受詞

例：

囝仔／騎／木馬

我／敬愛／國父

3. 動詞　+ 專有名詞(人名)

例：

回味／蔡惠風

找回／太魯閣

二、實體驅動 entirety driven：無動詞的

1. 修飾語　+ 中心語

表一、修飾語 + 中心語概念分析

修飾語 + 中心語概念分析	**例**
屬性／專有名詞(地名)　／中心語 屬性／專有名詞(人名)　／中心語	*垂花／蓬萊／葛* *黃文／獻公／集*
專有名詞(地名)／屬性／中心語	*城武縣／稅／銀* *湘潭縣／銀／錠*
年代／材質／中心語	*光緒年／銀／錠* *漢　／青玉／璲*
年代／性質／中心語	*全蜀／藝文／志* *漢／人物畫／像*
年代／結構／中心語	*全唐／聲律／論* *漢魏／叢書／選*
材質／功能／中心語	*光纖／連接／器* *網結草／雨／衣*
材質／樣式／中心語	*瑪瑙／雙耳／杯* *黃瑪瑙／煙／壺*
性質／中心語	*印花稅／條例* *壞死性／腸炎*

2. 中心語(+連接詞或介系詞) 中心語

例：

靈芝／和／牛樟

回顧／與／展望

四、概念擴展與群集縮減

(一) 概念擴展與群集

在透過成分詞將標題詞對應至 BOW 概念之後，便可以擴展得到一組標題詞及其成分概念，以及成分概念連結到 SUMO 上的概念節點集合。而因知識本體本身所具備的結構系統，所以可使用節點間的概念距離定義出相似概念，進而計算相似群集，使得成分詞得以擴展對應至知識本體上並透過群集的方式形成群內差異小，群間差異大的概念群集。本研究設定相似概念節點為在 SUMO 架構中距離各成分概念距離為 2 以內的概念節點。

每個經由 BOW 擴展的知識概念都可以在 SUMO 中找到對應節點，由於 SUMO 的樹狀結構，這些概念節點集合所形成的子樹便可視為一個群集。表二以 “瑪瑙雙耳杯” 作為例子，說明成分詞經過對應至 BOW 的成分概念擴展之後，藉由 SUMO 概念節點位置計算相似度距離所得到的概念群集。由本例中可知瑪瑙雙耳杯的成分詞對應到 SUMO 時，共可在知識本體樹狀結構中形成七個主要群集。

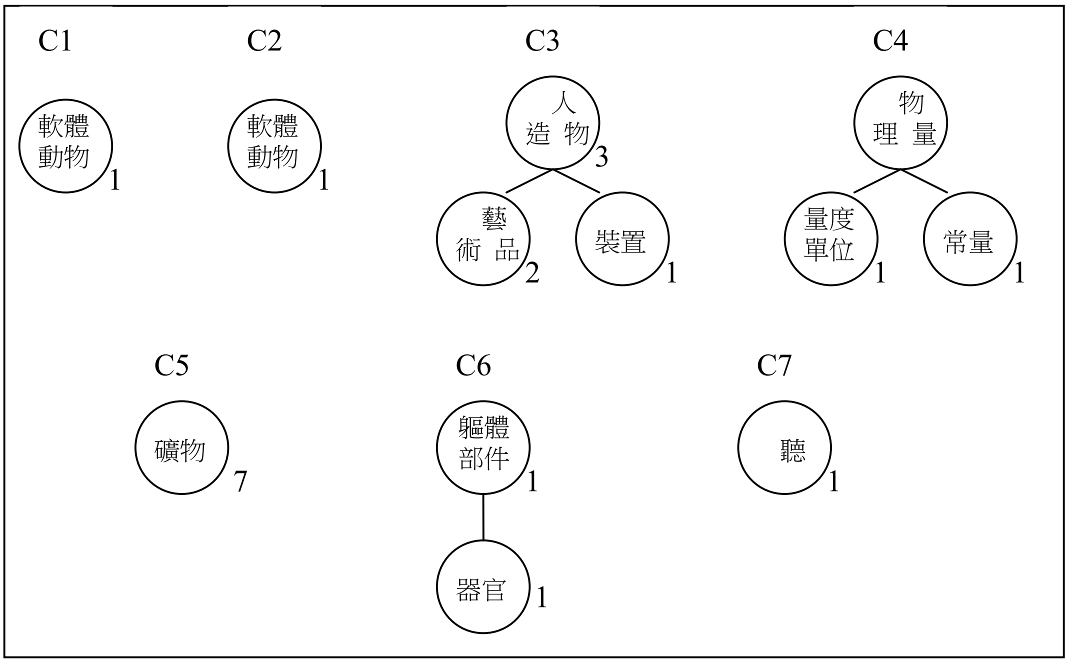
表二、概念群集範例：瑪瑙雙耳杯

瑪瑙雙耳杯					
BOW 擴展結果	WN ID	SUMO node	SUMO 中譯	SUMO 概念節點位置	概念群集
瑪瑙貝	01466296N	mollusk	軟體動物	1.1.1.1.2.4.8.14.15.9.,mollusk,軟體動物,C,_mollusk	cluster 1
瑪瑙貝	01466296N	mollusk	軟體動物	1.1.1.4.11.25.46.57.,mollusk,軟體動物,C,_mollusk	cluster 2
瑪瑙花紋的搪瓷鐵器	02169007N	art work	藝術品	1.1.1.1.2.5.13.,art work,藝術品,C,_art_work	cluster 3
瑪瑙紋搪瓷器	02169007N	art work	藝術品	1.1.1.1.2.5.13.,art work,藝術品,C,_art_work	cluster 3
雙耳瓶	02185088N	artifact	人造物	1.1.1.1.2.5.,artifact,人造物,C,_artifact	cluster 3
雙耳黏土窄瓶	02185088N	artifact	人造物	1.1.1.1.2.5.,artifact,人造物,C,_artifact	cluster 3
雙耳平底酒杯	03291208N	artifact	人造物	1.1.1.1.2.5.,artifact,人造物,C,_artifact	cluster 3
雙耳式耳機	02809404N	device	裝置	1.1.1.1.2.5.16.,device,裝置,C,_device	cluster 3

BOW 擴展結果	WN ID	SUMO node	SUMO 中譯	SUMO 概念節點位置	概念群集
瑪瑙線	09870127N	unit of measure	測量單位	1.2.3.11.50.,unit of measure,量度單位,C,_unit_of_measure	cluster 4
瑪瑙線	09870127N	constant quantity	常量	1.2.3.11.48.,constant quantity,常量,C,_constant_quantity	cluster 4
瑪瑙	10543998N	mineral	礦物	1.1.1.1.1.2.4.,mineral,礦物,C,_mineral	cluster 5
(礦)苔紋瑪瑙	10544179N	mineral	礦物	1.1.1.1.1.2.4.,mineral,礦物,C,_mineral	cluster 5
瑪瑙	10617402N	mineral	礦物	1.1.1.1.1.2.4.,mineral,礦物,C,_mineral	cluster 5
紅瑪瑙	10617402N	mineral	礦物	1.1.1.1.1.2.4.,mineral,礦物,C,_mineral	cluster 5
彩紋瑪瑙	10740932N	mineral	礦物	1.1.1.1.1.2.4.,mineral,礦物,C,_mineral	cluster 5
紅條紋瑪瑙	10740932N	mineral	礦物	1.1.1.1.1.2.4.,mineral,礦物,C,_mineral	cluster 5
纏絲瑪瑙	10740932N	mineral	礦物	1.1.1.1.1.2.4.,mineral,礦物,C,_mineral	cluster 5
雙耳的	00236774A	body part	軀體部件	1.1.1.1.2.4.9.18.,body part,軀體部件,C,_body_part	cluster 6
雙耳心	04189008N	organ	器官	1.1.1.1.2.4.9.18.23.,organ,器官,C,_organ	cluster 6
雙耳的	02509854A	hearing	聽	1.1.2.8.38.84.87.92.,hearing,聽,C,_hearing	cluster 7

(二) 群集縮減

由於概念擴展後所得到的群集往往包含太過龐雜的內容，因此有必要建立一個群集縮減的機制以濾去與原始標題差異較大的概念群集，僅保留主要相關的概念群集。此時，前述由人工歸納之構詞樣式即可扮演過濾器之角色。以瑪瑙雙耳杯一例，本研究即以構詞原則區辨出此標題詞中主要的中心語以及修飾語，再以此構詞樣式至知識本體中相對應分支中過濾出具有高相關性之對應群集，由已擴展並分群後的概念群集中保留對應正確知識本體分支之群集，而拋棄其餘群集不符合構詞原則之群集。



圖四、SUMO 子樹群集與次數分布（以“瑪瑙雙耳杯”為例）

此方法中關鍵的判斷基礎為各群集所連接的 SUMO 概念節點與構詞樣式的搭配。由“瑪瑙雙耳杯”範例可得知在“修飾語 + 中心語”的構詞樣式下，群集 3 所指涉的 SUMO 概念(人造物、裝置、藝術品)恰可表達中心語的概念。而群集 5 所指涉的 SUMO 概念（礦物），以及群集 6 所指涉的 SUMO 概念（軀體部件、器官）則用來傳達本例的修飾語概念，分別為材質與樣式。因此透過構詞樣式所進行的群集縮減便可將原本擴展至七個群集的所有概念縮減成群集 3、群集 5 以及群集 6 等三個具有代表性的群集。圖四中各 SUMO 子樹節點右側之數字為成分詞相關概念數量，作為選取群集與否的加權條件。經群集縮減後，原始標題詞的相似概念延伸詞集便是兩個縮減後概念群集中的詞彙。

值得留意的是，基於群集中包含節點的數量多寡，我們指定給群集 5 較高的重要程度，這是以詞典觀念來進行的設定，因為詞典中包含較多的項目自然是概念上重要的群集項目。而修飾語的類別經本文分析可得到有屬性、專有名詞、年代、材質、性質、結構、功能、樣式、性質等九大類。

五、結論

本研究主要目的在於探討以概念擴展的方式將原先具有領域特殊性的數位博物館典藏品標題進行成分概念分析，並對應至知識本體上的節點。再以構詞樣式將概念群集進行縮減，得到關聯概念群集。建構群集並以構詞樣式篩選關聯群集的一個好處是群集間無互斥性，可避免競爭而犧牲有代表性的群集。結果將使不同領域之典藏品能藉由標題的連結而整合成一知識系統。在研究中以人工歸納方式整理出構詞樣式並提出具有代表性的範例作為說明，可作為未來大量自動化處理之基礎。

由本研究所提出之概念與研究設計可針對儲存大量多領域知識的單一典藏機構文字資料進行概念擴展，對於數位博物館相關研究可有所助益，特別是在漢語數位博物館的資訊檢索應用上。同時亦可作為查詢擴展相關研究之參考。而由於專有名詞辨識及處理上之困難，後續研究上可導入合適的名稱辨識方法以使處理範圍能更臻完整。

參考文獻

[1] C. Leacock and M. Chodorow, “Combining local context and WordNet sense similiarity for word sense disambiguation,” In WordNet, An Electronic Lexical Database. The MIT Press, 1998.

