

## Metaphorical Transfer and Pragmatic Strengthening<sup>1</sup>:

### On the Development of *V-diao* in Mandarin

呂維倫\*

Louis Wei-lun Lu

#### Abstract

In this synchronic study, I shall adopt a corpus-based approach to investigate the semantic change of *V-diao* in Mandarin. Semantically, *V-diao* constructions fall into three categories:

- A) Physical disappearance from its original position, with the V slot filled by physical verbs, such as *tao-diao* “escape,” *diu-diao* “throw away,” and so on.
- B) Disappearance from a certain conceptual domain, rather than from the physical space, with the V slot filled by less physically perceivable verbs, such as *jie-diao* “quit,” *wang-diao* “forget,” and the like.
- C) The third category of *V-diao* involves the speaker’s subjective, always negative, attitude toward the result. Examples include: *lan-diao* “rot,” *ruan-diao* “soften,” *huang-diao* “yellow,” and so forth.

It is claimed in this paper that the polysemy between types A and B is motivated by metaphorical transfer [Sweetser, 1990; Bybee, Perkins and Pagliuca, 1994; Heine, Claudi and Hunnemeyer, 1991]. Based roughly on Huang and Chang [1996], I demonstrate that a cognitive restriction on selection of the verb will cause further repetitive occurrence of negative verbs in the V slot. Finally, I shall claim that pragmatic strengthening [Hopper and Traugott, 1993; Bybee,

---

<sup>1</sup> An earlier version of this paper was presented at ROCLING XIV at National Chengkung University, Tainan, Taiwan. The author is especially grateful to Dr. Lily I-wen Su for her insightful comment on this paper. Without her valuable feedback, completion of this paper would have been impossible. I would also like to thank Dr. Shuanfan Huang, Dr. Wenyu Chiang, Dora Lu, Agnes Huang and two anonymous reviewers of ROCLING for their helpful suggestions. Any remaining errors are of course my own.

\* Graduate Institute of Linguistics, National Taiwan University, Taipei, Taiwan. Email: r89142004@ms89.ntu.edu.tw

Perkins and Pagliuca, 1994] contributes to the emergence of unfavourable meaning in Type C.

Hopefully, this research can serve as a valid argument for the interaction of language use and grammar, and the conceptual basis of human language.

**Keywords:** metaphorical transfer, pragmatic strengthening, conceptualization.

### 1. Semantic Classification of V-*diao*

V-*diao* is traditionally termed a resultative compound, indicating the result of an action [Li and Thompson 1981]. However, a close examination of linguistic data indicates that the semantics of V-*diao* cannot be calculated by simply putting its components together. In this paper, I shall focus on the semantics of *diao* and try to tackle V-*diao* at a lexical level to see whether such lexical analysis works.

The V-*diao* construction comprises a verb (be it action or stative) and a verbal suffix -*diao*. It gives the final state of the agent, if used intransitively, and of the receiver of the action, in transitive cases. It may represent: A) physical disappearance of an entity from its original position, B) disappearance from a certain conceptual domain, and C) the speaker's subjective evaluation of the result of an event, as in (1)-(3), respectively:

- |     |                                      |            |            |      |      |        |
|-----|--------------------------------------|------------|------------|------|------|--------|
| (1) | ta                                   | qiaoqiao   | pao-diao   | le   |      |        |
|     | he                                   | quietly    | run away   | CRS  |      |        |
|     | "He ran away quietly."               |            |            |      |      |        |
| (2) | ta                                   | jie-diao   | le         | nage | huai | xiguan |
|     | he                                   | get rid of | Perf       | that | bad  | habit  |
|     | "He got rid of that bad habit."      |            |            |      |      |        |
| (3) | diennau                              | zuotien    | huai-diao  | le   |      |        |
|     | computer                             | yesterday  | break down | CRS  |      |        |
|     | "The computer broke down yesterday." |            |            |      |      |        |

I shall begin this paper with a close look at the semantics of the foregoing types of V-*diao*, especially the last one. This is because the Type C construction involves an intriguing phenomenon: interpretation of a negative result cannot be arrived at by directly adding the suffix -*diao* to any verb. It is worth noting that, synchronically, the semantics of *diao* denote a downward movement. It is, thus, reasonable to claim that the negative interpretation may derive from the human experiential basis of space.

### 1.1 Type A: Physical Disappearance

It is reported that a suffix in a resultative verb compound in Mandarin indicates the sequel of an action [Li and Thompson 1981]. The first kind of *-diao* gives the final state, i.e., physical absence, of the agent or the patient. This kind of *-diao* is mostly affixed to easily perceivable physical action verbs such as *pao* "run," as in (1), *dū* "throw," *shāo* "burn," and so on.

### 1.2 Type B: Disappearance from a Conceptual Domain

The second sort of V-*diao* also denotes the result of an action. However, this differs from type A in the sense that it represents a less "concrete" disappearance. It is often attached to low transitive verbs, without obvious physical motion, and accompanies an abstract noun phrase. Consider example (2) again:

(2)	ta	jie-diao	le	nage	huai	xiguan
	he	get rid of	Perf	that	bad	habit

"He got rid of that bad habit."

A bad habit is an abstract entity. The abandonment of it by the agent is almost physically undetectable. But how can one perceive its existence and absence? Also, from where does the habit disappear?

This has everything to do with our conceptual system. We experience many things, through sight and touch, as having distinct physical shapes and boundaries. We thus tend to project physical shapes and boundaries on them, conceptualising them as entities and imposing on them physical characteristics such as existence and disappearance, even though we can never really feel them with our hands or sense them with our eyes or nose [Lakoff and Johnson 1980]. Further details concerning Type B and metaphorical transfer will be addressed in the next section.

In this case, a habit is conceptualised as a physical entity. It can fade out, can be done away with, and can finally disappear from our conceptual domain as physical things do from a physical space. Thus, Type B seems to represent the final state of, usually, a non-physical action, i.e., an abstract entity being done away with, finally disappearing from one's conceptual domain.

### 1.3 Type C: Evaluative Function from the Speaker

Type C V-*diao* denotes a somewhat negative evaluation of the result in question. It often co-occurs with verbs with negative connotation, such as *lan-diao* "rot," *si-diao* "die," *shu-diao* "lose," etc. However, its negative meaning does not seem to come from the preceding verb in every case. Consider the following instances (4) and (5):

- (4) binggan ruan-diao jiu bu hauchi le  
 cookie soften PARTICLE not tasty CRS  
 “Cookies won’t taste good if they become soft.”

- (5) cai huang-diao jiu bu xinxi le  
 vegetable yellow PARTICLE not fresh CRS  
 “Vegetables won’t be fresh if they turn yellow.”

In (4) and (5), the words *huang* “yellow” and *ruan* “soft” do not themselves carry negative meanings, but the entire phrase clearly involve one’s unfavourable attitude toward the final state of the vegetables and cookies. In the following sections, I shall examine the semantic change of *-diao* and try to account for the emergence of its unfavourable interpretation.

## 1.4 Data and Methodology

Two main sources provide examples discussed to illuminate this search. The written source mostly comes from the Academia Sinica Corpus, with a complete tagging system. The spoken source comprises the Taida Spoken Corpus, together with another eight hours of transcribed data<sup>2</sup>. The spoken part amounts to an entire length of sixteen hours of conversational Mandarin. In sum, we collected a total of one hundred and eighty-nine tokens of *-diao*, excluding its use as a main verb such as *xiao-diao-da-ya* (笑掉大牙), *diao-tao* (掉頭), and so on. Also, when our argument called for constructed examples, native speakers, inclusive of the author himself, were consulted.

Two interesting observations on the corpora are left unaddressed due to the limited scope of the current study. First, the approximate portion of main verbs is much higher in our written corpus than that in our spoken corpus (around 4:1). Second, the development of *-diao* seems to match the tendency of subjectification proposed by Traugott [1989, 1995]. However, these issues are not closely related to the current study and will, thus, be left out of this research.

## 2. Metaphorical Transfer

It is argued that, when a grammatical meaning is derived from its source, there often exists a metaphorical relation between the two meanings [Sweetser, 1990; Bybee, Perkins and Pagliuca, 1994]. Such a semantic change takes place to serve a certain functional end in grammar and discourse, as indicated by Heine, Claudi and Hunnemeyer [1991:48]:

We try to demonstrate that metaphorical transfer forms one of the main driving forces in the

---

<sup>2</sup> The second source of spoken data was offered by Dr. Lily I-wen Su.

development of grammatical categories; that is, in order to express more “abstract” functions, concrete entities are recruited.

The above corresponds to my observations of *V-diao*: a metaphorical transfer takes place when meaning proceeds from the physical domain to a conceptual domain, denoting metaphorical disappearance.

## 2.1 From Type A to Type B: Metaphor at Work

The above claim seems to be verified in the development of *-diao*. The meaning of Type A is the most concrete and physical one, since it indicates a salient result after some physical action is carried out. Type B, on the other hand, denotes disappearance from our mental space instead of from a physical space. Now consider (6) a typical instance of such metaphorical transfer:

- (6) a. ta           xiang       pao       keshi       pao-bu-diao  
           he         think      run       but         run-not-away

“He tried to escape but failed.”

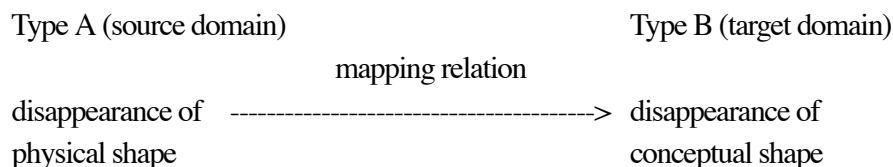
- b. zhuan       ge           shiwàn           pao-bu-diao  
      earn       PARTICLE a hundred thousand   run-not-away

“(Someone) should earn more than a hundred thousand dollars.”

*Pao-bu-diao* in (6a) denotes the unsuccessful outcome of the agent’s escape. The agent fails to escape and does not disappear. In (6b), the meaning is that the landmark “a hundred thousand” is certain to be met. However, not every single case of Type B has a counterpart in A. Actually, most Type B constructions do not. *Pao-bu-diao* is simply a case employed to illustrate the metaphorical relation of the polysemy between Type A and B. In most cases of Type B *V-diao*, the V slot is filled by less physical verbs, such as *jie* “get rid of” in (2), *hulue* “ignore,” *wang* “forget,” and so on.

## 2.2 Summary

In this section, I have shown that the physical “resultative compound” *V-diao* has undergone a metaphorical transfer and developed the sense of disappearance from a conceptual domain. Thus, it makes perfect sense to conclude that the polysemy in this case is at least partly contributed by metaphor, since disappearance is a common feature of Types A and B. The following figure indicates the mapping relation between Type A and Type B:



*Figure 1 Metaphorical Transfer Between Types A and B V-diao*

### 3. Pragmatic Strengthening

Other than metaphor, pragmatic strengthening is claimed to be a major mechanism of semantic change [Hopper and Traugott, 1993; Bybee, Perkins and Pagliuca, 1994]. In such changes, context plays a crucial role. Frequent use of a grammatical or lexical unit in a particular context may lead to the inference that the context is an incorporated part of its meaning. Goossens' research on Old English modals [1982] indicates that there rarely are "real" epistemic markers in OE, and that possibility markers frequently combine with adverbs to express epistemic functions. That is, speakers can generalise and extract the epistemic meanings from the context and impose them on modals. This suggests that frequent co-occurrence with a particular context may "colour" the semantics of a grammatical unit.

In this section, I will demonstrate that the final stage of development of *V-diao* is based on such a mechanism. Now let us see how language use and context collaborate to produce semantic change in the case of *V-diao*.

#### 3.1 From Type B to Type C: Semanticisation of Context

In Type C *-diao*, the sense of disappearance is retained, but there seems to exist something more than the combination of the verbal sense and disappearance. In general, these phrases involve unfavorable assessment on the part of the speaker. That is, the speaker obviously does not favour the change of state.

It is noteworthy that Type C can be further divided into two subtypes based on the verb in the V slot: 1) verbs with negative connotation, such as *lan* "rot," *si* "die," *po* "break," *shu* "lose," and so on; 2) neutral verbs, such as *huang* "yellow," *ya* "croak," *ruan* "soft," and so on. This classification highly pertains to the semantic change addressed in the current research. Let us see how.

Initially, only the former combinations are formed. They simply denote a metaphorical disappearance, labeled Type B. As the frequency of use increases, the speakers tend to associate the construction with the adverse image related to negative verbs. Such frequent collocation of negative verbs and *-diao* may invite the generalisation that the suffix is applied to express one's unfavourable appraisal of the situation at issue. The context is, thus, "semanticized" [Hopper and Traugott, 1993:75] and is transferred onto *-diao*. Consequently, the construction may accommodate neutral stative verbs in

the V slot and still gain a negative interpretation. See (4) and (5) again for the purpose of illustration:

- (4)    binggan    ruan-diao    jiu                    bu            hauchi            le  
       cookie    soften        PARTICLE    not            tasty            CRS  
       “Cookies won’t taste good if they become soft.”
- (5)    cai            huang-diao    jiu                    bu            xinxi            le  
       vegetable    yellow        PARTICLE    not            fresh            CRS  
       “Vegetables won’t be fresh if they turn yellow.”

*Huang* and *ruan* themselves do not signal negativity. The adverse meaning is subtly signalled and triggered by the repetitive occurrence of negative verbs in the position. In other words, the emergence of the speaker’s negative attitude derives neither from the suffix denoting disappearance, nor from the verb preceding it, but could have been generalised from the constant collocation of negative words and *-diao*. Now, even neutral verbs may fit into the V slot and yield negative assessment. However, no positive verbs may combine with *-diao*. Details of this co-occurrence restriction will be given in the next section.

### 3.2 Summary

Pragmatic strengthening is one of the driving forces of semantic change, and I have proven that it plays a crucial role in the development of *V-diao* as well. First, only verbs that result in physical and conceptual disappearance may occur in the construction. Among them, a group of verbs with negative connotation prompt the deduction of negative connotation. Consequently, the negative sense of the verb is transferred to the entire phrase, resulting in the speaker’s unfavorable appraisal of the result. The following figure illustrates the development path from Type B to Type C:



*Figure 2* Semanticisation of the Context in *V-diao*

### 4. Conceptual Structure and Selectional Restriction

As the polysemy of *V-diao* develop, its use broaden to increasingly wider contexts. At first, it only accommodates physical verbs and denotes physical disappearance. It then proceeds to tolerate less physical verbs and metaphorically allows a sense of conceptual disappearance. Finally, it may be applied

to a variety of stative verbs to express the speaker's attitude. Nevertheless, in spite of its seemingly free occurrence, some restrictions still exist. Consider the following pairs for the purpose of illustration:

(7) a. wo      zhengge      ren      sha-diao      le  
       I      entire      person      dumb-Suffix      CRS  
       "I was entirely stunned."

b. \*wo      congming-diao      le  
       I      smart-Suffix      CRS

(8) a. dongxi      langfei-diao      le  
       thing      waste-Suffix      CRS  
       "The thing is wasted."

b. \*dongxi      zhenxi-diao      le  
       thing      cherish-Suffix      CRS

From the above pairs, it is evident that the V slot does not allow verbs with positive connotation. It seems that the semantics of positive verbs clashes with that of the entire construction. Why is this the case? What is basis of this selectional restriction?