[2] P. Resnik, “Using information content to evaluate semantic similarity,” In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995.

[3] D. Lin, “An information-theoretic definition of similarity,” In Proceedings of the 15th International Conference on Machine Learning, Madison, WI, 1998.

[4] S. Gauch and J. B. Smith, “An Expert System for Automatic Query Reformation,” Journal of the American Society for Information Science, vol. 44, no.3, 1993.

[5] H. Chen, T. D. Ng, J. Martinez, and B. R. Schatz, “A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System,” Journal of the American Society for Information Science, vol. 48, no.1, pp.17-31, 1997.

[6] Suggested Upper Merged Ontology, http://www.ontologyportal.org/

[7] I. Niles and A. Pease, “Toward a Standard Upper Ontology,” In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine, 2001.

[8] 中央研究院中英雙語知識本體詞網 The Academia Sinica Bilingual Ontological Wordnet (Sinica BOW)，http://BOW.sinica.edu.tw

[9] WordNet, http://www.cogsci.princeton.edu/~wn/

[10] C. R. Huang, E. I. J. Tseng, D. B. S. Tsai, and B. Murphy, “Cross-lingual Portability ofSemantic relations: Bootstrapping Chinese WordNet with English WordNet Relations,” Language and Linguistics, vol. 4.3, pp. 509-532, 2003.

[11] C. R. Huang, R. Y. Chang, and S. B. Lee, “Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO,” 4th International Conference on Language Resources and Evaluation (LREC2004), Lisbon. Portugal, 2004.

[12] W. Y. Ma and K. J. Chen, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff," Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, pp. 168-171, 2003.

[13] W. Y. Ma and K. J. Chen, "A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction," Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, pp. 31-38, 2003.

結合線上翻譯服務的跨語言專利檢索系統

鄧舜元　Shun-Yuan Teng
華梵大學資訊管理學系暨研究所
Department of Information Management
Huafan University
wells0609@gmail.com

邊國維　Guo-Wei Bian
華梵大學資訊管理學系暨研究所
Department of Information Management
Huafan University
gwbian@cc.hfu.edu.tw

摘要

本論文提出一個結合線上翻譯服務網站的跨語言專利檢索系統(Cross-language Patent Retrieval System)，利用適合處理不同語言的 bi-gram 索引方法，透過檢索引擎處理多語言的專利文件集，並結合網路翻譯服務系統，利用查詢翻譯的方法，將原始的查詢加以翻譯，再進行專利檢索。本系統進一步利用檢索結果，根據專利分類體系之中的 International Patent Classification (IPC)分類，可將專利文件相關的 IPC 分類列出。

目前本系統可以處理英文與日文的單語專利檢索、以及日文檢索英文專利文件與英文檢索日文專利文件兩種跨語言檢索。使用者可以選擇查詢集來源，編輯與修改查詢，選擇三種不同的查詢集翻譯方式，選擇三種不同的翻譯服務網站進行查詢翻譯，並且可以選擇檢索的欄位及專利文件集的種類，進行跨語言的專利檢索。

Abstract

In this paper, we introduced a cross-language patent retrieval system which combined the various free web translators on the internet. The bi-gram indexing method was used to deal with the multilingual patent documents, and the query translation method was used to translate the query form the source language to the target language.

Currently, this system provides the functions of the monolingual and cross-language patent retrieval in English and Japanese. The users can input the queries and use the different translation systems to process the query translation. The different fields of the query topics and various patent document sets are selected to perform the cross-language patent retrieval from Japanese to English, and vice versa.

關鍵詞：跨語言檢索、專利檢索、Bi-gram、查詢翻譯

Keywords: Cross-Language Patent Retrieval, Bi-gram, Query Translation.

一、緒論

專利文件是極為重要的科技訊息來源，長期以來一直受到研發者或企業經營者的重

視。專利文件是目前唯一完全公開技術並能使用法律來保障專利發明人權益的一種方式，正因為其揭發技術方法能迅速反映最新科技動態及研究成果，因此專利的質與量是目前衡量國家創新能力的重要指標，由於產業界的國際競爭越趨激烈，全球企業莫不積極藉由專利的保護，維持技術領先優勢與市場利益。

根據世界智慧財產權組織（World Intellectual Property Organization, WIPO）[1]報導，專利文件包含全世界 90%～95%之研發成果，而其它的技術文件（論文或期刊等）中只僅含 5%～10%之研發成果，STN International [2]也指出有 70%~90%的專利資訊，根本沒有在其他的期刊或者雜誌發表過。此外 WIPO 還指出在研究工作中若能善於應用專利文件的話可以得到縮短 60%研發時程、同時減少 40%研發經費之效益。因此，閱讀與分析專利文件成為極為重要而且為不可或缺的一項工作，而使用專利檢索是分析專利文件中極為重要的一環，因為如果檢索出來的結果不正確，那麼依照錯誤的結果所做的分類、分析以及所有的數據、圖表等，都會無法正確的反應出隱含在專利文件中的知識，由此可知專利檢索對於企業或研發者都是很重要的一項工作。

當企業或研發者在開發新產品或申請專利時候，一定會先檢索目前的現況，才能預先知道已存在的研究有哪些、驗證產品開發計畫是否有誤、是否重複研發、是否抵觸他人之專利侵權，這樣才可以節省金錢和時間上的浪費，並能有效的推展研究發展的工作。

由於專利係採屬地主義，如果專利要受到保障的話，就必須要和各國來申請專利，當企業或研發者在檢索專利的時候，期望可以使用同樣的關鍵字來檢索美國，日本，跟台灣等國的專利，並取得相關專利文件資訊，不過目前大部分的專利檢索系統並不提供跨語言的檢索方式，所以使用者必須以三個不同語言，分別到三個不同的系統查詢，才能找到所有相關的資料。但問題是，並不是所有的使用者都具備足夠的語言能力，可以使用不同的語言來檢索專利，所以如果一開始能在界定的資料範圍內，提供了涵蓋兩種以上的語言，那麼系統就可以成為一個跨語言的專利檢索系統，讓使用者使用自己的語言，也可以檢索到英文或日文的專利。

本研究提出的跨語言專利檢索，其目的結合網路各種免費的翻譯資源，開發一個能處理多個語言的跨語言專利檢索統；由於取得資料上得限制，目前我們的系統處理的文件資料包含日文與英文專利文件集。本論文第二節介紹相關的研究，第三節說明系統架構與檢索程序，第四節介紹實驗與結果，最後為結論及未來的研究方向。

二、相關研究

依我國專利法的定義是為鼓勵、保護、利用發明與創作，以促進產業發展，我國專利主分為：發明專利（提供新的做事方式或對某一問題提出新的技術解決方案的產品或方法）、新型專利（對舊事物的形狀、構造或裝置提出新的技術性創作）、新式樣專利（在事物的外觀上追求美感的新創作）三者。

專利文件是經申請並通過審查後所授予的一種權利，全世界現有 100 多個專利局公佈專利文件，每年平均公佈 100 多萬件，它既是法律文件，又是重要的技術情報。據統計有 90%發明成果的技術內容只有在專利文件中才能找到，而且專利文件還具有對發明創造說明詳盡和公佈最早的特點，透過檢索查詢專利文件，可以把握市場科技開發方向，並且可以參考他人研究成果，節省研發經費與縮短投入的時間，同時也為廣大企業在國內外貿易中瞭解有關產品技術狀況，對預防侵權提供幫助，並且研擬市場競爭策略。

表 1.專利檢索的項目

項目	檢索時機	檢索目的
專利現況檢索	在進入某一研究領域或開發新產品之前	大量檢索出相關專利、了解目前的專利概況、並在了解之後做出正確的判斷。
可專利性的檢索	有新構想擬申請專利時	對申請專利的內容和技術做一新穎性的確認，調查有無相關前案有助於專利申請的通過。
侵權的檢索	為技術、產品引進或輸出入時進行	在一項新技術或新產品進入市場之前應進行有無侵權的檢索，以避免構成對他人專利權的可能侵犯。
專利有效性檢索	在異議或舉發別人的專利是否有效而進行	檢索出相同的技術或文獻以證明別人的專利無新穎性，防止競爭者佔領某一技術領域。
技術預測檢索	為預測未來的發展	能正確的運用專利加速開發創造。
具體專利技術檢索	為解決技術問題上	專利資料中之有關技術背景與問題，常比期刊或書籍中記載要來的詳細。

由於專利檢索有其重要性，在每個階段檢索目的都不同，表 1 整理了一般企業及研發者使用專利檢索，其檢索的項目與時機，並且說明其檢索的目的。

專利文件的分類通常採用 IPC 分類, IPC 分類表目前由世界智慧財產權組織負責出版，每五年修訂一次（如表 2），而現在使用的是第八版。從 2000.1.1 至 2005.12.31 使用的第七版 IPC 碼，在專利文獻上表示為：Int. CL.7；第七版共有 8 個部（Section）、120 個主類（Class）、628 個次類（Subclass）、69,000 個主目（Group）及次目（Subgroup）。一個完整之分類碼必須由代表部、主類、次類、主目或次目之符號結合構成（如圖 1），我們使用一個範例來說明 H04L 12/44 所代表的意義如下：

部：	H	電學
類：	H04	電氣通信技術
次類：	H04L	數位資訊之傳輸，例如電報通信
主目：	H04L 12/00	數據交換網路
次目：	H04L 12/44	星形或樹狀網路

表 2 國際專利分類表版本

版本	有效期間
第一版	1968/09/01～1974/06/30
第二版	1974/07/01～1979/12/31
第三版	1980/01/01～1984/12/31
第四版	1985/01/01～1989/12/31
第五版	1990/01/01～1994/12/31
第六版	1995/01/01～1999/12/31
第七版	2000/01/01～2005/12/31
第八版	2006/01/01 生效

（資料來源：國際專利分類檢索系統(第 8 版)使用指南）

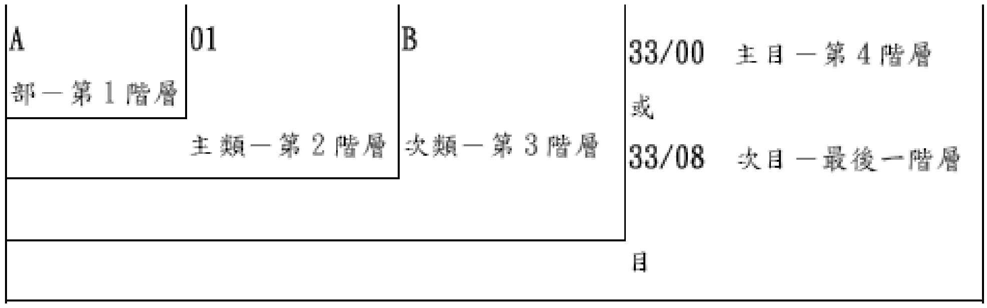


圖 1 完整之分類號組成

（資料來源：國際專利分類檢索系統(第 8 版)使用指南）

跨語言資訊檢索（Cross-Language Information Retrieval, CLIR）是使用某一種語言來查詢另外一種語言的文件，不過由於語言上的差異，通常都需要將查詢（Query）轉換成跟文件一樣的語言。目前大多數的使用者會在網際網路上使用搜尋引擎來查詢所需要的資料，當我們輸入中文的查詢字，執行檢索後我們可以發現結果可以包含其他語系與 Query 有關的相關資訊，這就是因為搜尋引擎會自動將您的輸入的 Query 翻譯成其他的語系並執行檢索的動作，由此可知跨語言的資訊檢索可以讓使用者方便使用自己熟悉的語言來檢索其他語系的文件。在跨語言資訊檢索相關的研究中，大部分採用的方法可歸納成文件翻譯（Document Translation）和查詢翻譯（Query Translation）兩種，兩種技術的目標都是要將查詢和文件的語言轉為一致。

使用文件翻譯的檢索方式必須將所有文件都翻譯和 Query 相同的語系，優點是文件與查詢都是使用相同的語言，使用者可以直接閱讀，缺點是翻譯所有的文件必須耗費大量的時間。

查詢翻譯需先將 Query 翻譯成和文件相同的語言，目前在跨語言資訊檢索中被廣泛的使用，此方法的好壞取決於 Query 是否被正確翻譯，而翻譯的方法有幾種被提出；有字典翻譯（Dictionary-based translation）方法[4]，語料庫翻譯（Corpus-based translation）方法[5]，混和（Hybrid）方法[6]，網路翻譯擷取（web-based translation extraction）方法[7]；由於網際網路上的資源眾多，很多的專家學者利用此優勢，使用網路查詢後再使用機率統計其結果，最後選擇最佳的翻譯當結果， Zhang 等人[8]指出使用網路擷取翻譯方式可以降低詞彙涵蓋度的問題。

綜觀以上方法，主要的目標都是將查詢和文件轉化成相同的語言，再進行資訊檢索，查詢的文字中某些關鍵字詞若無法被正確地翻譯，將會影響跨語資訊檢索的準確性。

在跨語言資訊檢索中，大部分的亞洲語言並不像英文一樣在每個單詞間都有分隔符號，因此斷詞這個步驟就顯得格外的重要，Shi＆Nie[9]針對亞洲語系的斷詞使用不同方法，指出在處理日文斷詞方法採用 bi-gram 加上 uni-gram 可得到更好的效用。

NTCIR（NACSIS Test Collections for IR）計畫[10]是由日本國家科學資訊系統中心（National Center for Science Information Systems, NACSIS）所策劃主辦的，其目的是希望能建立一個大型日文標竿測試集，作為資訊檢索與自然語言處理研究的基礎資料。NTCIR 從 1999 年開始舉辦，至今已經邁入第七屆，從第三屆（2001-2002）開始舉辦了第一次的專利檢索評比，提供大型的文件集，包含二年的日文專利全文、五年的日本專

利摘要及五年日本專利的英文摘要，檢索題目有英、日、中、韓等四種語言，作為跨語言的專利檢索。由於專利檢索有不同的目的：技術調查（technology survey）、前案檢索（invalidity search）、專利地圖（patent map）等，假使查詢的主題相同但不同的檢索目的就會出現不同相關專利的結果，需要不同的檢索模式與技巧， NTCIR 的專利檢索從技術調查（technology survey）、前案檢索（invalidity search）、專利地圖（patent map）、專利分類（patent classification）到今年的專利採礦（patent Mining）、專利翻譯（patent translation）每年都會有不同的任務。

根據 NTCIR-6 有關專利檢索的研究[11, 12, 13]，要提昇專利檢索的精確度，除了原本的查詢欄位外，必須加入其他的相關欄位，甚至把整份專利文件都當檢索的條件，都可以增加檢索的查全率（recall）及查準率（precision）[11]。

三、系統描述

本系統架構如圖 2 所示，首先將專利文件集資料，經過模組程式過濾不需要的特殊字元、控制碼…等後，採用 bi-gram 的方式來處理日文文件，建置索引資料庫。查詢集利用三種線上翻譯系統翻譯為目的語言，檢索模型使用 TF-IDF（term frequency-inverse document frequency）的方法將檢索到的文件評分並加以排序，分類模組程式將檢索的結果進一步作 IPC code 自動分類。圖 3 為處理日文查英文之跨語言專利檢索的過程，先將日文查詢集經線上翻譯系統翻譯成英文，將翻譯過後的查詢集進行詞彙與斷詞的處理，最後進行檢索作業；圖 4 是處理英文查日文之跨語言專利檢索的過程。

我們使用 Lucene[14]作為專利檢索的搜尋引擎，Lucene 是 apache 軟體基金會 jakarta 項目組的一個子項目，是一個使用 JAVA 語言開發且是開放原始碼的全文檢索引擎工具，提供資訊檢索所需要的重要功能：建立索引（index）和檢索（retrieval）。Lucene 針對軟體開發人員提供一個簡單易用的工具包，可以建立完整的全文檢索引擎，Lucene 除了有 JAVA 的版本之外，也陸陸續續的被開發成其他不同的版本，如 C#、C++、Delphi、Perl、Python、Ruby 和 PHP 等，本實驗所建立的專利檢索系統採用 C#的版本。

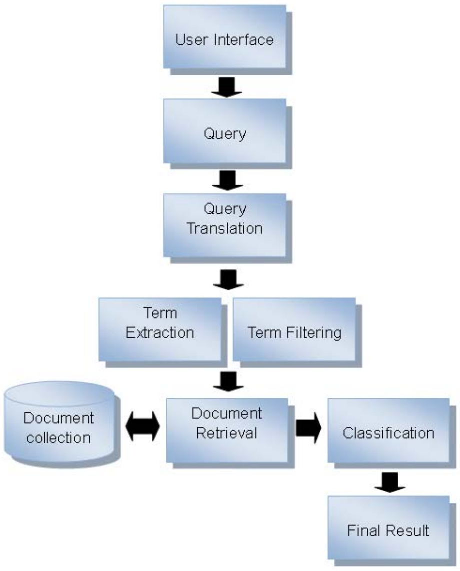


圖 2 系統架構圖

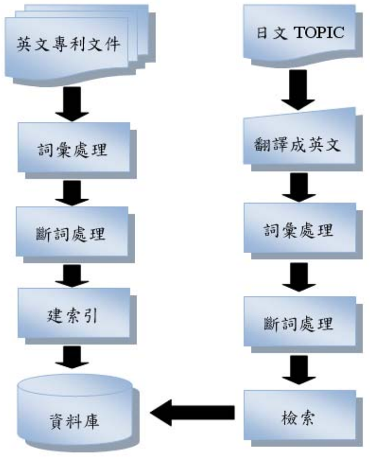


圖 3　日文查英文之跨語言專利檢索處理

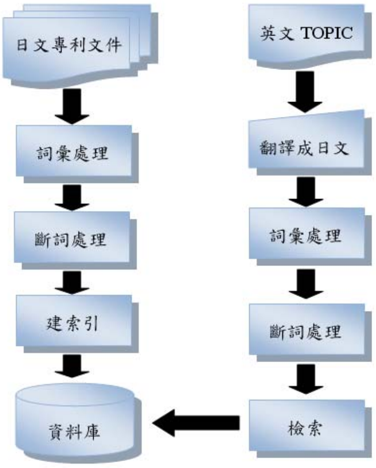


圖 4　英文查日文之跨語言專利檢索處理

3.1 詞彙處理

處理亞洲語系文件的首要工作是詞彙處理，因為大部分的亞洲語言並不像英文一樣在每個單詞間都有分隔符號，一般採用 N-gram 的技術處理不同的亞洲語言[15,16,17]，其中使用 bi-gram 的方法又比使用 uni-gram 的方式好，因此在本實驗中對於日文的斷詞方式採用 bi-gram 的方法，表 3 為日文句子採用 uni-gram 及 bi-gram 斷詞後的結果。

在資訊檢索時，標點符號、特殊字元與停用字（Stop Word）都是無意義的，因此在建立索引和檢索之前需把查詢集與文件集內的這些相關字元給去除。由於使用 Lucene 建立索引時，系統會自動將標點符號與停用字給過濾掉，因此我們只要注意特殊字元的處理即可。

表 3　日文採用 uni-gram 與 bi-gram 斷詞後結果

<table>
<tr><td>原始句</td><td>チリ産い貝の接着タンパク質の合成</td></tr>
<tr><td>uni-gram</td><td>チ|リ|産|い|貝|の|接|着|タ|ン|パ|ク|質|の|合|成</td></tr>
<tr><td>bi-gram</td><td>チリ|リ産|産い|い貝|貝の|の接|接着|着タ|タン|ンパ|パク|ク質|質の|の合|合成</td></tr>
</table>