#### 4.1 Metaphorical Basis of Selectional Restriction

I have argued for metaphor as the driving force of semantic change in the development of *V-diao*. The metaphorical transfer discussed in section two must obey the orientational metaphor GOOD IS UP; BAD IS DOWN proposed by Lakoff and Johnson [1980:16]:

Physical basis for personal well-being: Happiness, health, life, and control– the things that principally characterize what is good for a person– are all UP.

Also, C. R. Huang's previous studies on Mandarin *-qilai* constructions indicate that the development of grammatical units cannot contradict the metaphor that they are based on, and that the collocations of *-qilai* and verbs are conceptually restricted on a semantic basis [Chang 1994, Huang and Chang 1996]. The following observations concerning *V-diao* correspond to this claim.

The physical and experiential basis for DOWN IS BAD is also evident in our language use and conceptual system. Synchronically, the most basic meaning of *diao* is physical dropping / falling, signaling downward movement. It follows that *diao* can relate to something bad in our conceptual system. Whether it is grammaticalised or not, *diao* should never override the conceptual restriction to modify something good. In other words, if the metaphor DOWN IS BAD is truly at work, it seems



rather natural for *V-diao* not to accommodate a verb with positive connotation. Thus, the conceptual / cognitive restriction can fully account for the intrinsic incompatibility of positive verbs with *V-diao*.

The above semantic restriction is critical in the development from Type B to Type C *V-diao*; without it, later unfolding would be impossible. Language users generalise the negative meaning of *-diao* from a previous existing pattern. The constraint must have existed prior to the semanticisation of context. Otherwise, without such a selectional restriction, the meaning would fail to emerge, since positive verbs would intervene. Therefore, it is safe to say that this constraint metaphorically shapes, or at least partly contributes to, the semantic shift of *V-diao*.

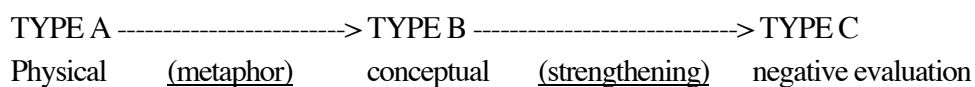
## 4.2 Summary

In this section, the incompatibility of positive verbs with *-diao* has been explored from a semantic viewpoint. The meaning of *diao* conceptually constrains the verbs it co-occurs with, which proves the metaphorical nature of our conceptual system. Also, this selectional restriction results in the existing pattern, which in turn results in the negative meaning of *-diao*. This metaphorical condition, thus, reflects interaction between the grammar and conceptual system.

## 5. Conclusion

In this study, I have classified *V-diao* constructions according to their semantics. In the second section, metaphorical transfer has been proposed as an important mechanism involved in the development of *V-diao*. Further, I have discussed how pragmatic strengthening enables language users to arrive at the negative meaning of *-diao*. Figure 3 shows different stages of *V-diao* and the change of mechanism.

Finally, I have shown that a selectional restriction on the V slot exists. The exclusion of positive verbs is conceptually conditioned by the semantics of *diao*. This suggests that the semantic change and grammaticalisation process of a grammatical unit is conditioned by human experiential basis. Hopefully, this study will serve as a valid argument for the interaction between our language use and grammar, and for a conceptual basis of human language.



*Figure 3 Different Stages of V-diao and Change of Mechanism*

## References

- Bybee, Joan L., Revere Perkins, and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: The University of Chicago Press.
- Chang, Shen-ming. 1994. *V-qi-lai Constructions in Mandarin Chinese: A Study of Their Semantics and Syntax*. M. A. Thesis. National Tsing Hua University.
- Fillmore, Charles J., Paul Kay, and Mary Catherine O'Connor. 1988. "Regularity and Idiomaticity in Grammatical Constructions: The Case of *Let Alone*." *Language* 64:501-38
- Goossens, Louis. 1982. "On the Development of the Modals and of the Epistemic Functions in English." *Papers from the Fifth International Conference on Historical Linguistics*, ed. by Anders Ahlqvist, 74-84. Amsterdam: Benjamins.
- Heine, Bernd, Ulrike Claudi, and Friederike Hunnemeyer. 1991. "From Cognition to Grammar -- Evidence from African Languages." *Approaches to Grammaticalization*. eds. by Traugott and Heine, Vol. 1, 149-87.
- Hopper, Paul, and Elizabeth C. Traugott. 1993. *Grammaticalization*. Cambridge: Cambridge University Press.
- Huang, Chu-ren and Shen-ming Chang. 1996. "Metaphor, Metaphorical Extension, and Grammaticalization: A Study of Mandarin Chinese -qilai." *Conceptual Structure, Discourse, and Language*. ed., by Adele Goldberg. CSLI.
- Lakoff, George, and Mark Johnson. 1980. *Metaphors We Live by*. Chicago: University of Chicago Press.
- Li, Charles, and Sandra Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Los Angeles: University of California Press.
- Sweetser, Eve Eliot. 1990. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge: Cambridge University Press.
- Traugott, Elizabeth Closs. 1989. "On the Rise of Epistemic Meanings in English: An Example of Subjectification in Semantic Change." *Language* 65:31-55.
- Traugott, Elizabeth Closs. 1995. "Subjectification in Grammaticalisation." *Subjectivity and Subjectivisation*, eds. by Dieter Stein and Susan Wright, 31-55. Cambridge: Cambridge University Press.

## A Simple Method for Chinese Video OCR and Its Application to Question Answering

Chuan-Jie Lin\*, Che-Chia Liu\*, Hsin-Hsi Chen\*

### Abstract

Captions in videos contain valuable information for video retrieval. Although texts in captions can be obtained easily in the new image compression formats like MPEG2, there still are many video programs encoded in older formats. Thus, video OCR is indispensable for content-based video retrieval. This paper proposes a simple video OCR method for Chinese captions, including image capturing, caption region deciding, background removing, character segmentation, OCR and post-processing. We employed Discovery Channel films as training and testing corpus. In an outside test, the accuracy of the video OCR was 84.1%. The hardware used in the experiment consisted of a computer with a P4-1.7G CPU, 256MB RAM and a 40G, 7200rpm hard disk. On average, it took 29 minutes and 11 seconds to process a film 495MB in size. We also applied the results of video OCR to video retrieval and question answering.

**Keywords:** digital library, question answering, Chinese video OCR, video retrieval

### 1. Introduction

In the new information era, multimedia is widely used, and the amount of existing video data is huge. How to extract the content of video data for further application has become an important issue. The well-known project "Informedia" [Wactlar, 2000] in digital library is a typical example. Captions in videos contain valuable information for video retrieval. Although texts in captions can be easily obtained in the new image compression formats like MPEG2, there still are many video programs encoded in older formats. Thus, video OCR is indispensable for content-based video retrieval. This paper proposes a simple video OCR method for Chinese captions and demonstrates its application in video search and question answering.

---

\* Department of Computer Science and Information Engineering, National Taiwan University, Taipei, TAIWAN, R.O.C.

E-mail: {cjlin, jjliu}@nlg2.csie.ntu.edu.tw, hh\_chen@csie.ntu.edu.tw

OCR research started very early and has achieved many good results. In a traditional OCR system, textual data is scanned and saved as images, and then transformed into text files [Lee and Chen, 1996]. There have also been many researches on handwriting OCR. In contrast, video OCR is more challenging than traditional OCR because we have to recognize small characters on a colorful background instead of black characters on a white background.

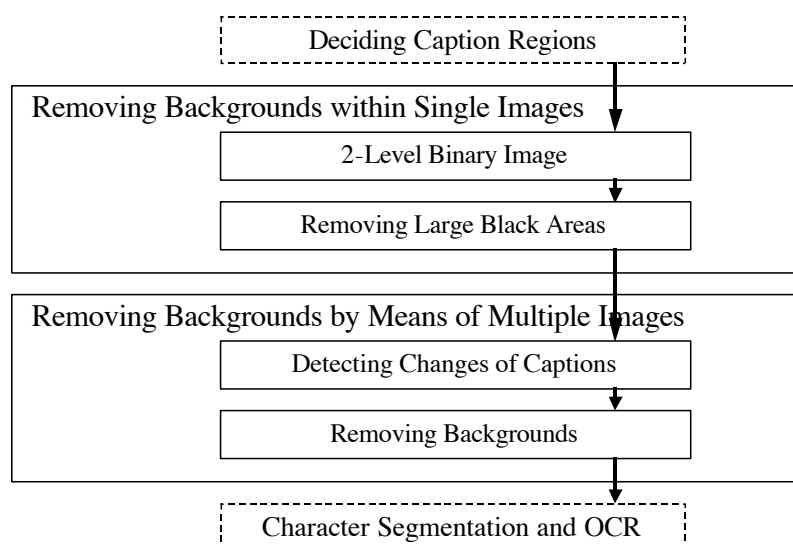
Several approaches have been proposed to video OCR. Wu *et al.* [1997, 1998] tried to find characters in pictures by means of connected components. Their method performs well on pictures but not films because the background of a film is more complicated, and text will also connect with other objects in the film. Lienhart *et al.* [1998, 2000] found text by means of color segmentation, contrast segmentation, geometry analysis, and texture analysis. Li, Doermann and Kia [2000] adopted a neural network to detect strings in images. Li and Doermann [1999] also employed multiple images to enhance resolution. Smith and Kande [1997] used text and object shifting, and facial recognition to reduce the size of images. Sato *et al.* [1998] achieved higher OCR correctness by means of image improving and multi-frame integration.

This paper focuses on Chinese captions in videos. Section 2 introduces several issues concerning video OCR and the architecture of our system. Sections 3 to 8 describe each strategy and each module in detail. The performance was evaluated using films made by the Discovery Channel. Section 9 demonstrates an application for question answering. Section 10 presents conclusion.

## 2. Architecture

There are two kinds of texts in videos, i.e., captions and image texts. Captions often appear at specific positions, such as a textual line in the lower part of a screen, or a vertical text line in the left or right part of a screen. Image texts consist of characters appearing in an image, such as shop signs, automobile registration numbers, *etc.* They are themselves part of the image, so they change their positions when the camera moves. Captions are narratives or dialogues in a film, so they often carry valuable information. The focus of this paper is how to extract texts in captions.

Complex backgrounds often show up behind captions; thus, the first problem is how to remove backgrounds. After backgrounds are removed, the remaining captions are black characters on a white background. That will make the following OCR task easier. We also apply a post-processing procedure to improve OCR performance. Figure 1 shows the architecture of the whole system.



**Figure 1** The Architecture of the Video OCR System.

To evaluate the performance of the system, some films produced by the Discovery Channel were used as experimental materials. Their topics vary widely from natural science to history, military, adventures and human life.

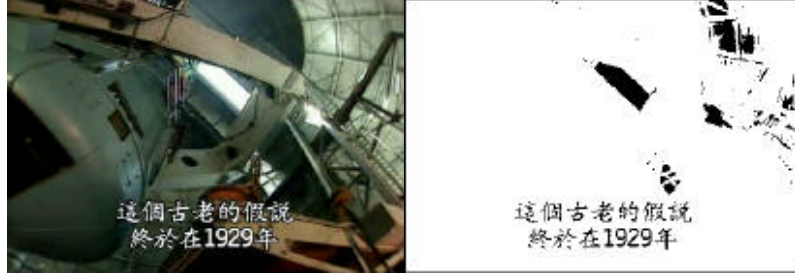
### 3. Deciding Caption Regions

The characteristics of captions are: (1) they are always in a straight line from left to right or up to down; (2) the characters usually have colors which contrast with the background, and often have perceivable borders; (3) they are always in the foreground of the image; (4) they usually consist of two or more characters; (5) the height of the caption region is not often higher than one third of the height of the image, because characters cannot be too large or too small for reading; (6) they have fixed height, width, and size; (7) they have fixed colors. We employ these characteristics to locate captions.

#### 3.1 Binary Image

Before processing, we first transform the original images into binary images. This technique is often used in video processing. It helps to simplify the background and make the retrieval of captions much easier.

When extracting images from a film, we take 2 pictures in a second and save them in the BMP format. In a BMP file, the color of each point is recorded as its *RGB*-value, (*red-value*, *green-value*, *blue-value*). Each value ranges in brightness from 0 to 255. Here, 0 indicates the darkest value and 255 the brightest value.



**Figure 2** An Example of Binary Image Transformation.

Using the *RGB*-values, we can transform an image into a binary image using the following method:

Let the binary-threshold be *SegColorScore*

For each point (*red-value*, *green-value*, *blue-value*) in an image:

**IF** *red-value*, *green-value*, and *blue-value* are larger than *SegColorScore*  
**THEN** change the color of this point to black, i.e., (0, 0, 0)  
**ELSE** change the color of this point to white, i.e., (255, 255, 255).

In our experiment, *SegColorScore* was set to 190. Figure 2 shows an example of binary image transformation. The captions are clearly separated from the background. The result is black characters on a white background.

### 3.2 Deciding Caption Regions

After performing binary image transformation, we decide where the captions are. Here, we employ another characteristic of captions: if we draw a horizontal line across a caption, the line will go through many vertical lines of Chinese characters. As in printed characters, these vertical lines are often of the same width.

Consider every point at the same height  $height_i$ . A sequence of black points is called a *segment*. In this way, a horizontal line at  $height_i$  is composed of a set  $SEGMENT_i=(segment_{i1}, segment_{i2}, \dots)$  of segments. If the difference between the numbers of black points in two neighboring segments is not larger than a predefined threshold (e.g., 3 in this paper), then we say these two segments belong to the same group. Thus, we have a set  $GROUP_i=(group_{i1}, group_{i2}, \dots)$  at  $height_i$ .  $Seg(group_{ij})$  is defined as the number of segments in  $group_{ij}$ . Now we define *Score As Caption Region* (abbreviated as *SACR* hereafter) of  $height_i$  as

$$SACR_i = \sum_{j=1}^{|GROUP_i|} Seg(group_{ij}) \times \log_2 Seg(group_{ij}). \quad (1)$$

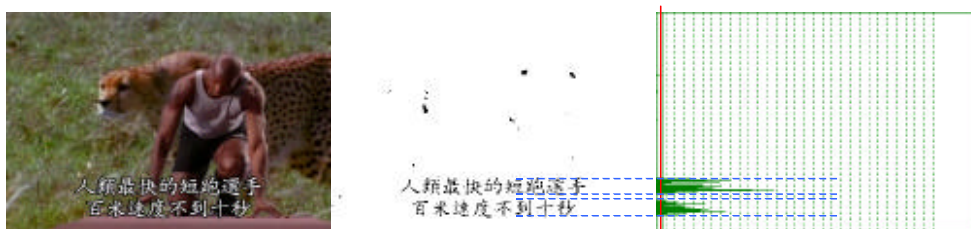


Figure 3 Examples of Deciding Caption Regions (1).

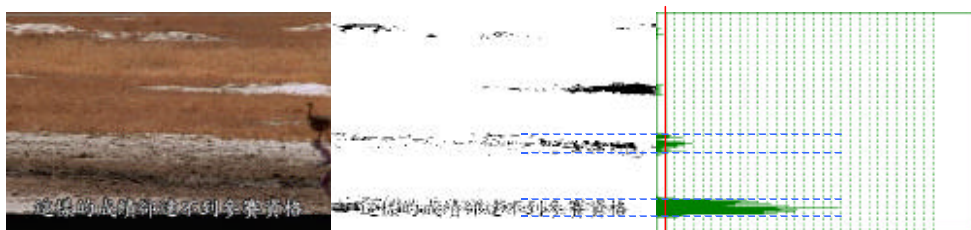


Figure 4 Examples of Deciding Caption Regions (2).