3.2 查詢翻譯處理

本系統的查詢翻譯是將原始語言的查詢集利用不同的線上翻譯網站將其翻譯成目標語言，再進行單語專利檢索；例如：英文的查詢集經過線上翻譯系統翻譯成日文後，再進行日文專利檢索。由於不同的線上翻譯系統所翻譯的結果均不相同，因此我們採用了 Google Translation[18]、Yahoo Babel Fish[19]及 Excite[20]三種不同的線上翻譯系統來彌補翻譯不足的問題。

在實驗中，我們使用的線上翻譯系統並非完全針對資訊檢索的查詢集(Topics)而設計的，所以若我們直接把查詢集的文件傳送到線上翻譯系統，發現傳回來的結果與文件格式會有所錯誤，表 4 列出直接使用線上翻譯系統時可能產生之的問題，這些錯誤必須人工檢視翻譯後的結果，並修正其錯誤，方可進行後續的檢索處理。

我們採取另一種的方式可以有效的讓這些錯誤發生的問題降低。首先先將查詢集的文件轉成 XML 的文件格式，因為 XML 的文件具有欄位的特性，我們將每個欄位分別傳送到翻譯系統，再按欄位依序取回翻譯的結果，雖然這樣可以正確的取回結果，不過發現部份結果會帶有 HTML tag，由於我們不需要這些 HTML tag，將這些 HTML tag 去除後，得到我們需要的翻譯結果。

本系統使用三種線上翻譯系統進行查詢翻譯，分別為 Google Translation、Yahoo 線上翻譯、與 Excite 線上翻譯，分別介紹於下:

Google Translation 是一個免費的線上翻譯網站，提供多種語系的翻譯，其中也包含使用英文對日文的翻譯服務，Google 線上翻譯系統有別於其他翻譯系統的作法，是採用統計式的作法，由電腦進行網頁比對找出翻譯機率，當作文件翻譯之用。

表 4　　直接使用線上翻譯可能產生之格式錯誤

翻譯後格式錯誤結果之範例	錯誤說明
<TOPIC> <TOPIC-ID> 100 </ TOPIC-ID> PB <TITLE> sound transmission in mobile communications processing system </ TITLE>	翻譯後內容與 Tag 的位置不正確
<TOPIC-ID> 101 </ TOPIC-ID> <TITLE> Artificial boundary-derived lipids and proteins and biological hybrid by RIPOSOMUWAKUCHIN </ TITLE>	Tag 內有多餘的空白產生
<topic-id> 100 < / トピック- ID を> <title> dtmf （デュアルトーンマルチ周波数）伝送方式は、移動体通信システム< / タイトル>	Tag 的文字也被翻譯了
A. B hepatitis, B. genetic engineering techniques, C. vaccine, vaccination	傳回的結果會增加一些不必要的 Html Tag

表 5 維基百科與 Google、Yahoo 翻譯比較表

	人名	專有名詞
	日文/英文對照	日文/英文對照
維基百科	ヘルベルト・フォン・カラヤン / Herbert von Karajan	航空交通管制 / Air traffic control
Google 翻譯	ヘルベルトフォンカラヤン / Herbert von Karajan	航空管制 / Air traffic control
Yahoo 翻譯	ハーバートフォン Karajan / [heruberuto] phone Karajan	航空管制 / Flight control

在實際使用上， Google Translation 對於專有名詞、人名、術語等的翻譯上有其獨特的地方，我們在維基百科（Wikipedia）上隨機選擇一個人名與專有名詞來作比較，在維基百科上的日文與英文別為ヘルベルト・フォン・カラヤン、Herbert von Karajan，我們使用Google翻譯的結果為ヘルベルトフォンカラヤン、Herbert von Karajan而 Yahoo 翻譯的結果為ハーバートフォン Karajan、[heruberuto] phone Karajan，其結果如表5所示。

Yahoo 翻譯網站提供多國語系的翻譯，它是採用 Alta Vista 和 Systran 合作提供的翻譯服務「Babel Fish」。Yahoo 日本網站使用的翻譯服務與 Babel Fish 是不同的技術，而且僅提供英文、中文、日文及韓文的翻譯服務。由於這兩個網站都有提供英日相互翻譯的功能，我們選擇使用 Yahoo 翻譯作為我們系統的其中一種翻譯系統。

Excite 翻譯網站提供的翻譯語系較上述兩種系統為少，它僅提供日文對英文、英文對日文、日文對中文、中文對日文、日文對韓文、韓文對日文等六種翻譯方式，但此網站是多數人在翻譯日文時推薦使用的線上翻譯網站，因此也納入作為其中一種的翻譯系統。

3.3 分類處理

本系統的分類處理採用下列步驟決定 IPC 分類碼：

(1) 將專利文件集建立索引

(2) 使用 Topics (Query)進行檢索

(3) 對步驟二取出之前 3000 份專利文件分別抽取相對應之 IPC code

(4) 步驟三得到之 IPC code 分別使用下列公式計算 Score

Score(IPC)=Σ(專利文件對於 Query 的相似度分數)

例如：從檢索結果中取出的 Top 3000專利文件，當中含有“A61B_5_02”這個 IPC code 的專利文件編號分別為 PATENT-US-GRT-2000-06152884 、PATENT-US-GRT-1993-05181521與 PATENT-US-GRT-1998-05772600，對於 Query 的相似度分數分別為21.264、17.724,、20.125，則：

Score (A61B_5_02) = 21.264 + 17.724 + 20.125 = 59.113

(5) 根據得分高低排序輸出 IPC codes

四、實驗

實驗資料來源是採用 NTCIR 提供的文件集 (Document Sets)與查詢集(Topics)，資料包含 1993 年到 2002 年的 Unexamined Japanese patent applications、USPTO patent data、Patent Abstracts of Japan 與 NTCIR-1 與 NTCIR-2 CLIR task Test Collection 等相關專利文件，表 6 為這些文件集的數量與所佔儲存容量。

我們使用 Lucene 建立文件集的索引，由於日文的文件集都是採用日語語系的 EUC 文件編碼方式，造成讀取文件與建立索引時有亂碼產生的問題，由於 Lucene 支援 UTF-8 的編碼方式，所以預先將所有的日文文件集使用工具將 EUC 編碼轉換成 UTF-8 編碼，去除標點符號、特殊字元與停用字，再使用 bi-gram 的斷詞處理，接下來使用 Lucene 建立索引檔，建立完索引檔後，便可以使用 Lucene 提供的檢索功能檢索資料。

表 7 所列的是使用日文專利文件集建立索引所花的時間及所佔儲存容量大小，表 8 是使用英文專利文件集建立索引所花的時間及所佔儲存容量大小，建立索引時所使用機器為：Pentium 4 2.66GHz，記憶體大小為 1GB，作業系統為 Microsoft Windows XP。

表 6 NTCIR-7 的文件集數量

類別	語言	文件數	容量（MB）
NTCIR-1	日文	332,918	312
	英文	187,080	218
NTCIR-2	日文	403,240	600
	英文	134,978	200
Unexamined Japanese patent applications	日文	3,496,252	96,768
Patent Abstracts of Japan	英文	2,543,488	4,102
USPTO patent data	英文	1,315,470	53,351

表 7 NTCIR-7 的日文文件集索引

類別	文件數	term 數量	Index 容量	Index 時間
NTCIR-1	332,918	1,596,747	312 MB	6.98 小時
NTCIR-2	403,240	2,021,914	710 MB	9.56 小時
Unexamined Japanese patent applications	3,496,253	7,596,840	46,445 MB	308.08 小時

表 8　　NTCIR-7 的英文文件集索引

類別	文件數	term 數量	Index 容量	Index 時間
NTCIR-1	187,080	1,033,575	211 MB	4.11 小時
NTCIR-2	134,978	785,607	227 MB	2.78 小時
Patent Abstracts of Japan	2,543,488	3,191,893	676 MB	22.07 小時
USPTO patent data	1,315,470	2,653,186	1942 MB	27.69 小時

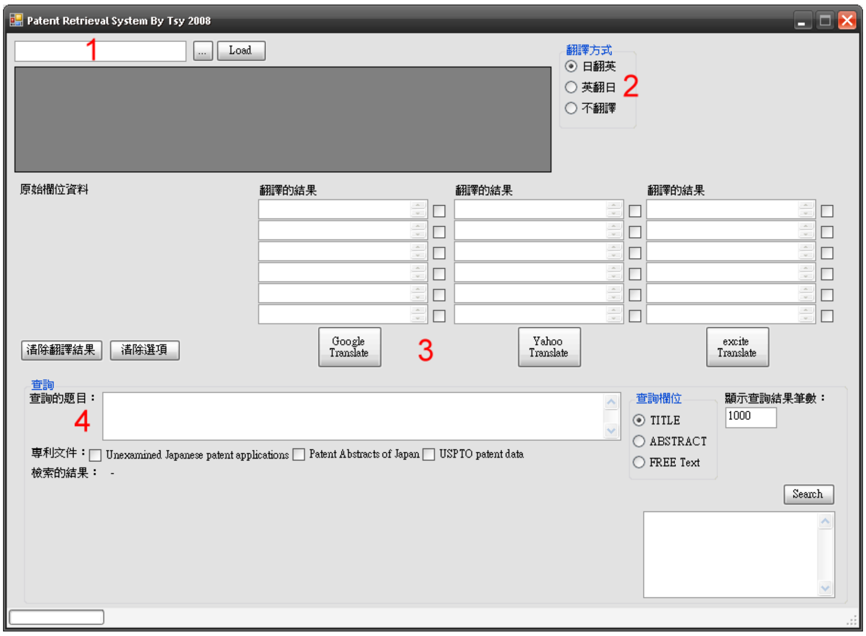
在檢索之前，首先把查詢集的文件均轉成 UTF-8，再將檔案格式改為 XML 的格式，接下來將查詢集裡的特殊字給去除，最後得到我們所需的查詢集。系統畫面如圖 5 所示，主要步驟包括：