Consider the following example. Here, 0 denotes a white point and 1 a black point.

```

points:    001110111110011000111011111110111111100111101110110111111
segments:  --111-22222--33---444-5555555-666666666--7777-888-99-AAAAAA
groups:    |-----1-----| |-----2-----| |-----3-----| |--4--|
Seg(group):      4              2              3              1
    
```

SACR in this example is  $4\log_2 4 + 2\log_2 2 + 3\log_2 3 + 1\log_2 1 = 14.75$ .

Assume that the height of an image is  $m$ . We calculate  $m$  SACR's for the height levels and compute the average  $\overline{SACR}$ . The height levels that have SACR's higher than the average one are in the caption region. Figures 3 and 4 show two examples. On the left side is the original image; in the middle is its binary image; and on the right side is the corresponding SACR of each height level, where the x-axis denotes the height, the y-axis denotes the SACR value, the solid vertical line is  $\overline{SACR}$ , and the horizontal dashed lines denote the caption regions.

### 3.3 Evaluation

The experiment was performed on three Discovery films: "Lightening," "Animals in the Wild," and "Whales." There were 69, 66, and 41 sentences in captions, respectively. The first 500 images of each film were extracted as experiment data. As shown in Table 1, the precision rates obtained were 76.7%, 39.8% and 82.0%, respectively, but the recall rates were nearly 100%. Errors occurred in cases like the stone road shown in Figure 4. The white stone road in the image had many black segments of the same width, so it was misjudged as a

caption region. Such misjudgments can be filtered out in the OCR processing stage. Hence, the recall rate is more important here for retrieving all the captions.

**Table 1.** Evaluation of Caption Region Deciding.

Films	Actual	System Decided	Correct	Precision	Recall
<i>Lightening</i>	69	90	69	76.7%	100.0%
<i>Animals in the Wild</i>	66	161	64	39.8%	97.0%
<i>Whales</i>	41	50	41	82.0%	100.0%

#### 4. Removing Backgrounds within Single Images

When we adjusted the binary image threshold *SegColorScore*, we found an interesting phenomenon: if *SegColorScore* was set too low, the background could not be removed very well; on the other hand, if it was set too high, the background was removed, but the captions were too unclear to do OCR. The value 190 used in the previous module resulted in very unclear images.

To do OCR more precisely, we have to keep the character clear while removing all the background. In this section, we will propose a method for removing backgrounds within single images by employing the difference between the captions and the background. How information from multiple images is used to remove backgrounds will be discussed in the next section.

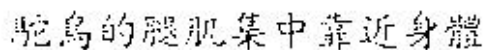
##### 4.1 2-Level Binary Image

During transformation into binary images, the values of *SegColorScore* will affect the clearness of the remaining images of captions. As shown in Figures 5 and 6, captions are clearly seen when *SegColorScore* is set to 140, but more background parts remain. The situation is reversed when it is set to 180.

Here, we propose a new method, called **2-level binary image transformation**, which employs two different *SegColorScore* values to keep captions clear and to remove backgrounds at the same time. The method is described in the following.



**Figure 5** Binary Image of *SegColorScore*=140.



**Figure 6** Binary Image of *SegColorScore*=180.



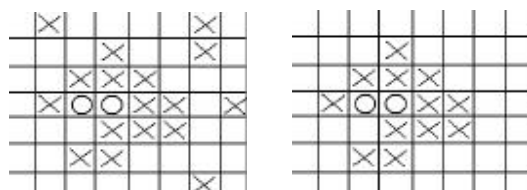


Figure 7 Illustration of 2-Level Binary Image Transformation.

### 駝鳥的腿肌集中靠近身體

Figure 8 2-Level Binary Image of Figure 5 and Figure 6.

Given a picture, we overlap two binary images obtained using two different *SegColorScore* values (let *HiSegColorScore* be the higher one, and *LowSegColorScore* the lower one). Consider the example shown in Figure 7. ‘O’ denotes a black point in both binary images, and ‘x’ a black point only in the binary image obtained using a lower *SegColorScore* value. We keep only those ‘x’ areas adjacent to a ‘O’, because those areas are regarded as black points, and change the other areas into white points. The resulting image is shown on the right side of Figure 7. Figure 8 shows the 2-level binary image result obtained from Figures 5 and 6, which is a clearer caption image.

#### 4.2 Removing Large Black Areas

Consider the image shown in Figure 9, which contains large black areas. It is not easy to remove a background area with a high brightness value using the above method. Thus, another method shown below is proposed to clean such an area if it is large and wide. We will try to deal with small fragments in the next section by using multiple images of the same caption texts.

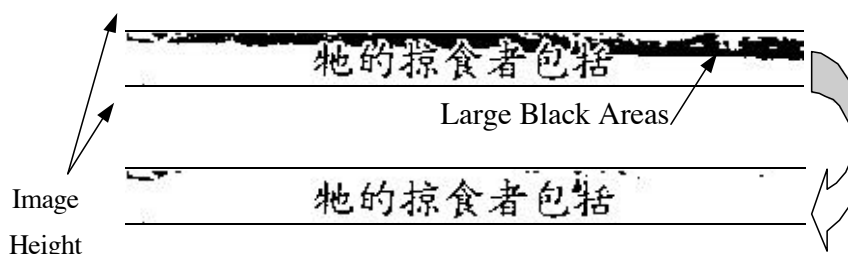


Figure 9 An Example of Removing Large Black Areas.

---

Range = (height of the caption region)  $\div$  4;

Total = Range  $\times$  Range  $\times$  0.9;

**CHECK** each black point in the caption region

Look at a square with edge of Range and with an upper-left corner at this point

**IF** the number of black points in this square  $\geq$  Total (i.e., 90% of the points are black)

**THEN** clear all the points in the area adjacent to this point

**END**

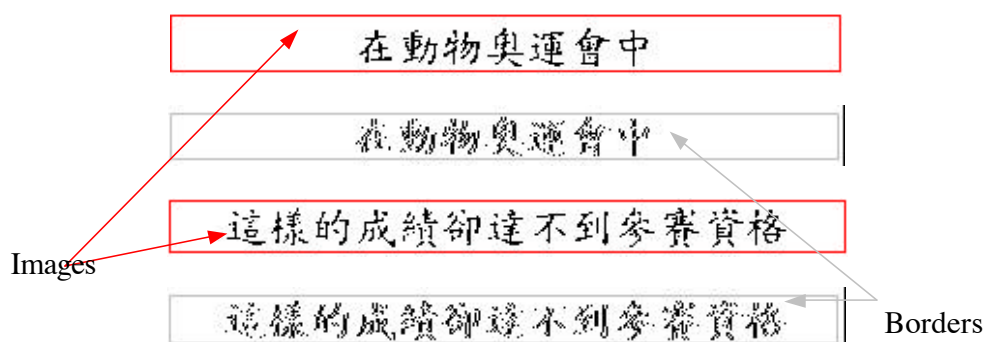
---

## 5. Removing Backgrounds by Means of Multiple Images

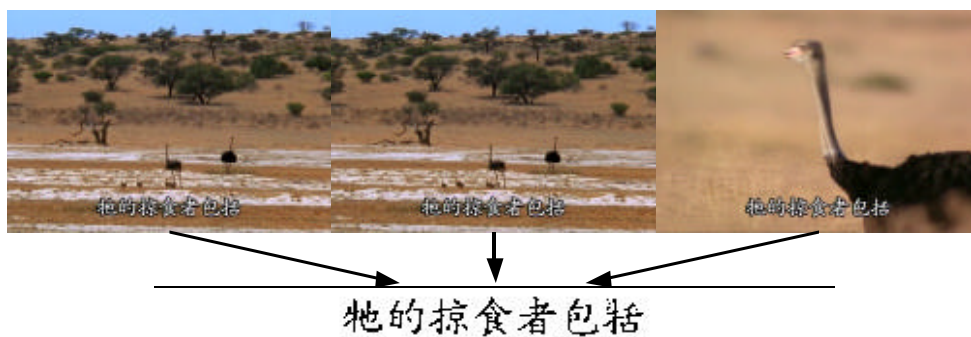
We employ another characteristic of captions to remove small and bright backgrounds; i.e., the positions of the images of captions will not change with the camera, but the background will. We overlap all the images with the same caption texts. Those black points which appear in almost all the images are considered as caption texts. In the next two subsections, we will introduce the method we use to detect the changes of caption texts and the method we use to remove backgrounds by means of multiple images.

### 5.1 Detecting Changes of Captions

The first task in removing backgrounds with multiple images is to decide which images have the same caption texts. Refer to the example shown in Figure 10. We record the border information of all the black areas. After reading the next image, we compare the border information with that of the previous one. If the difference is larger than a threshold, say,



*Figure 10 An Example of Detecting the Change of Captions.*



**Figure 11** An Example of Removing Backgrounds by Multiple Images.

*SceneChangeScore*, we postulate that the caption texts are different. In the experiment, the value of *SceneChangeScore* was set to 0.6.

The same three films used to evaluate the method used to determine caption regions were also used to evaluate this method. Table 2 shows that the performance was quite good.

**Table 2.** Evaluation of Detecting Changes of Subtitles.

Film	Number of Changes	Number of False Alarms	Correctness
<i>Lightening</i>	69	0	100.0%
<i>Animals in the Wild</i>	66	3	95.5%
<i>Whales</i>	41	0	100.0%

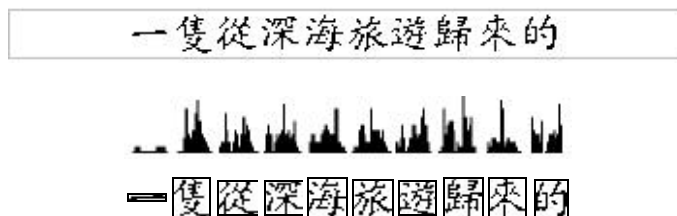
## 5.2 Removing Backgrounds by Means of Multiple Images

After detecting a sequence of images with the same caption texts, we use the following method to remove the backgrounds. Let *NumFrames* be the total number of sequential images. We consider each point in the caption region. If it is black in 90% of the images (i.e.,  $NumFrames \times 0.9$ ), then we set the point as black. Otherwise, it is set as a white point. Figure 11 shows an example. The background is removed more clearly than that is in Figure 9.

## 6. Character Segmentation

At this point, there exists a binary image that has black characters on a white background for each sentence in a caption. We next apply traditional OCR techniques to retrieve caption texts. The first step in performing OCR is to decide the boundaries of each character.

We first decide the left and right boundaries. The most popular way to perform character segmentation is to use projection profiles [Lu, 1995]. As shown in Figure 12, we project every black point onto a horizontal line. Intuitively, the projection for the space between Chinese characters is zero. However, there is also space inside a Chinese character. We employ another cue to resolve this problem. The width of Chinese characters is often



*Figure 12 An Example of Character Segmentation.*


approximately equal to their height. Let the height of a caption region be *ImageHeight*. The gap that is a distance of  $ImageHeight \times 0.7 \sim ImageHeight \times 1.4$  from the previous gap will be regarded as a possible segmentation point.

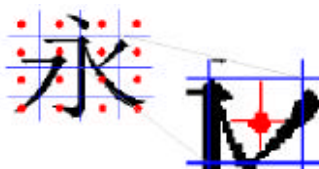
After deciding the left and right boundaries, we use the same method to decide the upper and lower boundaries of each character.

## 7. Optical Character Recognition

We adopt a statistical model similar to that of Oka [1982] to perform Chinese OCR. Figure 13 shows an example. Each character image is separated into 16 equal parts. Starting from the center of each part, we observe its up, down, left, and right directions. If there is a black point in a given direction, the corresponding signature value is set to 0. Otherwise, it is set to 1. In this way, we will have 64 (16 parts  $\times$  4 directions) values (called a signature) for each character image.

A set of character images that were retrieved from the Discovery Channel films formed a corpus for collecting the signatures of a standard character corpus. When recognizing a new image, we first extract its signature and then compare it with the ones in the standard character corpus. The similarity is measured by counting how many values are matched. Therefore, the similarity score will be between 0 and 64. The higher the score is, the more similar the two patterns are. If the highest score of a new image is less than 16, it is regarded as non-character image.

The following is an example. A new image  is compared with '傳' and '博' in the standard character corpus. The corresponding signatures are as follows:



*Figure 13 Signature of Image “永”.*



Figure 14 The First Ten Candidates of OCR.

傳 1010100010001101110100100000101111101010010001011111111111001111  
 傳: 1010100000001101010101100000100111101010010001011111111111001111  
 博: 1110100000011111010001000100100111101100010001011111111111101101

The similarity of the image 傳 with ‘傳’ and ‘博’ is 60 and 50, respectively, so ‘傳’ is ranked as the first candidate of this image.

Figure 14 illustrates the first ten candidates of each character image in 探索遺傳學的奇異世界 after OCR is performed. The correct rate of the top one is very high, and most of correct answers appear in the top ten. Table 3 shows the top one performance. The film "Genetics" was used in inside test, and "King of the Pyramids" and "The Real Cleopatra" were used in the outside tests. The results show that the correct rates in the inside test was 91.5%, and that the performance of the outside tests was 78.5% and 81.5% for the two films, respectively.

Table 3. Experiment Results of OCR.

Films	TOTAL	CORRECT	ERROR	MISS
<i>Genetics</i>	809	739 (91.5%)	69 (8.5%)	0
<i>King of the Pyramids</i>	684	537 (78.5%)	110 (16.1%)	37 (5.4%)
<i>The Real Cleopatra</i>	750	611 (81.5%)	86 (11.5%)	53 (7.1%)

## 8. OCR Post-Processing

We found that nearly 95% of the correct answers were in the top ten candidates, and Table 3 shows that the top one achieved 91.5% performance in the inside test. This section will touch on how to promote the correct answer which is not ranked first initially to the first position to improve the overall performance.

## 8.1 Basic Model

In Figure 14, a value enclosed in parentheses before a candidate denotes its similarity score. First, we filter out those candidates whose scores are lower than the score of the top one candidate by a threshold. The filtered characters are shadowed in Figure 14. Only the characters with larger scores are retained. This will reduce the number of possible candidates. Then, we perform the following steps. Consider three characters denoted  $ABC$  sequentially. Generate all the possible candidate pairs for  $ABC$ , e.g.,  $A_iB_j$  or  $B_mC_n$ . Check if a candidate pair is in a dictionary (i.e., a two-character word), or is a part of a three-character word. If it is, we multiply the OCR similarity scores of these two candidates. Otherwise, their score is set to zero. Next, we find the pair with the highest score. If it is  $A_iB_j$ , then  $A_i$  and  $B_j$  are selected, and we start the next iteration from  $C$  (i.e.,  $CDE$ ). If it is  $B_mC_n$ , then  $A_i$ , i.e., the top one candidate of  $A$ , is selected, and we start the next iteration from  $B$  (i.e.,  $BCD$ ).

## 8.2 Strategies Used in Experiments

For the above algorithm, several issues had to be evaluated in the experiments. For example, should we consider all the combinations of characters? Is the top one candidate more important than the others? Are longer words in the dictionary more helpful? We applied 3 strategies to the basic model to examine these factors. The experimental results were compared with those obtained using the Select-First and Longest-First models.

- [Strategy 1] All pairs of candidates are considered.
- [Strategy 2] Only pairs consisting of at least one ranked first candidate are proposed. In other words, when  $AB$  are recognized, only  $A_1B_1, A_1B_2, \dots, A_2B_1, A_3B_1, \dots$  are considered.
- [Strategy 3] 4- or 3-character words in the dictionary are proposed first. Then, Strategy 2 is considered.

## 8.3 Evaluation

The standard character corpus was collected from six Discovery films (i.e., "Natural Born Winners," "Snakes," "Genetics," "The Southern Rockies," "Great Quakes: Kobe, Japan," and "Galapagos: Beyond Darwin"). There were in total 7,818 character images, and only 2,256 signatures of distinct characters were recorded. Tables 4 to 6 show the experimental results for the three different films. Among them, "Genetics" was used in the inside test; "King of the Pyramids" and "The Real Cleopatra" were used in the outside tests. The first 700 images of each film were extracted as experiment data. The notations used in the tables are defined

as follows:

- TOTAL: total number of characters in captions;  
 CORRECT: number of characters recognized correctly;  
 ERROR: number of characters collected in the standard corpus but recognized incorrectly;  
 MISS: number of characters which are not collected in the standard character corpus;  
 Improve: improvement relative to the baseline;  
 Select-First: (baseline) select the top one candidate;  
 Longest-First: select the longest candidate combination which is collected in the dictionary.

**Table 4.** Experimental Results of Post-Processing for the Film "Genetics".

	TOTAL	CORRECT	ERROR	MISS	Improve
Select-First	809	739 (91.5%)	69 (8.5%)	0	-----
Longest-First	809	753 (93.1%)	56 (6.9%)	0	1.6%
Strategy 1	809	751 (92.8%)	58 (7.2%)	0	1.3%
Strategy 2	809	759 (93.8%)	50 (6.2%)	0	2.3%
Strategy 3	809	762 (94.2%)	47 (5.8%)	0	2.7%

**Table 5.** Experimental Results of Post-Processing for the Film "King of the Pyramids"

	TOTAL	CORRECT	ERROR	MISS	Improve
Select-First	684	537 (78.5%)	110 (16.1%)	37 (5.4%)	-----
Longest-First	684	544 (79.5%)	103 (15.1%)	37 (5.4%)	1.0%
Strategy 1	684	546 (79.8%)	101 (14.8%)	37 (5.4%)	1.3%
Strategy 2	684	559 (81.7%)	88 (12.9%)	37 (5.4%)	3.2%
Strategy 3	684	563 (82.3%)	84 (12.3%)	37 (5.4%)	3.8%

**Table 6.** Experimental Results of Post-Processing for the Film "The Real Cleopatra".

	TOTAL	CORRECT	ERROR	MISS	Improve
Select-First	750	611 (81.5%)	86 (11.5%)	53 (7.1%)	-----
Longest-First	750	614 (81.9%)	83 (11.1%)	53 (7.1%)	0.4%
Strategy 1	750	635 (84.5%)	62 ( 8.3%)	53 (7.1%)	3.0%
Strategy 2	750	640 (85.3%)	57 ( 7.6%)	53 (7.1%)	3.8%
Strategy 3	750	644 (85.9%)	53 ( 7.1%)	53 (7.1%)	4.4%

Tables 4, 5, and 6 show that Strategy 3 was the best one. The correct rates were 82.3% and 85.9% in the outside tests, and 94.2% in the inside test. 5.4% and 7.1% of the characters could not be found in the dictionary in the outside tests, respectively.

We further compare the experimental results obtained using Strategy 3 and the Select-First Model in Table 7, where "T→F" is the number of characters recognized correctly

using Select-First but incorrectly using Strategy 3, and “F→T” is the number of characters recognized correctly using Strategy 3 but incorrectly using Select-First. From Table 7, we can find that “T→F” case was only 0.7%, but that 3.0% to 5.2% of more characters could be recognized correctly. This leads us to the conclusion that post-processing is helpful.

**Table 7.** Comparison of Strategy 3 and Select-First.

Film	Total	Result	T→T	T→F	F→T	F→F
<i>Genetics</i>	809	94.2%	738 (91.2%)	6 (0.7%)	24 (3.0%)	41 (5.1%)
<i>King of the Pyramids</i>	647	87.0%	532 (82.2%)	5 (0.8%)	31 (4.8%)	79 (11.2%)
<i>The Real Cleopatra</i>	697	92.4%	608 (87.2%)	3 (0.4%)	36 (5.2%)	50 (7.2%)

Table 8 shows the experimental results for the three whole films. The main error in the outside test was that about 7~10% of the characters were not collected in the standard character corpus. The signatures of the standard character corpus were collected from the real images of the six films, and only those of 2,256 distinct characters were included.

**Table 8.** Experimental Results for the Entire Films Obtained Using Strategy 3.

Film	Real Answers	Reported by System	Correct (Recall)	Error	Miss
<i>Genetics</i>	9189	8834	8105 (88.2%)	1481(16.1%)	26(0.3%)
<i>King of the Pyramids</i>	7976	7878	6582 (82.5%)	851(10.7%)	543(6.8%)
<i>The Real Cleopatra</i>	8862	8874	7365 (83.1%)	636(7.18%)	861(9.7%)

To solve this problem, we tried to collect the signatures from the existing font types. We experimented on 標楷體 and 華康中楷體. The experimental results are listed in Table 9. The first experiment was the same the experiment reported in Table 8. In the second and the third experiments, we used 5,401 frequently used Chinese characters as the standard character corpus in 標楷體 and 華康中楷體, respectively. Comparatively speaking, the results were worse, and using 華康中楷體 was better than using 標楷體.

In addition, we prepared another standard character corpus for the fourth and the fifth experiments, in which 2,256 signatures came from the original corpus, and the other Chinese characters came from the 華康中楷體 images. The performance was improved, but it was still not as good as that obtained in the inside test. Meanwhile, the whole character set (13,060) did not perform better than the set of frequently used characters.



**Table 9.** Experimental Results on Different Standard Character Corpora ("King of the Pyramids").

	Real Answers	Reported by System	Correct (Recall)	Error	Miss
2,256, Original	7976	7878	6582 (82.5%)	851 (10.7%)	543(6.8%)
5,401, 標楷體	7976	7092	2648 (33.2%)	5325 (66.8%)	3(0.0%)
5,401, 華康中楷體	7976	7380	3265 (40.9%)	4708 (59.0%)	3(0.0%)
5,401, Original+華康中楷體	7976	7885	6701 (84.0%)	1272 (15.9%)	3(0.0%)
13060, Original+華康中楷體	7976	7885	6612 (82.9%)	1272 (15.9%)	0(0.0%)

## 9. Question Answering (QA) System

Since caption texts in video can be extracted successfully using the procedures proposed in the previous sections, we tried to integrate the IR and QA techniques to develop a video question answering system in the next step.

### 9.1 Video QA System

Figure 15 shows the interface of the Video QA System. Users issue questions in the submission window. The system finds answers in a film corpus and shows them in the answer window with several indicative pictures extracted from the video for each answer. If the user wants to watch the original film for an answer, he can click on that picture, and the system will play the film starting from the answer fragment.

The technique used for QA was proposed by Lin *et al.* [2001]. It implements a question answering system on heterogeneous collections including video. The correctness of Video OCR is not 100% yet (82.3% or better is shown in Tables 5 and 6), so pattern matching in traditional techniques (i.e., matching keywords or synonyms, or searching in other semantic trees) has to take OCR similarity into account. The score for extracting answers can be calculated as follows:

$$score(qw_i, pw_j) = 0 \quad \text{if } |qw_i| \neq |pw_j|$$

$$\text{else} = \left( \frac{\sum_{k=1}^{|qw_i|} Ocr(qc_k, pc_k)}{|qw_i|} \right) \times weight(qw_i), \quad (2)$$

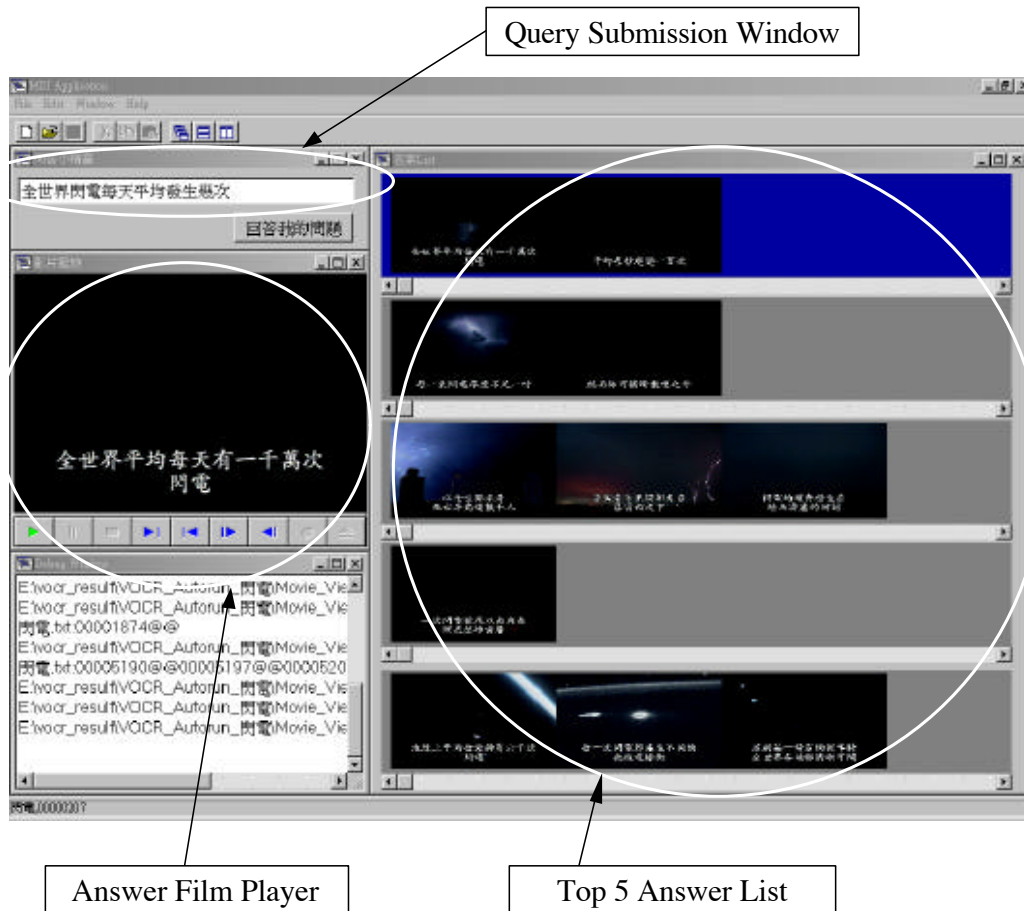


Figure 15 The Interface of Video QA System.

where  $|qw_i|$  denotes the number of characters in  $qw_i$ , and  $qc_k$  is the  $k^{\text{th}}$  character in  $qw_i$  (the same convention is used for  $pw_j$ ).  $\text{Ocr}(qc_k, pc_k)$  is the OCR similarity of characters  $qc_k$  and  $pc_k$ .

## 9.2 Evaluation

### 9.2.1 Questions

Testing questions were collected from "Assignment Discovery" at the web site of Discovery, traditional Chinese version (<http://chinese.discovery.com/sch/index.html>). "Assignment Discovery" is a project that provides many learning lessons from Discovery programs. This project provides lesson plans, activities, and comprehension questions and answers for teachers to use in designing study programs for students.

We selected the comprehension questions for six films as our testing questions to do the evaluation. We collected questions from this website in order to avoid bias. The films were "Elephants," "On Jupiter," "Hubble: Secrets from Space," "Eye of the Serpent," "Whales," and "Lightning."

### 9.2.2 Performance

The performance of the QA system was measured in MRR (Mean Reciprocal Rank), which was used in the QA evaluation of TREC QA-Track [Voorhees, 2000].

There were 43 questions in total for these six films. The experiment results are listed in Table 10. The MRR result was 0.1848  $(=(4+5/2+3/3+1/4+1/5)/43)$ . 32.6% (14/43) of the questions were answered correctly.

**Table 10.** Evaluation of the Video QA System.

Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Not Answered
4	5	3	1	1	29

From our investigation, the main sources of errors were as follows:

- (1) Characters in keywords were not collected in the standard character corpus,

for example, “雷” in the question “冰雹如何形成?”

- (2) Paraphrase problem.

For the question “木星繞行太陽一週需時多久?”, the answer text is “木星環繞太陽一周，須地球時間十二年。” The two phrases “繞行一週” and “環繞一周” are paraphrases.

- (3) More precise rules for deciding question focus are required.

Consider the question “閃電可以到達多熱的程度?” It is classified as “QUANTITY,” so all quantity expressions become possible candidates. But we should only look for temperature expressions as answers.

- (4) World knowledge is needed.

Consider the question “歷史上第一位做閃電實驗的人是誰?” The correct answer mentions that Franklin did an experiment in 1752, but “the first” is not mentioned. Therefore, it is hard to decide whether he was the first experimenter.

We only employ information consisting of question foci, question keywords, and Named Entities in our Chinese QA system. From the above observations, world knowledge and semantic analysis are needed to answer these questions, especially “How” and “Why” questions. This is a challenging problem.

## 10. Conclusion

This paper has introduced a Chinese video OCR system, including image capturing, caption regions deciding, background removal, character segmentation, OCR, and NLP post-processing. The correctness achieved is above 90% for the inside test, and above 80% for the outside test. Its application to video retrieval and a QA system have also been discussed.

There are mainly four kinds of OCR errors: (1) the standard character corpus is not complete; (2) the background is not clear enough; (3) character segmentation errors; and (4) errors in OCR post-processing. In our standard character corpus, there are only 2,256 characters. But there are 5,401 frequently used Chinese characters, not to mention 7,659 less frequently used characters. This is why many characters could not be recognized. In our experiments, most of the backgrounds could be cleared successfully. But if the objects do not move, or if small fragments appear behind the captions, it is not easy to remove them using our method. This will affect the performance of character segmentation and OCR. The OCR errors may also propagate to the post-processing module. For example, a character image that is not in the standard character corpus will not have a correct answer among its candidates, and these ten candidates will affect the choice of other characters.

## References

- Discovery Channel, <http://chinese.discovery.com/>.
- Lee, Yue-Shi and Hsin-Hsi Chen, "Analysis of Error Count Distribution for Improving the Postprocessing Performance of OCCR," *Communication of Chinese and Oriental Languages Information Processing Society*, 1996, pp. 81-86.
- Li, Huiping and David Doermann, "Text Enhancement in Digital Video Using Multiple Frame Integration," *Proceedings of SPIE, Document Recognition IV*, 1999, pp. 1-8.
- Li, Huiping; David Doermann and Omid Kia, "Automatic Text Detection and Tracking in Digital Video," *IEEE Transactions on Image Processing*, 9(1) 2000, pp. 147-156.
- Lienhart, Rainer and Axel Wernicke, "On the Segmentation of Text in Videos," *Proceedings of IEEE International Conference on Multimedia and Expo (ICME2000)*, 3 2000, pp. 1511-1514.
- Lienhart, Rainer and Effelsberg Wolfgang, "Automatic Text Segmentation and Text Recognition for Video Indexing," *Technical Report TR-98-009, Praktische Informatik IV*, 1998.
- Lin, Chuan-Jie, Hsin-Hsi Chen, Che-Chia Liu, Jin-He Tsai and Hong-Jia Wong, "Open-Domain Question Answering on Heterogeneous Data," *Proceedings of Workshop on Human Language Technology and Knowledge Management, ACL*, 2001, pp. 79-85.
- Lu, Y., "Machine Printed Character Segmentation – An Overview," *Pattern Recognition*, 28, 1995, pp. 67-80.