(1) 首先載入查詢集

(2) 根據要翻譯的方式，選擇日翻英、英翻日或者不翻譯

(3) 接下選擇使用的線上翻譯工具

(4) 系統根據使用者選擇的檢索內容、檢索的專利文件集（日文專利集或英文專利集）、檢索的欄位（TITLE、ABSTRACT、Free Text）、檢索結果的顯示筆數，顯示最後的檢索結果



翻譯方式：● 日翻英　○ 英翻日　○ 不翻譯

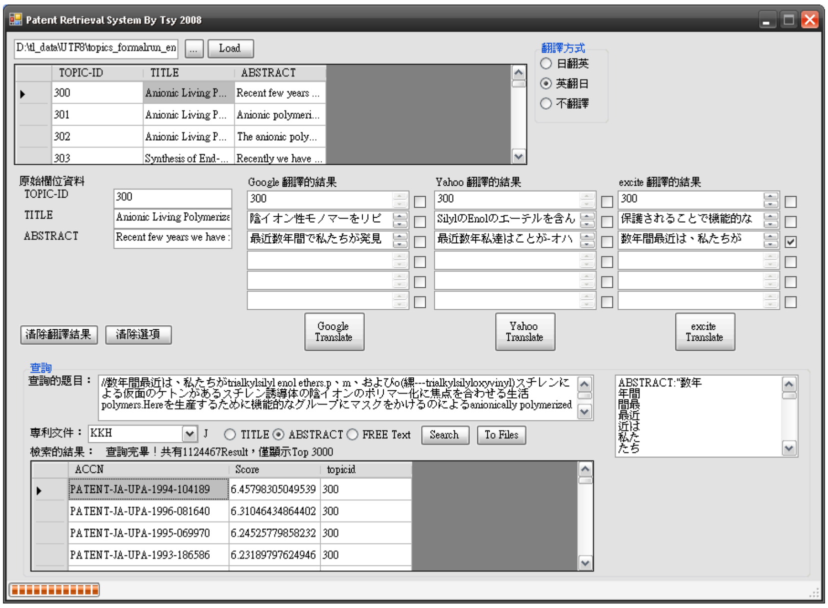
專利文件：□ Unexamined Japanese patent applications □ Patent Abstracts of Japan □ USPTO patent data

檢索的結果： -

查詢欄位：● TITLE　○ ABSTRACT　○ FREE Text

顯示查詢結果筆數：1000

圖 5 系統主畫面



D:\tl_data\UTF8\topics_formalrun_en

翻譯方式：○ 日翻英 ● 英翻日 ○ 不翻譯

TOPIC-ID	TITLE	ABSTRACT
300	Anionic Living P...	Recent few years ...
301	Anionic Living P...	Anionic polymeri...
302	Anionic Living P...	The anionic poly...
303	Synthesis of End-...	Recently we have ...

原始欄位資料		Google 翻譯的結果		Yahoo 翻譯的結果		excite 翻譯的結果	
TOPIC-ID	300	300	□	300	□	300	□
TITLE	Anionic Living Polymeriz	陰イオン性モノマーをリピ	□	SilylのEnolのエーテルを含ん	□	保護されることで機能的な	□
ABSTRACT	Recent few years we have :	最近数年間で私たちが発見	□	最近数年私達はことが-オハ	□	数年間最近は、私たちが	☑
			□		□		□
			□		□		□
			□		□		□

查詢

查詢的題目：//数年間最近は、私たちがtrialkylsilyl enol ethers.p、m、およびo(縲---trialkylsilyloxyvinyl)スチレンによる仮面のケトンがあるスチレン誘導体の陰イオンのポリマー化に焦点を合わせる生活polymers.Hereを生産するために機能的なグループにマスクをかけるのによるanionically polymerized

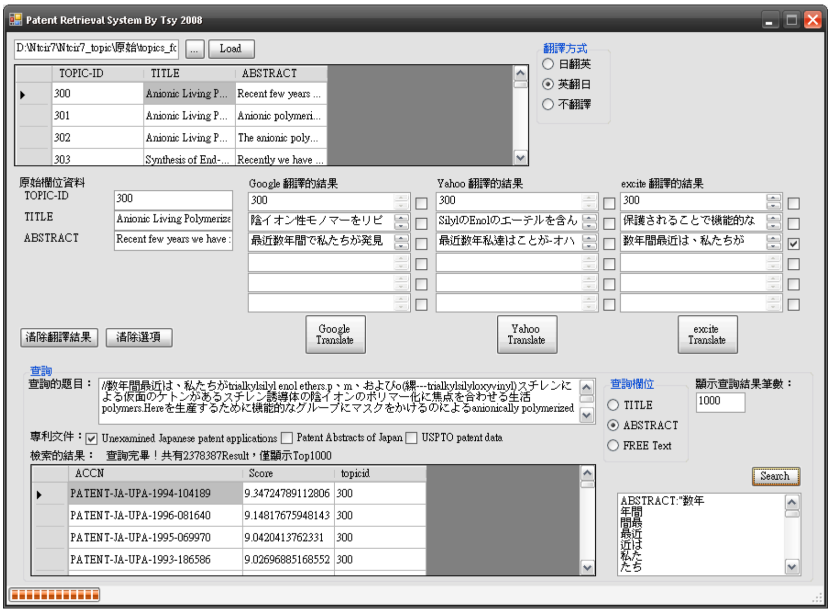
ABSTRACT:"数年
年間
間最
最近
近は
私た
たち

專利文件：KKH　J　○ TITLE ● ABSTRACT ○ FREE Text

檢索的結果：　查詢完畢！共有1124467Result，僅顯示Top 3000

ACCN	Score	topicid
PATENT-JA-UPA-1994-104189	6.45798305049539	300
PATENT-JA-UPA-1996-081640	6.31046434864402	300
PATENT-JA-UPA-1995-069970	6.24525779858232	300
PATENT-JA-UPA-1993-186586	6.23189797624946	300

圖 6　　英文查詢集檢索日文專利文件集之範例



D:\Ntcir7\Ntcir7_topic\原始\topics_fc

翻譯方式：○ 日翻英 ● 英翻日 ○ 不翻譯

TOPIC-ID	TITLE	ABSTRACT
300	Anionic Living P...	Recent few years ...
301	Anionic Living P...	Anionic polymeri...
302	Anionic Living P...	The anionic poly...
303	Synthesis of End-...	Recently we have ...

原始欄位資料		Google 翻譯的結果		Yahoo 翻譯的結果		excite 翻譯的結果	
TOPIC-ID	300	300	□	300	□	300	□
TITLE	Anionic Living Polymeriz	陰イオン性モノマーをリピ	□	SilylのEnolのエーテルを含ん	□	保護されることで機能的な	□
ABSTRACT	Recent few years we have :	最近数年間で私たちが発見	□	最近数年私達はことが-オハ	□	数年間最近は、私たちが	☑
			□		□		□
			□		□		□
			□		□		□

查詢

查詢的題目：//数年間最近は、私たちがtrialkylsilyl enol ethers.p、m、およびo(縲---trialkylsilyloxyvinyl)スチレンによる仮面のケトンがあるスチレン誘導体の陰イオンのポリマー化に焦点を合わせる生活polymers.Hereを生産するために機能的なグループにマスクをかけるのによるanionically polymerized

查詢欄位：○ TITLE ● ABSTRACT ○ FREE Text

顯示查詢結果筆數：1000

專利文件：☑ Unexamined Japanese patent applications □ Patent Abstracts of Japan □ USPTO patent data

檢索的結果：　查詢完畢！共有2378387Result，僅顯示Top1000

ACCN	Score	topicid
PATENT-JA-UPA-1994-104189	9.34724789112806	300
PATENT-JA-UPA-1996-081640	9.14817675948143	300
PATENT-JA-UPA-1995-069970	9.0420413762331	300
PATENT-JA-UPA-1993-186586	9.02696885168552	300

ABSTRACT:"数年
年間
間最
最近
近は
私た
たち

圖 7 日文查詢集檢索英文專利文件集之範例

由於要提昇專利檢索的精確度，除了原本的查詢欄位外，必須加入其他的相關欄位，甚至把整份專利文件都當檢索的條件，都可以增加檢索的查全率（recall）及查準率（precision）[11]，所以我們在設計系統時，也將全欄位納入我們的檢索條件內，圖 6 為使用英文查詢集來檢索日文專利文件，首先載入英文查詢集，選擇英翻日，選擇查詢題目編號為 300 的題目，接者使用 Excite 線上翻譯系統，使用 ABSTRACT 為檢索條件，專利文件集則選擇日文專利文件集，檢索欄位為 ABSTRACT，結果顯示筆數為 1000 筆，點選 search 則會顯示檢索結果，其中 ACCN 為專利文件編號、Scroe 為適合度分數；圖 7 為使用日文查詢集來檢索英文專利文件，使用日文查詢集，日翻英方式、查詢題目編號 1393、Google 線上翻譯、TITLE 為檢索條件、Free Text 為檢索欄位、英文專利文件集、顯示筆數為 1000 筆。