- Oka, R. I., "Handwritten Chinese-Japanese Characters Recognition by Using Cellular Feature." *Proceedings 6th International Joint Conference on Pattern Recognition*, 1982, pp. 783-785.
- Sato, Toshio, Takeo Kanade, Ellen K. Hughes, Michael A. Smith and Shin'ichi Satoh, "Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Caption," *ACM Multimedia Systems*, 7(5) 1999, pp. 385-395.
- Smith, Michael A. and Takeo Kande, "Video Skimming and Characterization Through the Combination of Image and Language Understanding Technique," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 775-781.
- Voorhees, Ellen, "Overview of the TREC-9 Question Answering Track," *Proceedings of the Ninth Text Retrieval Conference (TREC-9)*, 2000, pp. 71-80.
- Wactlar, Howard, "Informedia - Search and Summarization in the Video Medium," *Proceedings of Imagina 2000 Conference*, 2000.
- Wu, Victor and Edward M. Riseman, "TextFinder: An Automatic System to Detect and Recognize Text in Images," *IEEE Transactions on pattern analysis and machine intelligence*, 21(11) 1998, pp. 1224-1229.
- Wu, Victor; Manmatha, R. and Riseman, Edward. M., "Finding Text in Images," *Proceedings of the 2nd intl. conf. on Digital Libraries*, 1997, pp. 1-10.



## Pitch Marking Based on an Adaptable Filter and a Peak-Valley Estimation Method

Jau-Hung Chen and Yung-An Kao\*

### Abstract

In a text-to-speech (TTS) conversion system based on the time-domain pitch-synchronous overlap-add (TD-PSOLA) method, accurate estimation of pitch periods and pitch marks is necessary for pitch modification to assure optimal quality of synthetic speech. In general, there are two major tasks in pitch marking: pitch detection and location determination. In this paper, an adaptable filter, which serves as a bandpass filter, is proposed for use in pitch detection to transform voiced speech into a sine-like wave. The pass band of the adaptable filter can be adapted based on the fundamental frequency. Based on the sine-like wave, a peak-valley decision method is proposed to determine the appropriate parts (positive part and negative part) of voiced speech for use in pitch mark estimation. In each pitch period, two possible peaks/valleys are searched, and dynamic programming is performed to obtain pitch marks. Experimental results indicate that our proposed method performs very well if correct pitch information is estimated.

### 1. Introduction

In past years, the concatenative synthesis approach has been adopted for use in many text-to-speech (TTS) systems [Hamon *et al.* 1989][Iwahashi *et al.* 1995][Shih *et al.* 1996][Chen *et al.* 1998][Chou *et al.* 1998][Charpentier *et al.* 1986]. Concatenative synthesis uses real recorded speech segments as synthesis units and concatenates them together during synthesis. In addition, the time-domain pitch-synchronous overlap-add (TD-PSOLA) [Charpentier *et al.* 1986] method has been employed to perform prosody modification. This method modifies the prosodic features of a synthesis unit according to the target prosodic information. Generally, the prosodic information of a speech unit includes its pitch (the fundamental frequency,  $f_0$ ), intensity, duration, etc. For a synthesis scheme based on the

---

\* Advanced Technology Center, Computer and Communication Research Laboratories, Industrial Technology Research Institute, Chutung 310, Taiwan  
Email: chenjh@itri.org.tw, kya@itri.org.tw

TD-PSOLA method, it is necessary to obtain a pitch mark for each pitch period in order to assure optimal quality of synthetic speech. The pitch mark is a reference point for the overlap between speech signals.

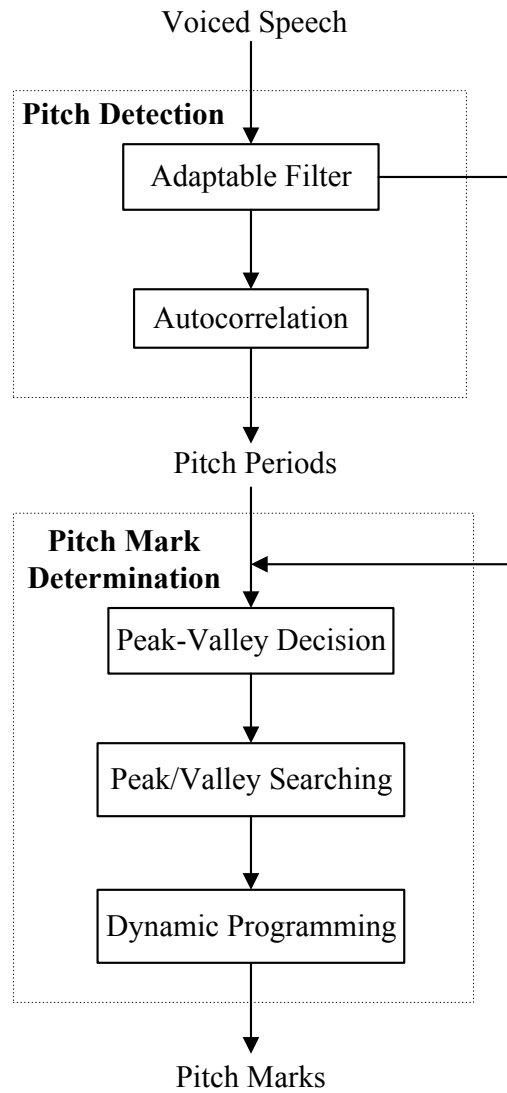
A speech synthesizer with various voices is useful for speech synthesis. Sometimes, it is also important for a service-providing company to have a synthesizer with the voice of its own employee or its favorite speaker. For conventional TTS systems, however, it is a demanding and tedious job to create a new voice. Recently, corpus-based TTS systems have been developed which use a large number of speech segments. Some approaches select speech segments as candidates for synthesis units. Establishing synthesis units involves speech segmentation, pitch estimation, pitch marking, and so on. Moreover, pitch marking is very labor-intensive task if no automatic mechanism is available.

In general, there are two major tasks in pitch marking: pitch detection and location determination. Compared to the literature on pitch detection [Rabiner *et al.* 1976][Rabiner 1977][Noll 1967][Markel 1972][Barnard *et al.* 1991][Kadambe *et al.* 1991][Barner 2000][Huang *et al.* 2000], few papers have focused on pitch marking [Moulines *et al.* 1990][Kobayashi *et al.* 1998], which is also a difficult problem because of the great variability of speech signals. Moulines *et al.* [Moulines *et al.* 1990] proposed a pitch-marking algorithm based on the detection of abrupt changes at glottal closure instants. In each period, they assumed that the speech waveform could be represented by the concatenation of the response of two all-pole systems. On the other hand, Kobayashi *et al.* [Kobayashi *et al.* 1998] used dyadic wavelets for pitch marking. The glottal closure instant was detected by searching for a local peak in the wavelet transform of the speech waveform.

In this paper, we propose a pitch-marking method based on an adaptable filter and a peak-valley estimation method. The block diagram of our method is shown in Fig. 1. The input signals are limited to voiced speech because only the periodic parts are of interest. We introduce an adaptable filter, which serves as a bandpass filter, to transform voiced speech into a sine-like wave. FFT (Fast Fourier Transform) is used to transform voice to the frequency domain, and the filter's pass band is determined by finding the spectral peak of the fundamental frequency. Consequently, the pass band can be adapted based on the fundamental frequency. The autocorrelation method is then used to estimate the pitch periods on the sine-like wave. In addition, a peak-valley decision method is employed to determine which part of the voiced speech is suitable for pitch mark estimation. The positive part (the speech with positive amplitude) and the negative part (the speech with negative amplitude) are investigated in this method. This is demonstrated by Fig. 3(a), which shows an example of a waveform having a negative part that reveals explicit periodicity. In general, it is possible to achieve better speech quality if the pitch marks are labeled at the positions of the extreme



points (peaks and valleys) of speech. In each pitch period, two possible peaks/valleys are searched. Finally, the pitch marks are obtained through dynamic programming by calculating the degree of pitch distortion.



**Figure 1** Block diagram of the proposed pitch-marking method.

## 2. Pitch Detection Using an Adaptable Filter Followed by Application of the Autocorrelation Method

The proposed adaptable filter serves as a bandpass filter in which the pass band extends from 50 Hz to the detected fundamental frequency, up to 500 Hz, of the voiced speech. First, we will define the following symbols, which are used in this algorithm:

$N$ : frame size in sample. Consecutive frames do not overlap.

$s_m[n]$ : the voiced speech of the  $m$ -th frame, where  $0 \leq n < N$ .

$SF_m[k]$ : the frequency response of  $s_m[n]$ , where  $0 \leq k < N$ .

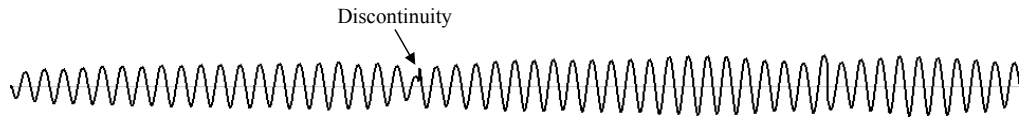
$YF_m[k]$ : the pass band frequency response of  $SF_m[k]$ , where  $0 \leq k < N$ .

$o_m[l]$ : the adaptable filter's output signal of the  $m$ -th frame, where  $0 \leq l < N$ .

The algorithm of the adaptable filter is described as follows:

- Step 1. Use FFT to transform the signal  $s_m[n]$  to obtain the frequency response  $SF_m[k]$ .
- Step 2. Find the position  $k_p$  of the spectral peak of the fundamental frequency for  $SF_m[k]$  by searching the first forty points of  $|SF_m[k]|$ .
- Step 3. Decide on the filter's pass band. Let  $YF_m[k] = SF_m[k]$  if  $3 \leq k \leq k_p + 2$  or  $3 \leq N - k \leq k_p + 2$ ; otherwise, let  $YF_m[k] = 0$ .
- Step 4. Normalize  $YF_m[k]$  by multiplying a scale of  $Max_k(|YF_m[k]|) / |YF_m[k_p]|$ .
- Step 5. Use IFFT (Inverse FFT) to transform the normalized  $YF_m[k]$  to the time domain. Let  $o_m[n]$  be the real part of the time domain signal.

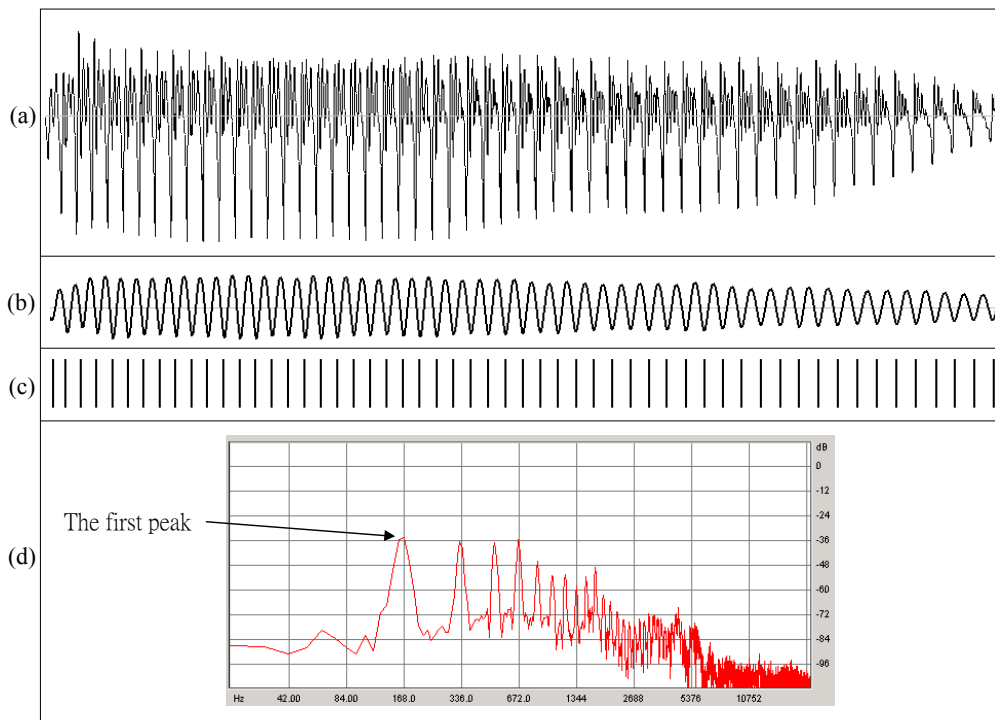
Finally, the refined pitch periods are obtained by analyzing the filtered speech  $o[n]$  using the conventional autocorrelation method. The waveform of  $o_m[n]$  after IFFT may be discontinuous at the frame boundaries. A typical example is shown in Fig. 2. However, such waveform discontinuity is not very significant and does not significantly affect the results of pitch period estimation.



**Figure 2** A typical example of waveform discontinuity after IFFT.

An example of an adaptable filter is displayed in Fig. 3. Panels (a) and (b) show the waveforms of the original speech and the filtered speech, respectively. It can be seen that the

filtered speech is generally a sine-like wave with clear periodicity than the original speech waveform. For a frame in the middle of the voiced speech, the spectral contour is depicted in panel (d). Note that the frequency axis is not linearly plotted to allow inspection of the first spectral peak. The first peak was found at 168 Hz, which was the fundamental frequency. Finally, the pitch periods were obtained by analyzing the filtered speech using the conventional autocorrelation method.



**Figure 3** Results obtained using the adaptable filter and pitch mark determination. (a) Waveform of the voiced speech with explicit periodicity in the negative part. (b) Waveform of the filtered speech. (c) Detected pitch marks. (d) Spectral contour (note that the frequency axis is not linearly plotted).

### 3. Pitch Mark Determination Using a Peak-Valley Decision Method and Dynamic Programming

#### 3.1 Peak-Valley Decision

From observations, we have found that voiced speech,  $s[\cdot]$ , is synchronous with filtered

speech,  $o[\cdot]$ , either at peaks or at valleys. The cases illustrated in Figs. 3 (a) and 2 (b) are synchronous at valleys having explicit periodicity instead of at peaks. As a result, the pitch marks can be more easily determined in the negative part than in the positive part. In the following, the peak-valley decision method is used to calculate two costs by summing the amplitudes of  $s[q]$ , where  $q$  represents the position of the local extreme point of  $o[\cdot]$  over each pitch period:

$$C_{peak} = \frac{1}{N_{peak}} \cdot \sum_{n=1}^{N_{peak}} s[Pos_{peak}[n]], \quad (1)$$

$$C_{valley} = \frac{-1}{N_{valley}} \cdot \sum_{n=1}^{N_{valley}} s[Pos_{valley}[n]], \quad (2)$$

where the symbols are defined as follows:

$C_{peak}$  : cost estimated at the peaks of  $o[\cdot]$ .

$C_{valley}$  : cost estimated at the valleys of  $o[\cdot]$ .

$N_{peak}$  : total number of the peaks of  $o[\cdot]$ .

$N_{valley}$  : total number of the valleys of  $o[\cdot]$ .

$Pos_{peak}[n]$  : position of the  $n$ -th peak of  $o[\cdot]$ .