表 9 日文查詢集編號 1393 的檢索結果

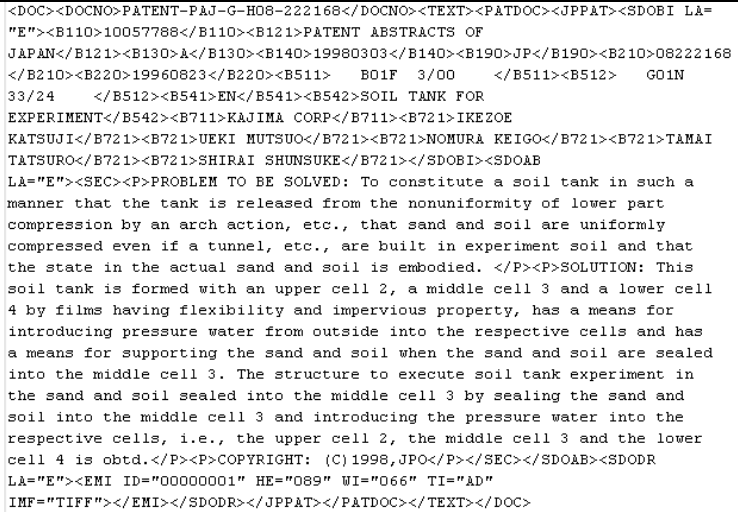
原始 Query 的 Title	鉛直遮水壁の封じ込め効果に関する透水土槽実験（その3）－透水土槽実験と事後解析－
翻譯後的結果	Vertical wall impervious soil permeability effects of the containment tank experiment (3) - and subsequent laboratory analysis permeable earth tank --
檢索結果 Top 1 的內容	 <DOC><DOCNO>PATENT-PAJ-G-H08-222168</DOCNO><TEXT><PATDOC><JPPAT><SDOBI LA="E"><B110>10057788</B110><B121>PATENT ABSTRACTS OF JAPAN</B121><B130>A</B130><B140>19980303</B140><B190>JP</B190><B210>08222168</B210><B220>19960823</B220><B511> B01F 3/00 </B511><B512> G01N 33/24 </B512><B541>EN</B541><B542>SOIL TANK FOR EXPERIMENT</B542><B711>KAJIMA CORP</B711><B721>IKEZOE KATSUJI</B721><B721>UEKI MUTSUO</B721><B721>NOMURA KEIGO</B721><B721>TAMAI TATSURO</B721><B721>SHIRAI SHUNSUKE</B721></SDOBI><SDOAB LA="E"><SEC><P>PROBLEM TO BE SOLVED: To constitute a soil tank in such a manner that the tank is released from the nonuniformity of lower part compression by an arch action, etc., that sand and soil are uniformly compressed even if a tunnel, etc., are built in experiment soil and that the state in the actual sand and soil is embodied. </P><P>SOLUTION: This soil tank is formed with an upper cell 2, a middle cell 3 and a lower cell 4 by films having flexibility and impervious property, has a means for introducing pressure water from outside into the respective cells and has a means for supporting the sand and soil when the sand and soil are sealed into the middle cell 3. The structure to execute soil tank experiment in the sand and soil sealed into the middle cell 3 by sealing the sand and soil into the middle cell 3 and introducing the pressure water into the respective cells, i.e., the upper cell 2, the middle cell 3 and the lower cell 4 is obtd.</P><P>COPYRIGHT: (C)1998,JPO</P></SEC></SDOAB><SDODR LA="E"><EMI ID="00000001" HE="089" WI="066" TI="AD" IMF="TIFF"></EMI></SDODR></JPPAT></PATDOC></TEXT></DOC>

表 9 為使用日文 Topic 編號為 1393 的 TITLE 當作 Query 的題目，經過 Google Translation 將日文翻譯為英文，查詢英文專利文件的結果我們取出 Top 1 的內容並顯示。表 10 為使用日文查詢集編號 305 時，三種線上翻譯系統的結果比較，以使用 Google Translation 的準確度較佳。

由於專利文件的資料相當的龐大，我們花了很多的時間在轉換文件編碼，處理原始資料上一些沒用的資訊，如特殊字、標點符號…等，當這些處理完畢後，才能開始建立索引。而在查詢翻譯處理中，使用的線上翻譯系統並非完全針對資訊檢索的查詢集(Topics)而設計的，在翻譯結果上會產生格式上的錯誤，我們採取的方式可以有效解決這些格式問題。

根據專利分類體系之中的 International Patent Classification (IPC)進行自動分類，因此當完成檢索之後，需將檢索結果加入 IPC code 並加以評分，得到最後之結果。分類子系統畫面如圖8所示，主要的步驟包括：

(1) 載入檢索結果

(2) 選擇要加入 IPC code 的種類

(3) 進行比對並加入 IPC code

例如載入編號300的檢索結果，IPC code 的分類方式為 PAJ＆USPTO，點選執行後產生加入 IPC code 後的結果（如圖9）；其中 topicid 為檢索題目的編號、IPC 為 IPC code、Score 為分數、IPC-rank 為順序。

表 10　　日文查英文項目，日文查詢集編號 305 的翻譯結果

	末端に官能基を有するポリマーの合成〔VIII〕官能基を保護したω-ハロ化合物とアニオンリビングポリスチレン、ポリイソプレンの反応
Google	Functionalized end of synthetic polymers VIII] [ω-functional group to protect the compound and ANIONRIBINGUPORISUCHIREN Hello, polyisoprene reaction
Yahoo	ω-[haro] chemical compound and the anionic living polystyrene which protect the synthetic (viii) functional group of the polymer which possesses the functional group in end, the reaction of the polyisoprene
Excite	ω-hello the reaction of the compound, the anion living polystyrene, and polyisoprene that protects synthesis VIII functional group of Polymer that has the functional group in the end.

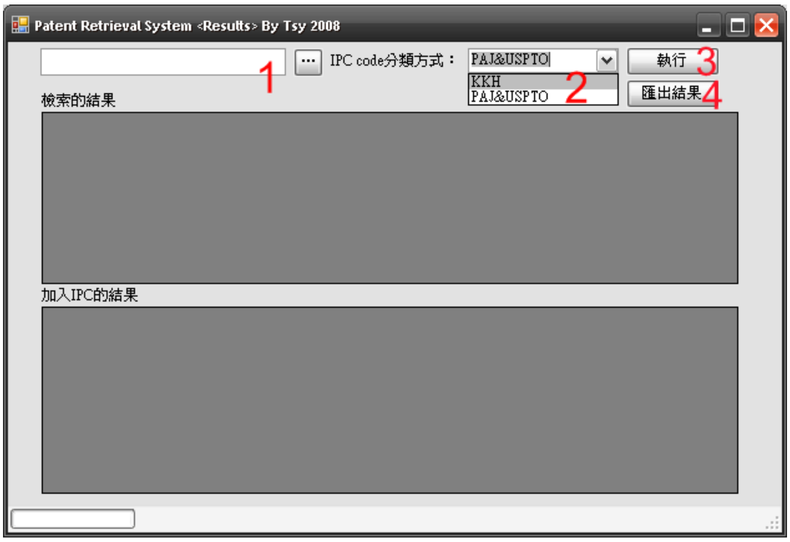
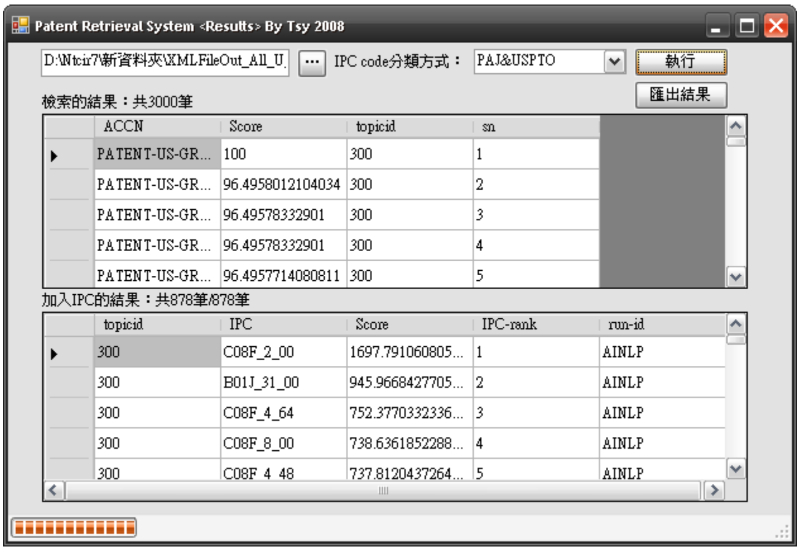


圖 8　　分類子系統的主畫面


Patent Retrieval System <Results> By Tsy 2008
D:\Ntcir7\新資料夾\XMLFileOut_All_U
...
IPC code分類方式：
PAJ&USPTO
執行
匯出結果

檢索的結果：共3000筆

ACCN	Score	topicid	sn
PATENT-US-GR...	100	300	1
PATENT-US-GR...	96.4958012104034	300	2
PATENT-US-GR...	96.49578332901	300	3
PATENT-US-GR...	96.49578332901	300	4
PATENT-US-GR...	96.4957714080811	300	5

加入IPC的結果：共878筆/878筆

topicid	IPC	Score	IPC-rank	run-id
300	C08F_2_00	1697.791060805...	1	AINLP
300	B01J_31_00	945.9668427705...	2	AINLP
300	C08F_4_64	752.3770332336...	3	AINLP
300	C08F_8_00	738.6361852288...	4	AINLP
300	C08F_4_48	737.8120437264...	5	AINLP

圖 9　　檢索結果加入 IPC code 之範例

五、結論

本論文提出一個 CPRS 跨語言專利檢索系統（Cross-language Patent Retrieval System）的架構，採用適合處理不同語言的 bi-gram 索引方法，透過 Lucene 檢索引擎處理多語言的專利文件集（Document Sets）與查詢集（Topics），並結合網路翻譯系統，利用查詢翻譯的方法，將原始的查詢加以翻譯，再進行專利檢索。並且建置一個能處理多語專利文件的跨語言專利檢索系統，進一步利用檢索結果，根據專利分類體系之中的 International Patent Classification (IPC)分類，得到相關的之 IPC 分類。

目前我們的系統可以處理英文與日文的單語檢索、以及日文檢索英文專利文件與英文檢索日文專利文件兩種跨語言檢索。使用者可以選擇三種不同的翻譯方式來翻譯查詢集，並且可以選擇檢索的欄位及專利文件集的種類，進行跨語言專利檢索。

致謝

本論文的完成感謝日本 NTCIR（NACSIS Test Collections for IR）提供專利文件，以及華梵大學先進製造研究中心提供部份的研究經費支援。

參考文獻

[1] WIPO, http://www.wipo.int/portal/index.html.

[2] STN International, http://www.stn-international.de.

[3] 國際專利分類檢索系統(第 8 版)使用指南， URL: http://newweb.tipo.gov.tw/ch/MultiMedia_FileDownload.ashx?guid=5dc74ecb-4be5-42c7-ada2-dcd37ad908fb

[4] Ballesteros, L. and Croft, W.B. “Dictionary–based Methods for Cross-Lingual Information Retrieval.” Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, 791-801, 1996.

[5] Yang, C.C. and LI, K.W. “Mining English/Chinese Parallel documents from the world wide web”, Proceedings of the 11th international World Wide Web Conference, Honolulu, Hawaii, May, 188-192, 2002.

[6] Bian, G.W. and Chen H.H. “The Study of Query Translation and Document Translation in a Cross-Language Information Retrieval System” Ph.D. Dissertation, National Taiwan University, Taipei, Taiwan, 1999.

[7] Cheng, C.C.; Shue, R.J.; Lee, H.L.; Hsieh, S.Y.; Yeh, G.C. and Bian, G.W. “AINLP at NTCIR-6: Evaluations for Multilingual and Cross-Lingual Information Retrieval”, Proceedings of NTCIR-6 Workshop, Japan, 2007.

[8] Zhang, Y.; Vines, P. and Zobel, J. “Chinese OOV translation and post-translation query expansion in chinese--english cross-lingual information retrieval”, ACM Transactions on Asian Language Information Processing, Vol.4, No.2, June, 55-77, 2005.

[9] Shi, L. and Nie, J.Y. “Using Unigram and Bigram Language Models for Monolingual and Cross-Language IR", Proceedings of NTCIR-6 Workshop, 2007.

[10] Makoto, I.; Atsushi, F. and Noriko, K. “Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task”, Proceedings of NTCIR-6 Workshop, 2007.

[11] Fujii, A. “Integrating Content and Citation Information for the NTCIR-6 Patent Retrieval Task”, Proceedings of NTCIR-6 Workshop, 2007.

[12] Mase, H. and Iwayama, M. “NTCIR-6 Patent Retrieval Experiments at Hitachi”, Proceedings of NTCIR-6 Workshop, 2007.

[13] Tseng, Y.H. ; Tsai, C.Y. and Juang, D.W. “Invalidity Search for USPTO Patent Documents Using Different Patent Surrogates”, Proceedings of NTCIR-6 Workshop, 2007.

[14] Lucene, http://lucene.apache.org/java/docs/index.html

[15] 鄭貞信，「英中日韓文的跨語言檢索之比較」，華梵大學資訊管理學系碩士論文，民國九十六年。

[16] Kwok, K-L and Dinstl N. “NTCIR-6 Monolingual Chinese and English-Chinese Cross Language Retrieval Experiments using PIRCS”, Proceedings of NTCIR-6 Workshop, Japan, 2007.

[17] Su, C.Y.; Lin, T.C. and Wu, S.H. “Using Wikipedia to Translate OOV Term on MLIR”, Japan, Proceedings of NTCIR-6 Workshop, 2007.

[18] Google Translation, http://www.google.com.tw/translate_t

[19] Yahoo Babel Fish, http://babelfish.yahoo.com/

[20] Excite, http://www.excite.co.jp/world/english/

Generating Patterns for Extracting Chinese-Korean Named Entity Translations from the Web

Chih-Hao Yeh†, Wei-Chi Tsai†, Yu-Chun Wang‡, Richard Tzong-Han Tsai†*

†Department of Computer Science and Engineering, Yuan Ze University, Taiwan
‡Department of Electrical Engineering, National Taiwan Univeristy, Taiwan

*corresponding author

{s941539, s941537}@mail.yzu.edu.tw
r95921024@ntu.edu.tw
thtsai@saturn.cse.yzu.edu.tw

Abstract

One of the main difficulties in Chinese-Korean cross-language information retrieval is to translate named entities (NE) in queries. Unlike common words, most NE's are not found in bilingual dictionaries. This paper presents a pattern-based method of finding NE translations online. The most important feature of our system is that patterns are generated and weighed automatically, saving considerable human effort. Our experimental data consists of 160 Chinese-Korean NE pairs selected from Wikipedia in five domains. Our approach can achieve a very high MAP of 0.84, which demonstrates our system's practicability.

摘要

中韓跨語檢索上的困難之處，即在於query term中的專有名詞，由於人類語言的變化無法如同大多數的一般名詞一樣，能夠在雙語辭典中找到其對應的翻譯詞。本論文提出一種基於翻譯模板在Web中尋找翻譯詞的方式。由於Web的資料幾乎涵蓋人類到目前為止的所有知識，且會隨時更新，因此能確保找到翻譯詞的recall。本研究的一大特點在於，所有用於擷取翻譯詞的模板，均為自動生成，因此不需耗費大量人力來建構。此外，我們會利用訓練資料集來評估各模板的權重，藉以給與各候選詞適當的信心值。我們採用維基百科的中韓專有名詞pair做為本方法所需之訓練集與測試集。經實驗過後，我們的方法可以達到MAP 0.84的高分，證明本論文提出方法的實用性。

Keywords: Chinese-Korean named entity translation, the Web, pattern
關鍵詞：中韓專有名詞翻譯, 網路語料, 模板

1 Introduction

In recent years, South Korean's entertainment industry has established itself as one of the most important emerging markets on the planet. In 2006, South Korean programs on Chinese government TV networks accounted for more than all other foreign programs combined [1]. South Korean actors such as Lee Young-ae (이영애, 李英愛), Bae Yong Joon (배용준, 裴勇俊), Rain, and Song Hye Gyo (송혜교, 宋慧喬) became very popular superstars in the great China area, making text and multimedia information related to them turned hot. Such information is firstly written or tagged in Korean. Unfortunately, most users in this area cannot directly specify queries in Korean. Therefore, it is necessary to translate queries from Chinese to Korean for Chinese-Korean (C-K) information retrieval systems. The main challenge involves translating named entities (such as names of shows, movies and albums) because they are usually the main concepts of queries.

Named entity (NE) translation is a challenging task because, although there are many online bilingual dictionaries, they usually lack domain specific words or NEs. Furthermore, new NEs are emerged everyday, but bilingual dictionaries cannot update their contents frequently. Therefore, it is necessary to construct a named entity translation (NET) system. In [2], the authors romanized Chinese NEs and selected their English transliterations from English NEs extracted from the Web by comparing their phonetic similarities with Chinese NEs. Yaser Al-Onaizan [3] transliterated an NE in Arabic into several candidates in English and ranked the candidates by comparing their counts in several English corpora. Chinese-Korean NET is much more difficult than NET considered in previous works because a Chinese NE may not have similar pronunciation to its Korean translation.

In this paper, we propose an effective pattern-based NET method which can achieve very high accuracy. All patterns are automatically generated and weighed, saving considerable human effort.

2 Difficulties in Chinese-Korean Named Entity Translation

To translate an NE originated from Chinese into Korean, we begin by considering the two C-K NET approaches. The older is based on the Sino-Korean pronunciation and the newer on the Mandarin. For example,“臺灣” (Taiwan) used to be transliterated solely as “대만” (Dae-man). However, during the 1990s, transliteration based on Mandarin pronunciation became more popular. Presently, the most common transliteration for “臺灣” is “타이완” (Ta-i-wan), though the Sino- Korean-based “대만” is still widely used. For Chinese personal names, both ways are used. For example, the name of Chinese actor Jackie Chan (“成龍” Cheng-long) is variously transliterated as “성룡”Seong-ryong (Sino-Korean) and “청룽” Cheong-rung (Mandarin). Translating Chinese NEs by either method is a major challenge because each Chinese character may correspond to several different Hangul characters or character sequences that have similar pronunciations. This results in thousands of possible combinations of Hangul characters (e.g., “張韶涵’ Zhang Shao-han can be transliterated to “장사오한” Jang-sa-o-han or “장샤오한” Jang-sya-o-han), making it very difficult to choose the most widely used one.

NEs originated from Japan may contain Hiraganas, Katakanas, or Kanjis. For each character type, J-C translation rules may be similar to or very different from J-K translation rules. Some of these rules are based on Japanese pronunciation, while some are not. For NEs composed of all Kanjis, their Chinese translations are generally exactly the same as their Kanji written forms. In contrast, Japanese NEs are transliterated into Hangul characters. Take “小泉純一郎” (Koitsumi Junichiro) for example. Its Chinese translation “小泉純一郎” is exactly the same as its Kanji written form, while its pronunciation (Xiao-quan Chun-yi-lang) is very different from its Japanese pronunciation. This is different from its Korean translation, “고이즈미 준이치로” (Ko-i-jeu-mi Jun-i-chi-ro). In this example, we can see that, because the translation rules in Chinese and Korean are different, it is ineffective to utilize phonetic similarity to find the Korean translation equivalent to the Chinese translation.

3 Pattern-Based Named Entity Translation

In this section, we describe our C-K NET method for dealing with the problems described in Section 2. We observed that an NE and its translation may co-exist in a sentence. Such sentences may have structural similarity. For example, for ”李明博” and its Korean translation “이명박”, the following sentences both contain the structure “NE (translation)”:

“2007년 12월19일 밤 $\underline{\text{李明博}}_{\text{NE}}$ ($\underline{\text{이명박}}_{\text{translation}}$) 후보의 당선이 사실상”

“$\underbrace{\text{李明博}}_{\text{NE}}$ ($\underbrace{\text{이명박}}_{\text{translation}}$) 대통령 중국 방문 의미와 과제”

These local structures can be treated as surface patterns. In previous work, [4] employ five hand-crafted patterns to extract NE translations. However, it is time-consuming to manually create most patterns. Therefore, we aim to develop an approach that learns patterns automatically.

3.1 Learning Translation Patterns

To generate translation patterns (TP), we need to prepare sentences containing at least one Chinese NE and its Korean translation. For each pair of sentences x and y, we apply the Smith-Waterman local alignment algorithm [5] to find the longest common string. During the alignment process, positions where x and y share the same word are counted as a match. x's ith character and y's jth character are denoted as x_i and y_i, respectively. The algorithm firstly constructs an $|x| \times |y|$ matrix S. Each element $S_{i,j}$ represents the similarity score of an optimal local alignment ending at x_i and y_j, which can be calculated by the following formula:

$$S_{i,j} = \max \begin{cases} 0 \\ S_{i-1,j-1} + s(x_i, y_i) \\ S_{i-1,j} - d \\ S_{i,j-1} - d \end{cases},$$

where $s(x_i, y_i)$ is the similarity function of x_i and y_i; d is the gap penalty. After S are calculated, we backtrack from the optimal element to generate the output TP τ. τ is initialed to be empty. The backtracking process iterates as follows. Suppose the current element is S_{ij}, our algorithm selects the largest one from $\{S_{i-1,j-1}, S_{i,j-1}, S_{i-1,j}\}$ as the next element. If $S_{i-1,j-1}$ is selected, that is, x_{i-1} and y_{i-1} are identical, x_{i-1} will be attached in front of τ. If either $S_{i,j-1}$ or $S_{i-1,j}$ are selected, a wild card will be attached in front of τ. This process stops until it arrives at the first zero-valued element and τ is output.