$Pos_{valley}[n]$  : position of the  $n$ -th valley of  $o[\cdot]$ .

The peak-valley decision is made as follows: If  $C_{peak} > C_{valley}$ , then the positive part (peak) of  $s[\cdot]$  is adopted for evaluation of the pitch marks. Otherwise, the negative part (valley) of  $s[\cdot]$  is adopted.

### 3.2 Pitch Mark Determination Based on Dynamic Programming

Once the peak or valley, say the peak, has been adopted, the positions of the pitch marks are determined by picking the peaks of  $s[\cdot]$ . For a speech segment with a length of one pitch period, the PSOLA method can be used to synthesize good quality speech if the pitch mark is denoted at the signal with the largest amplitude. However, the largest peak may not correspond to the largest one in the next period (as shown in Fig. 4). This inconsistency will result in unpleasant speech after the PSOLA method is used. Therefore, the two highest peaks in each period are searched in pitch mark determination. We do not use three peaks or more because this would improve the performance very little. In this paper, we consider that a peak is located at the signal with the largest amplitude among consecutive positive signals. Among

peaks, the highest peak is the one with the largest amplitude. The second highest peak is the highest of the two peaks, the left-side and the right-side peaks, neighboring the highest peak.

For the  $i$ -th pitch period,  $P_i$ , suppose the highest and the second highest peaks are located at  $L_{i1}$  and  $L_{i2}$ , respectively. It might occur that the second one is absent. In this case, we let  $L_{i2} = L_{i1}$ . For all the detected peaks, pitch mark determination is then performed based on dynamic programming. The distortion of the pitch period,  $d_i(j,k)$ , and its accumulation,  $A_i(j)$ , are defined as follows:

$$d_i(j,k) = \left| L_{ij} - L_{(i-1)k} \right| - P_i + g(j,k), \text{ for } i=2, \dots, PN, \quad (3)$$

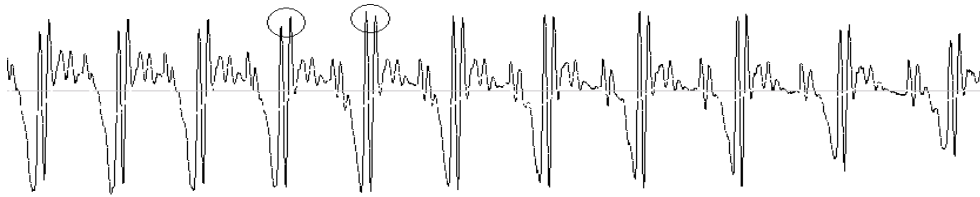
$$A_i(j) = \min \left\{ \begin{array}{l} d_i(j,1) + A_{i-1}(1), \\ d_i(j,2) + A_{i-1}(2) \end{array} \right\}, \text{ for } i=2,3, \dots, PN, \quad (4)$$

where  $PN$  is the total number of pitch period and  $j, k=1,2$ . In Equation (3),  $g(j,k)$  is a penalty function represented by

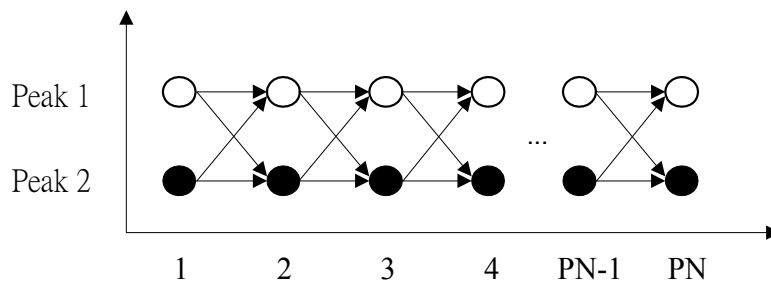
$$g(j,k) = \begin{cases} 0, & \text{if } j = 1 \text{ or } k = 1 \\ \frac{1}{PN}, & \text{otherwise} \end{cases} . \quad (5)$$

The penalty function is introduced here due to the preference for the highest peak.

The search path of the dynamic programming is illustrated in Fig. 5. The peak locations (pitch marks) can be obtained by back tracing the peak sequence corresponding to the smallest values of  $A_i(1)$  and  $A_i(2)$ . An example of the results of pitch marking is shown in Fig. 3(c). A procedure similar to that described above can be applied for the case of a ‘‘valley.’’



**Figure 4** An example of a waveform (syllable /a/ of tone 3), in which the largest peak does not correspond to the largest one in the next period (indicated by the circles).



**Figure 5** Illustration of the peak-picking search path of the dynamic programming.

## 4 Experiments and Results

### 4.1 Experimental Environment

A continuous speech database was established which provides the basic synthesis units of our Mandarin Chinese TTS system. This database is composed of 70 phrases, and their lengths are from 4 to 6 Chinese characters. It includes a total of 436 tonal syllables comprising the required 413 basic synthesis units. A native female speaker read them in normal speaking style. The speech signals were then digitized by a 16-bit A/D converter at a 44.1k Hz sampling rate. Syllable segmentation was done manually in order to obtain the precise boundaries of the voiced speech and unvoiced speech. The total duration of the 436 voiced speech segments was about 2.1 minutes. For each syllable, the voiced speech was used to test the proposed methods. The frame size used in the adaptable filter was set to 4096 speech samples (92.8 ms). We used large frame size so that we could deal with signals with very low  $f_0$  values.

For the voiced speech, the waveforms along with the pitch marks obtained using our pitch-marking program were visually displayed. The pitch marks were then checked and corrected by an experienced person through a friendly interface. For evaluation of the experiments, we obtained 436 sets of human-labeled pitch marks, denoted as  $H$ , which comprises 23,868 pitch marks.

### 4.2 Performance of the Pitch Marking Method

The peak-valley decision results were verified by human judgment based on visual displays. A success rate of 99.1% was obtained (4 of the 436 results disagreed). For the female speaker, we found that 97.2% of the voiced segments revealed clear periodicity in the negative parts.

The proposed method generated 23,860 pitch marks, denoted as  $I$ , without any duplication. The success rate of the pitch marking method is calculated as follows:

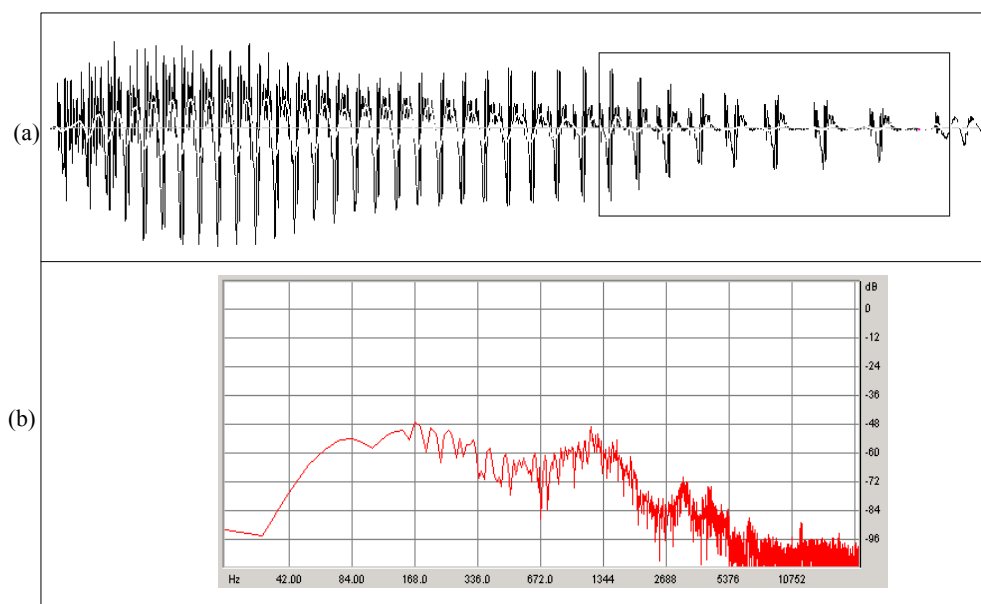
$$\text{Correct rate} = \frac{|\{x \mid x \in I \text{ and } x \in H\}|}{|H|} \times 100\% . \quad (6)$$

As shown in Table 1, a success rate of 97.2% was obtained (baseline), in contrast with 95% and 97% success rates obtained using the methods proposed in [Moulines *et al.* 1990] and [Kobayashi *et al.* 1998], respectively. Moreover, we found that most of the errors resulted from incorrect pitch detection results. Most of the pitch errors were due to large changes of pitch located at the boundaries of the voiced speech. With correct pitch information provided, our method achieved a success rate of 99.5%.

The tone type of voice significantly affects the results of the detection of  $f_0$ . The main reason for error detection of  $f_0$  is dependent on the tone types of voice. There are five tones in Mandarin speech, including a high-level tone (Tone 1), a mid-rising tone (Tone2), a mid-falling-rising tone (Tone 3), a high-falling tone (Tone 4), a neutral tone (Tone 5). In our system, it is easy to detect  $f_0$  for tone 1 and tone 2 since the spectral peak of  $f_0$  is prominent (Fig. 3 (d)). For tone 3, tone 4 and tone 5,  $f_0$  may be erroneously detected at the end of the voice segment if the consecutive pitch periods change abruptly (Fig. 6 (a)). For this case, the spectral peak of  $f_0$  is unclear (Fig. 6 (b)), which may result in error detection.

**Table 1.** Success rate of the pitch-marking method.

Condition	Baseline	Using correct pitch
Success rate	97.2%	99.5%



**Figure 6** An example of unclear spectral peaks. (a) Waveform of the syllable /a/ of tone 3. (b) Spectral contour corresponding to the end part of the waveform (note that the frequency axis is not linearly plotted).

## 5 Conclusions

In this paper, a preliminary work on pitch marking has been proposed. We have presented an adaptable filter which can be combined with the autocorrelation method to perform pitch detection. On the other hand, a peak-valley decision method has been proposed to select either the positive or the negative part for pitch mark evaluation. Also, a dynamic-programming-based pitch mark determination method has been demonstrated, where two peaks/valleys are searched in each period. In the experiments, our pitch-marking method achieved a 97.2% success rate. Furthermore, a high success rate of 99.5% was obtained when correct pitch information was provided.

## Acknowledgement

This paper is a partial result of Project 3XS1B11 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C.



## **References**

- Hamon, C., E. Moulines, and F. Charpentier, "A diphone synthesis based on time-domain prosodic modifications of speech," *International Conference on Acoustics, Speech, and Signal Processing*, 1989, pp.238-241.
- Iwahashi, N. and Y. Sagisaka, "Speech segment network approach for optimization of synthesis unit set," *Computer Speech and Language*, 1995, pp.335-352.
- Shih, C. L. and R. Sproat, "Issues in text-to-speech conversion for Mandarin," *Computational Linguistics and Chinese Language Processing*, vol.1, 1996, pp.37-86.
- Chen, S. H., S. H. Hwang and Y. R. Wang, "An RNN-based prosodic information Synthesizer for Mandarin text-to-speech," *IEEE Transactions on Speech and Audio Processing*, 6(3), 1998, pp. 226-239.
- Chou, F. C. and C. Y. Tseng, "Corpus-based Mandarin speech synthesis with contextual syllabic units based on phonetic properties," *International Conference on Acoustics, Speech, and Signal Processing*, 1998, pp.893-896.
- Charpentier, F. J. and M. G. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," *International Conference on Acoustics, Speech, and Signal Processing*, 1986, pp. 2015-2020.
- Rabiner, L. R., M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A Comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics., Speech and Signal Processing*, 24, 1976, pp. 399-417.
- Rabiner, L. R., "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics., Speech and Signal Processing*, 25, 1977, pp. 24-33.
- Noll, A. M., "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, 47, 1967, pp. 293-309.
- Markel, J. D., "The SIFT algorithm for fundamental frequency estimation," *IEEE Transactions on Audio Electroacoustics*, Au-20, 1972, pp. 367-377.
- Barnard, E., R. A. Cole, M. P. Veal, and F. A. Alleva, "Pitch detection with a neural-net classifier," *IEEE Transactions on Signal Processing*, 39(2), 1991, pp. 298-307.
- Kadambe, S., G. F. Boudreaux-Bartels, "A comparison of a wavelet functions for pitch detection of speech signals," *International Conference on Acoustics, Speech, and Signal Processing*, 1991, pp. 449-452.
- Barner, K. E., "Colored L-l filters and their application in speech pitch detection," *IEEE Transactions on Signal Processing*, 48(9), 2000, pp. 2601-2606.
- Huang, H. and F. Seide, "Pitch tracking and tone features for Mandarin speech recognition," *International Conference on Acoustics, Speech, and Signal Processing*, 2000, pp.1523-1526.

- Moulines, E., F. Emerard, D. Larreur, J. L. Le Saint Milon, L. Le Faucheur, F. Marty, F. Charpentier, and C. Sorin, "A real-time French text-to-speech system generating high-quality synthetic speech," *International Conference on Acoustics, Speech, and Signal Processing*, 1990, pp.309-312.
- Kobayashi, M., M. Sakamoto, T. Saito, Y. Hashimoto, M. Nishimura, and K. Suzuki, "Wavelet analysis used in text-to-speech synthesis," *IEEE Transactions on Circuits and Systems-II, Analog and Digital Signal Processing*, 45(8), 1998, pp. 1125-1129.

## 統計式片語翻譯模型

### Statistical Translation Model for Phrases

張俊盛\*

游大緯\*

李俊仁\*+

Jason S Chang, David Yu, Chun-Jun Lee

#### 摘要

機器翻譯是自然語言處理研究上最重要的課題之一，在過去運用機器翻譯比較成功的例子，多是特定的領域文件的翻譯。近來因為網際網路與搜尋引擎的盛行，大家開始重視機器翻譯在跨語言檢索（Cross Language Information Retrieval）中的角色。在跨語言檢索的問題上，通常是對查詢字詞或片語，進行翻譯（Query Translation）。然而翻譯的結果必須和欲搜尋的文件庫有高度的相關性，才能達到檢索的效果。目前翻譯查詢關鍵詞的做法，無論是採用現成的翻譯軟體，或者使用一般性的雙語詞典，都很難確保產生和文件相關的翻譯。因此我們希望能夠透過統計式片語機器翻譯（Statistical Phrase Translation Model, SPTM）的做法來進行查詢關鍵詞的翻譯，以提高跨語言檢索的效率。在這篇論文中，我們提出新的統計式片語翻譯模型，並進行實驗。實驗中我們利用 BDC 雙語電子辭典實驗以 SPTM 進行片語內的詞彙對應。以 SPTM 產生對應分析，比較快速，而且正確率比較高。

#### Abstract

Machine Translation is one of the most difficult problems in the field of natural language processing. In the past, MT has been applied to professional communication in the process of translating technical and corporate document on a specific domain. Recent years saw the rapid development of Internet as a new form of communication and information exchange, and the need to access information across the language barrier became apparent. People began to look into the role that MT can play in Cross Language Information Retrieval. The prevalent approach to CLIR is based on translation of query, in particular query

---

\*國立清華大學資訊工程研究所 E-mail: jschang@cs.nthu.edu.tw

+中華電信研究所

phrases. However, for CLIR there is an additional new objective of translating into something that is relevant to the collection being searched upon. Therefore, the current approach of using general bilingual word list or an off-the-shelf commercial MT software is bound to be very ineffective in terms of retrieving relevant documents. We propose a new approach to Statistical Phrase Translation Model (SPTM), aimed at achieving a tighter estimation of phrase translation. Experiments were conducted using bilingual phrases in BDC Electronic Chinese-English Dictionary. Preliminary results shows the approach is much faster and produces better word alignment for phrases, which has not been possible using previous approaches.

**Keywords:** Statistical Machine Translation; Phrase Translation; Cross-language Information Retrieval.

## 1. 簡介

機器翻譯是自然語言處理研究上最重要的課題之一，有助於幫助使用者跨越語言與文化的障礙。在過去運用機器翻譯比較成功的例子，多是特定的領域文件的翻譯，如技術性的使用手冊、氣象報告、國際機構的官方文件。近來因為網際網路與搜尋引擎的盛行，大家開始重視機器翻譯在機器輔助翻譯（Machine Assisted Human Translation）[Lange, Gaussier, and Daille, 1997]、跨語言檢索（Cross-Language Information Retrieval）[Gey and Chen, 1997] 及電腦輔助語言學習（Computer Assisted Language Learning）[Shei and Pain, 2001] 可能扮演的角色。

在特定領域的文件翻譯上，機器翻譯系統主要是以句子為單位，進行處理。在跨語言檢索的問題上，可以採取「文件翻譯」（document translation），或者「查詢資訊翻譯」（query translation）的做法 [McCarley 1999]。目前大部分的研究者都採取翻譯查詢關鍵詞的做法。例如，在 NTCIR-2 的英到中的資訊檢索評估活動 [Kando *et al.* 2001] 中的一個查詢主題中，就提供以下的英文關鍵詞，試驗參與的系統，找到相關中文新聞文件的能力：

- Assembly Parade Law
- Parade and Demonstration
- Constitution
- Freedom of speech
- Communism
- Council of Grand Justices
- Legislation
- Amendments

查詢關鍵詞的翻譯涉及詞彙語義解析 (Word Sense Disambiguation) 的問題 [Ide and Veronis 1998, Chen and Chang 1998] 與片語的翻譯 (Phrase Translation) 的問題, 和技術文件翻譯很重要的不同點, 在於翻譯的結果, 是要拿來在檢索系統的文件庫 (Text Collection) 中搜尋文件。所以翻譯的詞義解析與翻譯的詞彙選擇 (Lexical Choice) 必須和文件庫的語料有高度的相關性。以上述關鍵詞中的 demonstration 為例, 我們就必須翻譯成新聞中常見的「示威」而不能翻譯成「示範」。

目前學者研究跨語言檢索的主要做法, 大致為:

1. 利用市場上販售的翻譯軟體 [Gey and Chen 1997, Kwok 2001]
2. 使用一般性的雙語詞典 [Oard 1999, Kwok 2001]

這兩種做法, 很明顯的都不容易產生和文件庫相關的翻譯。這一點對於音譯的專有名詞, 特別明顯。Kwok 就指出使用現成翻譯軟體和一般性雙語詞典, 不能得到 Michael Jordan 在文件庫的正確音譯「麥可喬丹」, 顯然是跨語言檢索研究的一大問題。

為了提高翻譯和文件庫的相關性, Chen 等 [1999] 將詞彙共現機率 (occurrence statistics) 導入翻譯詞彙選擇的考慮中。有鑒於音譯專有名詞在跨語言檢索的重要性, 也有研究者提出了一些統計或規則式的做法, 將英文中音譯的日、中人名地名轉換回原文的專有名詞 [Knight and Graehl 1997, Chang *et al.* 2001]。這些做法, 雖然對於跨語言檢索有一定的效果, 但缺乏比較全面性, 也不具備嚴謹的理論基礎, 因此影響到改進檢索效率的空間。

我們認為要做好跨語言檢索中的查詢關鍵詞的翻譯, 必須有一套全面而嚴密的方法, 發展適用的機器翻譯模型。在機器翻譯的做法中, 範例為本做法 (Example-based Approach) 和統計式機器翻譯, 都是比較資料導向 (data-driven) 的做法, 比較能夠產生和資訊檢索文件庫相關的翻譯。統計式翻譯模型 (Statistical Translation Model) 的應用於文件的機器翻譯 [Jones and Havrilla, 1998]、翻譯語料之詞彙對應 [Gale and Church 1992, Melamed 2000]、字典建構 [Melamed 2000]。IBM Watson 研究中心的 Brown 等 [1988, 1990, 1993] 最早提出理論嚴謹的統計式機器翻譯做法。Wu [1997] 提出以無標記二元倒裝樹之句法結構為基礎的統計式翻譯模型。Wang [1998] 和 Och 等人 [1999] 採用片語與樣板, 導入句法結構於統計式的翻譯模型。Yamata 和 Knight [2001] 則提出完整的句法導向的統計式翻譯模型, 以規範例兩種語言的句法剖析樹的對應關係。其模型包括剖析樹之子樹的順序重排, 功能詞彙節點的增加與刪除。

我們希望能夠透過一種新的統計式對應與機器翻譯做法 (Statistical Alignment and Machine Translation) 來進行查詢關鍵詞的翻譯, 為跨語言檢索的查詢詞翻譯提供一個比較有效而且嚴謹的做法。在這篇論文中, 我們提出一種新的翻譯指派機率 (Assignment Probability) 的做法, 並進行實驗。實驗的結果證實新的模型的確能改進片語對應與翻譯的效率。

## 2. 統計式機器翻譯模型

機器翻譯早期是以逐字翻譯加上局部的位置調整的直接做法 (Direct Approach)，後來逐漸轉成主要是以句法分析為基礎的轉換式的做法 (Transfer Approach)。在 1980 年代末，研究的趨勢比較傾向實證式的做法 (Empirical Approach)，以翻譯的範例或平行語料庫為本，發展機器翻譯系統。Brown 提出的語料庫為本之統計式做法，在理論的架構最為完備。在 Brown [1993] 的統計式機器翻譯模型 Model 3 下，原文  $S$  和譯文  $T$  的翻譯機率 (Translation Probability)  $Pr(T|S)$ ，可以分解成以下的三個機率函數：

- (a) 詞彙翻譯機率 (Lexical Translation Probability)

$$Pr(T_j | S_i)$$

- (b) 孳生機率 (Fertility Probability)

$$Pr(a | S_i)$$

- (c) 位置扭曲機率 (Distortion Probability)

$$Pr(j | i, k, m)$$

其中

$S_i$  為  $S$  的第  $i$  個字

$T_j$  為  $T$  的第  $j$  個字

$a$  為  $T_j$  的長度

$k$  為  $S$  的長度

$m$  為  $T$  的長度

Brown 等使用加拿大國會議事錄的英法平行語料庫，證實透過反覆交替的「期望值估計」與「最佳化」演算法 (Expectation and Maximization Algorithm)，可以得到這三個簡單的機率函數的統計估計值。其「期望值估計」的步驟，就是在目前的機率函數估計值下，求取所有翻譯對應的機率值。而「最佳化」的步驟，就是以所有的雙語語料樣本的翻譯對應為根據，估計三個機率函數的最大概似估計值 (Maximum Likelihood Estimation)。

透過 EM 演算法，統計式機器翻譯模型中的翻譯機率函數的估計值可趨於收斂。在雜訊通道模型 (Noisy Channel Model) 下，結合翻譯機率函數，與目標語的 N-gram 語言模型 (Language Model)，可以用搜尋演算法，如束限搜尋法 (Beam Search) 求最佳機率值的方式，產生翻譯。

### 3. 適用於片語對應與翻譯的統計式模型

IBM Model 3 中的位置扭曲機率，是基於每一字的翻譯目標位置和其他字無關的假設。在獨立事件的假設下，某一個翻譯對應（alignment）方式的機率，在位置方面而言，是所有字的和對應字的位置形成的位置扭曲機率值的乘積。實際上，每一字的翻譯目標位置和其他字的翻譯位置有高度的相關性。如果  $S_i, i' \neq i$  都不對應到  $T_j$ ，則  $S_i$  對應到目標位置  $j$  的機率幾乎為 1

$$Pr(j|i, k, m) \approx 1 \text{ 若 } Pr(j|i', k, m) = 0, i' \neq i$$

因此獨立假設下的機率，幾乎大部分的情況下會造成過低的估計。即便是很可能的翻譯對應方式，其機率值經過一連串位置扭曲機率的乘積，常趨於不合理的低數值。例如，檢視三字英文與五字中文的互譯片語樣本，最可能翻譯對應  $A^*$  下的三個字  $S_1 S_2 S_3$  翻譯目標位置，分別是

$$S_1 \rightarrow \{T_1, T_2\}$$

$$S_2 \rightarrow \{T_3, T_4\}$$

$$S_3 \rightarrow \{T_5\}$$

也就是  $A^* = (0, 12, 34, 5)$ （第一個 0 代表所有的中文字都對應到一個英文字，沒有中文字無法對應的情況）。在  $k = 3$  及  $m = 5$  的片語樣本中，翻譯對應為  $A^*$  的情況約佔 35%。直接估計  $A^*$  的最大概似估計值（Maximum Likelihood Estimation），得到

$$Pr_{MLE}(A^*) = 0.35$$

然而在機率獨立的假設下

$$Pr(A^*) = P(1|1, 3, 5) P(2|1, 3, 5) P(3|2, 3, 5) P(4|2, 3, 5) P(5|3, 3, 5)$$

即使位置扭曲機率值以高數值（0.6）估計  $P(j|i, 3, 5)$ ，其乘積仍然遠低於合理的估計值：

$$Pr(A^*) < (0.6)^5 = 0.046656 \ll 0.35$$

為了更精確合理的估計翻譯目標位置的機率，我們提出了直接估計整體翻譯配對位置與字數的做法。在此做法下，孳生機率和位置扭曲機率合併成為指派機率（Assignment Probability）。因此不再獨立考慮個別的字的位置、翻譯目標位置、孳生的字數，而是以整體的對應來一併考慮。在這樣的想法下，我們將原文  $S$  和譯文  $T$  的翻譯機率  $Pr(T|S)$ ，分解成以下的兩個機率函數：

- (a) 詞彙翻譯機率（Lexical Translation Probability）

$$Pr(T(A_i) | S_i)$$

(b) 指派機率 (Assignment Probability)

$$Pr(A | k, m) = Pr(A_0, A_1, A_2, \dots, A_k | k, m)$$

其中

$S_i$  為  $S$  的第  $i$  個字

$T(A_i)$  為  $T$  中對應到  $S_i$  的部分

$A_0$  為  $T$  中沒有對應到  $S$  的部分的標號

$A_i$  為  $T$  中對應到  $S_i$  的部分的標號,  $i > 0$

$k$  為  $S$  的長度

$m$  為  $T$  的長度

#### 4. 實驗

我們進行了一系列的實驗，以驗證我們提出的新的片語翻譯模型的效果與可行性。透過實驗，我們想了解新模型有關的下列幾個問題：

1. 以指派機率替代孳生機率和位置扭曲機率，是否可以得到較正確的對應分析？
2. 指派的位置是否集中在幾種樣式，而不是許多個別對應目標位置的排列組合？指派機率的參數量，會不會過多，會不會導致估計的速度過慢？
3. 指派機率的參數量和樣本數量，相較之下，其機率值的統計可靠度會不會過低？
4. 訓練後的機器翻譯模型，應用到跨語言檢索的可行性高或低？

##### 4.1 實驗的設計與起始機率值的設定

由於不易取得大量雙語片語的語料，我們採用 BDC 漢英字典 [BDC 1992] 的片語條目作為實驗的原始材料。為了配合實驗的目標，並簡化問題，我們首先去掉英文多於 3 個詞的條目，但中文長度不限。因為，4 字詞（含）以上之條目僅佔訓練語料 4% 不到，不足以求得有意義且具代表性模型參數。另外我們也去掉中文的四字成語條目。這些條目的翻譯，常常不是字面翻譯，去掉之後，可以降低資料的雜訊。原始資料經過整理之後，我們得到 96,156 筆可用的英中片語翻譯的記錄。我們以  $(P_n, Q_n)$ ,  $n = 1, N$  來代表這組英中片語翻譯語料。

在試驗中，我們以 EM 演算法，來得到第三節所提出的辭彙翻譯機率、指派機率。我們採取了和一般不同，但類似 Och 等人 [2000] 對於 IBM 機率模型的改進實驗的做法。其目的都是希望加速機率的估計。



1. 開始的時候，我們採用 IBM Model 2，以詞彙翻譯機率與位置扭曲機率，來估計訓練統計模型的機率參數。在 EM 演算法的第二輪之後才開始使用新模型的指派機率。
2. 我們假設英中片語翻譯時，英文和中文字的順序一致的機會較高。所以第一輪運算機率模型的位置扭曲機率不用一般常用的平均分布  $Pr(j|i, k, m) = 1/m$ ，而採用參數式的統計法，根據片語翻譯傾向於保留原文順序的的經驗法則，令位置扭曲機率的值如下：

$$Pr(j|i, k, m) = 1 - \left| \frac{j-0.5}{m} - \frac{i-0.5}{k} \right| \quad (1)$$

其中  $i$  = 英文字位置， $k$  = 英文字總數， $j$  = 中文字位置， $m$  = 中文字總數。對於公式[1]的機率值，需要再調整，使得在  $i, k, m$  值固定時，對所有的  $j$  值， $Pr(j|i, k, m)$  的加總為 1。

表 1. 位置扭曲機率的起始估計值

$S_i$	$T_i$	$i$	$k$	$j$	$M$	$Pr(j i, k, m)$
flight	8	1	2	1	4	0.318
flight	字	1	2	2	4	0.318
flight	飛	1	2	3	4	0.227
flight	行	1	2	4	4	0.136
eight	8	2	2	1	4	0.136
eight	字	2	2	2	4	0.227
eight	飛	2	2	3	4	0.318
eight	行	2	2	4	4	0.318

對於每一筆雙語片語，我們假設每個英文字可以翻譯成其中任何一個中文字，但是其機率會因位置不同而異。例如某一筆記錄是 2 個英文字翻譯成 4 個中文字，我們可以得到 8 個英中文字的任意配對。每一個配對的位置扭曲機率如公式 1 的  $Pr(j|i, k, m)$  值。例如，對語料中雙語片語 (flight eight, 8 字飛行)，我們用公式 1 可以計算得到如表 1 的任意詞彙配對的位置扭曲機率。

表 2. 位置扭曲機率與詞彙翻譯機率的估計值

$S_i$	$T_j$	$i$	$k$	$j$	$m$	$Pr(j i, k, m)$	$Pr_{LEX}(C E)$	$Pr(T_j S_i)$
flight	8	1	2	1	4	0.318	0.00797	0.00253
flight	字	1	2	2	4	0.318	0.00797	0.00253
flight	飛	1	2	3	4	0.227	0.25770	0.05850
flight	行	1	2	4	4	0.136	0.16901	0.02299
eight	8	2	2	1	4	0.136	0.02903	0.00395
eight	字	2	2	2	4	0.227	0.04839	0.01098
eight	飛	2	2	3	4	0.318	0.06774	0.02154
eight	行	2	2	4	4	0.318	0.06774	0.02154

有了任意配對的位置扭曲機率後，我們就可據此估計語料庫片語中的任何英文字  $E$  和中文字  $C$  間的翻譯機率  $Pr_{LEX}(C|E)$ ，公式如下：

$$Pr(C|E) = \frac{\sum_{n=1}^N \sum_{i=1}^k \sum_{j=1}^m \delta(E, P_n(i)) \delta(C, Q_n(j)) Pr(j|i, k, m)}{\sum_{n=1}^N \sum_{i=1}^k \sum_{j=1}^m \delta(E, P_n(i)) Pr(j|i, k, m)} \quad (2)$$