The following is an example of a pair of sentences that contains “言承旭” (Jerry Yen) and its Korean translation, “언승욱” (Eon Seung-uk) :

- 대만 배우 언승욱 (言承旭) 요약정보.
- 배우 언승욱(言承旭)이 취재진의 질문에 답하고 있다.

After alignment, the pattern is generated as:

배우 <Korean NE slot>(<Chinese NE slot>)

This pattern generation process is repeated for each NE-translation pair.

3.2 Weighting Translation Patterns

After learning the patterns, we have to filter out some ineffective patterns and determine each TP's weight for ranking translation candidates. Each TP τ is evaluated by employing it to extract all possible Korean translations for each training-set NE e in from the sentences used to generate τ. In extracting e's translations, τ's <Chinese NE slot> is replaced with e. τ's extraction F-score over all training-set NEs is treated as its weight and calculated as follows:

$$\text{Precision} = \frac{\text{\# of correctly extracted translations}}{\text{\# numbers of extracted translations}}$$

$$\text{Recall} = \frac{\text{\# of correctly extracted translations}}{\text{\# of correct translations}}$$

$$\text{F-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.3 Extracting and Ranking Translation Candidates

To find a Chinese NE's Korean translation, we can apply the TPs to extract possible Korean translations. For the given input Chinese NE e, our system sends e to AltaVista and limits the result pages to Chinese and Korean. The snippets are collected and break them into sentences. Only the sentences containing e are retained.

We then use all C-K TPs, whose <Chinese NE slot> are replaced with e, to match the retained sentences. The strings matched by <Korean NE slot> are extracted as e's translation candidates. Each candidate c is scored by summing up the weights of TPs extracting c. This score is used to rank all candidates.

4 Evaluation and analysis

In this section, we conduct an experiment to evaluate our pattern-based C-K NET system on different NE types.

4.1 Data Sets

In this section, we illustrate how to prepare the experimental data for learning and testing TPs. As mentioned in Section 3, TP learning requires sentences containing a Chinese NE and its Korean translation.To prepare them, we firstly collect a list of NEs in Chinese, which comprises of 120 NEs originated in Chinese such as 言承旭 (Jerry Yan) and 周杰倫 (Jay Chou) as well as NEs originated in Korea such as 裴勇俊 (Bae Young-Jun) and 張娜拉 (Jang Na-Ra). These NEs are divided into five types based on its Wikipedia page's categorization, including person, location, organization, architecture, and others. Secondly, to acquire these NEs' Korean translation, each NE is sent to the Chinese Wikipedia, and the title of the matched article's Korean version is treated as the NE's translation in Korean. Thirdly, each NE and its Korean translation are attached in a query and then the query is sent to AltaVista. For instance, "言承旭" (Jerry Yan) and its Korean translation "언승욱" are used to produce a query "+言承旭 +언승욱". The returned snippets in the top 20 pages are split into sentences. Only the sentences that contain at least one NE and its Korean translation are retained in the test set.

The preparation of test data is similar to that of training data. 40 NEs (other than the 120 training NEs) are collected from the Wikipedia. The distribution among the five categories are exactly the same as that of the training NEs. Then, each NE is directly sent to Altavista. The returned snippets in the top 20 pages are split into sentences. Only the sentences that contain the NE are retained.

4.2 Evaluation Methodology

We use the Mean Average Precision (MAP) [6] and Top-1 Precision at the Top-1 to measure our NET system's performance. For evaluating the performance of translating each NE, we can calculate the average precision (AP), which is the average of precisions computed after

Table 1: Evaluation Results

NE Type	MAP	Top-1 Precision
Location	0.7854	0.8571
Person	0.9458	1.0000
Organization	0.8333	0.6667
Architecture	0.7736	0.8333
Others	0.8702	0.7500
All	0.8402	0.8500

truncating the list after each of the correct translations in turn:

$$AP = \frac{\sum_{r=1}^{N}(P(r) \times rel(r))}{\text{number of correct translations}},$$

where r is the rank, N the number of extracted candidates, $rel()$ a binary function on the correctness of a given rank, and $P()$ precision at a given cut-off rank. For evaluating the performance of translating all NEs, we can calculate the mean average precision (MAP), which is mean value of all the average precisions. Top-1 precision, used for evaluating the quality of our system's Top-1 candidates, is computed as the ratio of numbers that the correct translation is the top 1 candidate divided by the total number of NE queries.

For evaluating the translation result, our Korean expert manually checked if the candidates are correct translations. A candidate is judged as correct regardless it is generated based on the Sino-Korean or Mandarin pronunciation.

4.3 Evaluation Result

Table 1 shows the categorical and overall performance of our pattern-based NET method. The results show that our method can achieve close to 0.85 in both metrics. The scores of translating person names are even higher. Both metrics are over 0.9 and the Top-1 precision is equal to 1, that is, all the Top-1 candidates output by our system are correct. This indicates that our system's ranking for translation candidates is very accurate and can be further exploited in other applications that incorporate our NET system.

5 Discussion

From the evaluation results, we find that our method can translate most Chinese NEs effectively. However, there are still some cases dropping the performance. In the following subsections, we discuss TP's effectiveness and analyze error cases.

5.1 Effectiveness of Pattern-based NET

For most test NEs, our method can extract their correct translations and put them in the forepart of the candidate list. For example, for "金泰熙" (Jin Tai-xi), the Top-1 candidate is "김태희" (Kim Tae-Hee), which is exactly the correct translation. As mentioned in section 2, Korean transliterates Chinese NEs according to their Sino-Korean or Mandarin pronunciation. Our method can extract translations based on both methods. For example, for "謝長廷" (Xie Chang-ting), both the Sino-Korean transliteration "사장정" (Sa-jang-jeong) as well as the Mandarin transliterations such "셰창팅" (Sye-chang-ting), "세창팅" (Se-chang-ting), and "시에츠앙

팅” (Si-e-cheu-ang-ting) are extracted. It shows that our method can extract most Chinese NEs’ Korean translations regardless of how the Korean translations are generated.

5.2 Error Analysis

5.2.1 Relevant Terms

Our method may extract an NE’s relevant phrases in addition to its translations. For instance, for “親民黨” (People First Party), both the correct translation “친민당” (Chin-min-dang) and the relevant phrase “쑹추위” (Ssung Chu-ui, 宋楚瑜, the chairman of People First Party) are extracted. Although relevant phrases are not exact the translations, they might improve the performance of some NET applications such as cross-language information retrieval (CLIR). This is because the effect of adding relevant phrases in queries is similar to that of query expansion.

5.2.2 Different Phraseology in Chinese and Korean

Different phraseology in Chinese and Korean might make our NET method extract false positives. For example, the First Sino-Japanese War (1894–1895) is called Jiawu Zhanzheng (甲午戰爭) by Chinese but is called Cheong-il-jeong-jaeng (淸日戰爭) by Koreans. The Top-1 candidate output by our system is “갑오전쟁” (kap-o-jeon-jaeng), which is only annotated for Koreans to understand its pronunciation. The most widely used Korean translations for the First Sino-Japanese War is “청일전쟁” (Cheong-il-jeong-jaeng, 淸日戰爭). The other example is the Chinese query “浪漫滿屋” (Lang-man-man-u), a Korean drama’s name. The Top-1 candidate is “랑만만우” (rang-man-man-u), which is the transliteration of the Chinese characters “浪漫滿屋” based on the Mandarin pronunciation. However, the correct Korean translation is “풀하우스” (Full House), which is the transliteration based on the English pronunciation. These two queries show that different phraseology may rank the incorrect translation candidates higher. However, the correct translations, such as “청일정쟁” and “풀하우스”, are also extracted by our method. For some applications, such as CLIR, the inaccurate ranking does not influence the performance a lot [7].

6 Conclusion

In this paper, we have demonstrated several advantages of our pattern-based NE translation method. Our pattern-based method achieves higher recall because it extracts NE translations from the Web, which contains most of human knowledge. Even translations of novel NEs can be found. Second, our method can extract most translations for each NE. This feature makes similar effects of query expansion and very helpful for cross-language information retrieval because documents containing frequent or infrequent translations can be retrieved. Finally, the high MAP over all five domains establishes our method’s generality.

In the future, we plan to apply our method to other language pairs. We also hope to extract not only the translations but also relevant information to them. We believe these new features can be applied to other applications, such retrieving multimedia contents on the Web 2.0 platform whose tags are written in different languages from queries.

References

[1] Cho Hae-Joang, “Reading the “Korean wave as a sign of global shift””, *Korea Journal*, pp. 167–172, 2005.

[2] Hsin-Hsi Chen, Sheng-Jie Huang, Yung-Wei Ding, and Shih-Cbung Tsai, “Proper name translation in cross-language information retrieval”, *Proceedings of 17th COLING and 36th ACL*, pp. 232–236, 1998.

[3] Yaser Al-Onaizan and Kevin Knight, “Translating named entities using monolingual and bilingual resources”, *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*, pp. 400–408, 2002.

[4] Dong Zhou, Mark Truran, Tim Brailsford, and Helen Ashman, “NTCIR-6 experiments using pattern matched translation extraction”, *Proceedings of NTCIR-6 Workshop Meeting*, 2006.

[5] T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences”, *Journal of Molecular Biology*, vol. 147, pp. 195–197, 1981.

[6] Tefko Saracevic, Paul Kantor, Alice Y. Chamis, and Donna Trivison, “A study of information seeking and retrieving”, *Journal of the American Society for Information Science*, vol. 39, no. 3, pp. 161–176, 1988.

[7] Yu-Chun Wang, Richard Tzong-Han Tsai, and Wen-Lian Hsu, “Learning patterns from the web to translate named entities for cross language information retrieval”, *Proceedings of the third International Joint Conference on Natural Language Processing (IJCNLP)*, 2008.