其中  $P_n(i)$  為  $P_n$  之第  $i$  字， $Q_n(j)$  為  $Q_n$  之第  $j$  字， $k = |P_n|$ ， $m = |Q_n|$

$\delta(x, y) = 1$  若  $x = y$ ， $\delta(x, y) = 0$  若  $x \neq y$

公式 2 的用意在於加總  $E$  和  $C$  的在所有片語中的機率值，並除以  $E$  和所有中文  $C$  的機率值的總和，使得對所有的  $C$  值  $Pr(C|E)$  的機率值加總為 1。依據公式 1 和公式 2 的機率值，我們可以估計任何片語內任意字的配對  $(S_i, T_j)$  的機率值：

$$Pr(T_j|S_i) = Pr(j|i, k, m) Pr_{LEX}(C|E) \quad \text{其中 } C = T_j, E = S_i$$

以訓練語料中雙語片語 (flight eight, 8 字飛行) 為例。對於其中英文字 flight 與中文字「飛」的對應機率為  $Pr_{LEX}(\text{飛} | \text{flight}) Pr(3|1, 2, 4)$ 。表 2 列出表 1 的任意配對的詞彙翻譯機率。一般而言，起始的統計估計值相當的精確，如表二所顯示的機率值的估計，都相當合理。例如，正確的對應的對應機率如  $Pr(\text{飛} | \text{flight})$  和  $Pr(\text{行} | \text{flight})$  分別為 0.05850 與 0.02299，遠高於錯誤的對應的詞彙機率  $Pr(8 | \text{flight})$  和  $Pr(\text{字} | \text{flight})$  的 0.00253。

## 4.2 EM 演算法的第一輪計算

有了起始的機率函數估計值，我們就可以進行 EM 演算法來估計翻譯模型中的參數值。我們應用 Viterbi 式訓練的 EM 演算法。在每次重新估算時，依據每一個樣本的最佳對應，而不考慮每一個樣本的所有的可能對應。

### 第一次的對應最佳化

我們採取簡單的貪婪法 (Greedy Method) 來求取每一組雙語片語 ( $P_n, Q_n$ ) 的最佳對應。我們假設簡單的孳生模型：一個英文可以對應到 0 到多個中文字，而每個中文字只能對應到最多一個英文字。有了片語內的詞彙翻譯與位置扭曲機率的起始估計值與其乘積 (如表 2)，我們就可以對每一個中文字，逐次選取其對應到英文字機率值最高者，產生英文和中文字的配對，並根據假設的孳生模型，排除其他的英文字和此中文字的配對。反覆的執行上述步驟，直到沒有剩餘的中文字，或機率值低於某一個門檻值 (threshold) 為止。若有剩餘的中文字，就視為沒有對應到英文字。最低對應的機率門檻值，可以避免信賴度太低的錯誤對應，也有助於導入 0 對 1, 0 對多的孳生模式。經過實際抽樣觀察之後，以 0.008 為門檻值，可去掉大部分低信賴度的錯誤配對。再回到 “flight eight” 的例子，由表 2 的機率值，我們可得到如表 3 的對應方式 (0, 34, 12)。

表 3. (*flight eight*, 8 字飛行) 之最佳對應 (0, 34, 12)

$S_i$	$T_j$	$i$	$k$	$j$	$m$	$Pr(j i, k, m)$	$Pr_{LEX}(C E)$	$Pr(T_j S_i)$
flight	飛	1	2	3	4	0.227	0.25770	0.0585
flight	行	1	2	4	4	0.318	0.16901	0.05375
eight	字	2	2	2	4	0.227	0.04839	0.01098
eight	8	2	2	1	4	0.136	0.02903	0.00395

### 指派機率函數值的重新估算

經過機率最佳化求取最可能的對應方式後，我們就可以拋棄個別字的位置扭曲機率，導入新的翻譯指派機率模型，直接估計整個對應方式的機率值。我們依照片語的英中文字數，統計出英中文字數  $k$  與  $m$  固定下，各種指派方式  $A$  的機率：

$$\Pr(A|k, m) = \frac{\text{count}(A \text{ 為 } (S, T) \text{ 的對應})}{\text{count}(k = |S|, m = |T|)} \quad (3)$$

表 4. 兩字對四字片語的指派機率值最高的前 12 名

$k$	$m$	A			$Pr(A_{k,m})$
		$A_0$	$A_1$	$A_2$	
2	4	0	12	34	0.572025052
2	4	0	123	4	0.121317560
2	4	0	1	234	0.085479007
2	4	0	1234	0	0.078056136
2	4	0	0	1234	0.065066110
2	4	0	124	3	0.020992809
2	4	0	2	134	0.016585479
2	4	0	3	124	0.007886801
2	4	0	34	12	0.005915101
2	4	0	13	24	0.004059383
2	4	0	134	2	0.003363489
2	4	0	23	14	0.002551612

表 5. 二字到四字片語，最可能的 5 種指派方式的實例

	T	T(A <sub>0</sub> )	S <sub>1</sub>	T(A <sub>1</sub> )	S <sub>2</sub>	T(A <sub>2</sub> )
T-shaped antenna	T 形天線		T-shaped	T 形	antenna	天線
X-ray examination	X 光檢查		X-ray	X 光	examination	檢查
Irresistible force	不可抗力		irresistible	不可抗	force	力
Unwritten law	不成文法		unwritten	不成文	law	法
Central Asia	中亞細亞		Central	中	Asia	亞細亞
mutual non-interference	互不干涉		mutual	互	non-interference	不干涉
undesirable element	不良少年		undesirable	不良少年	element	
Unalterable truth	不易之論		unalterable	不易之論	truth	
come soon	不日放映		come		soon	不日放映
a desperado	不逞之徒		a		desperado	不逞之徒

在實驗中，EM 演算法的第一輪自動的發掘出 601 種指派方式。以兩字對四字片語而言，有 38 種方式。表 4 列出依照機率由高到低排列的前 12 名指派方式。請參考表 5 所列 2 對 4 字片語對應的實際例子。由表 4 可以觀察到幾點：

1. 機率估計的結果，和我們的認知沒有很大的出入：  
最可能的片語翻譯的順序是保留原文的順序。  
同一英文字翻譯的目標位置是連續的。  
一個英文字最可能翻譯到 2 個中文字。
2. 指派安排的機率值集中在少數的幾個樣式上。最可能的 3 種指派，佔了接近 80% 的機率。而前 5 種及 10 種指派方式，分別累積了 95% 及 99.5% 的機率值。這證明了應用指派機率，可以很有效的在雙語對應或機器翻譯時，限制搜尋的範圍，而不影響到精確性。
3. 指派機率函數收斂的速度很快。

表 6. “flight”翻譯成不同中文字串的機率

E	C	Pr ( C E )
flight	飛行	0.6480231012
flight	飛	0.1411528654
flight	航空	0.0602616768
flight	\$empty\$	0.0296114718
flight	航	0.0296114718
flight	分	0.0041786956
flight	分隊	0.0041786956
flight	飛班機	0.0041786956
flight	飛航	0.0041786956
flight	飛機	0.0041786956
flight	航飛	0.0041786956
flight	黑	0.0041786956
flight	群	0.0041786956
flight	\$any\$	0.0000009248

#### 詞彙翻譯機率值的重新估算

在統計指派方式的機率的同時，我們同樣的也拿 4.2 節最佳化的結果，估計英文字翻譯成不同中文字串的機率。我們採取和第一輪不一樣的做法，不再考慮英文字對應到中文單字的機率，而是考慮每一個英文字在片語中，所對應到的中文字串。這些中文字串大部分的情況是連續的，而且是詞典裡常見的詞項。當然也有少數的例子，英文的對應目標是空字串、不連續字串、不能獨用的詞素 ( bound morpheme ) 等等情況。我們以“\$empty\$”來代表英文字對應到空字串的情況。考慮資料不足 ( data sparseness ) 的可能，我們導入“\$any\$”來代表英文字對應到訓練外的任意中文字串的情況，並採用 Good-Turing 的平滑化方法 ( smoothing method ) 來估計\$any\$的翻譯機率。

表 6 列出 flight 翻譯成不同中文字串的機率，包括一般的詞、詞素、\$empty\$、\$any\$。在這一輪的期望值估計中，flight 對應到\$empty\$的機率估計值 0.0296114718 仍然過高。只要指派機率如表 4 的 ( 0,0,1234 ) 和 ( 1,0,234 ) 的機率，以及\$any\$機率的估計值估計得合理，我們期望在 EM 演算法的以後的幾個輪迴中，兩者互相競爭的情況下，\$empty\$機率的估計值會逐漸的降低，而趨近合理的區段。

### 4.3 EM 演算法的第二輪計算

在第一輪的期望值估計之後，我們可以再次的求取片語的最可能對應方式。在第二輪的運算當中，我們不再使用公式 1 的位置扭曲機率，而是採用已經估計出來的整體性的指派機率。

表 7.  $Pr(\text{8 字飛行} | \text{flight eight})$  機率值最高之前 5 名

S	T	$A_0$	$A_1$	$A_2$	$T(A_0)$	$T(A_1)$	$T(A_2)$	$Pr(\mathbf{T}, \mathbf{A}   \mathbf{S})$
flight eight	8 字飛行	0	34	12		飛行	8 字	0.0000788100
flight eight	8 字飛行	0	3	124		飛	8 字行	0.0000000051
flight eight	8 字飛行	12	34	0	8 字	飛行	\$empty\$	0.0000000007
flight eight	8 字飛行	12	3	4	8 字	飛	行	0.0000000003
flight eight	8 字飛行	2	3	14	字	飛	8 行	0.0000000001

第二輪運算中，我們對每一筆雙語片語 (S, T)，依據其英文和中文字數，考慮相符的所有的指派方式 A，計算其翻譯機率  $Pr(\mathbf{T}, \mathbf{A} | \mathbf{S})$ 。對於某一指派方式 A， $Pr(\mathbf{T}, \mathbf{A} | \mathbf{S})$  為 A 的機率和由 A 所決定的詞彙配對  $(S_i, T(A_i))$  的機率乘積：

$$Pr(\mathbf{T} | \mathbf{S}) = \max_A Pr(\mathbf{T}, \mathbf{A} | \mathbf{S}) = \max_A Pr(\mathbf{A} | k, m) \prod_{i=1}^k Pr(T(A_i) | S_i)$$

因此最可能的指派  $A^*$  可由下列公式決定

$$\begin{aligned} A^* &= \arg \max_A Pr(\mathbf{T}, \mathbf{A} | \mathbf{S}) \\ &= \arg \max_A Pr(\mathbf{A} | k, m) \prod_{i=1}^k Pr(T(A_i) | S_i) \end{aligned} \quad (4)$$

其中  $k = |\mathbf{S}|, m = |\mathbf{T}|$

以  $(\mathbf{S}, \mathbf{T}) = (\text{flight eight}, \text{8 字飛行})$  為例，對於不同的指派 A，其翻譯機率的計算如下：

$A = (0, 12, 34)$ ：

$$Pr(\mathbf{T}, \mathbf{A} | \mathbf{S}) = Pr(0, 12, 34 | 2, 4) Pr(\text{8 字} | \text{flight}) Pr(\text{飛行} | \text{eight})$$

$A = (0, 34, 12)$ ：

$$Pr(\mathbf{T}, \mathbf{A} | \mathbf{S}) = Pr(0, 34, 12 | 2, 4) Pr(\text{飛行} | \text{flight}) Pr(\text{8 字} | \text{eight})$$

$A = (0, 3, 124)$ ：

$$Pr(\mathbf{T}, \mathbf{A} | \mathbf{S}) = Pr(0, 3, 124 | 2, 4) Pr(\text{飛} | \text{flight}) Pr(\text{8 字行} | \text{eight})$$

$A = (2, 34, 1)$  :

$$Pr(\mathbf{T}, \mathbf{A} | \mathbf{S}) = Pr(2, 34, 1 | 2, 4) Pr(\text{飛行} | \text{flight}) Pr(\text{8 leight})$$

$A = (12, 34, 0)$  :

$$Pr(\mathbf{T}, \mathbf{A} | \mathbf{S}) = Pr(12, 34, 0 | 2, 4) Pr(\text{飛行} | \text{flight}) Pr(\text{\$empty\$leight})$$

在計算過程中，我們採取 Branch and Bound 的搜尋法，以降低搜尋的時間。我們會記錄該筆雙語片語翻譯機率  $Pr(\mathbf{T}, \mathbf{A} | \mathbf{S})$  的在搜尋過程的最大值，當逐一考慮相符的所有的指派方式  $A$  時，若指派機率值已小於當時的翻譯機率最大值時，該指派方式  $A$  則不予計算，因為再繼續乘以詞彙翻譯機率只會讓機率更小，所以可以予以捨棄，如此可以大幅的增進 EM 演算法的效率。

表 7 列出 (flight eight, 8 字飛行) 的幾個最高翻譯機率值的指派方式。表 7 的數值顯示第二輪的統計估計值已經相當的收斂，最佳的對應 (0, 34, 12) 的機率值 0.0000788100，遠超過次佳的對應 (0, 3, 124) 的機率值 0.0000000051。

## 5. 實驗結果與評估

我們進行的實驗，證明了新的統計式片語翻譯模型確實可行，能產生相當正確的對應分析。新模型中導入的指派機率的參數不會過度的膨脹，因此 10 萬筆的資料就可以估計出相當可靠的各項機率值。由於新的模型，避免了許多機率值的乘積，EM 演算法的花費的時間較少，機率函數的收斂速度也比較快。

表 8. 第二輪運算之後對應結果由錯誤轉為正確的例子

S	T	第一輪結果			第二輪結果		
		T(A <sub>0</sub> )	T(A <sub>1</sub> )	T(A <sub>2</sub> )	T(A <sub>0</sub> )	T(A <sub>1</sub> )	T(A <sub>2</sub> )
association football	A 式足球			A 式足球		A 式	足球
delay flip-flop	D 型正反器			D 型正反器		D 型	正反器
I demodulator	I 信號解調器			I 信號解調器		I 信號	解調器
Disgraceful act	不友好行動			不友好行動		不友好	行動
disregard to	不拘於		不拘於			不拘	於
secret ballot	不記名投票			不記名投票		不記名	投票
bearer stock	不記名股票		不記名票	股		不記名	股票
false retrieval	不實檢索			不實檢索		不實	檢索
used car	中古車		中古車			中古	車
infix operation	中序運算		中序運算			中序	運算



### 5.1 實驗結果分析

由實驗的結果觀察，以指派機率替代孳生機率和位置扭曲機率，確實可以掌握每一字的翻譯目標位置和其他字的翻譯位置有高度的相關性，而得到比較正確的對應分析。所以在對應的問題比較困難的幾個情況，仍有可能做出正確的分析：

1. 比較偏離常態的罕用翻譯，如 association 通常翻譯成「協會」、「學會」。而 association football 中卻翻譯成類似音譯的「A 式」。雖然 association 不常翻譯成「A 式」，但是指派機率可以比其他模型，給予第一和第二字的「A 式」較高的機率。
2. 翻譯非常分散，沒有定譯的虛詞或輕動詞 (light verb)，如 make、take、to 等，在指派機率的模型下也都可以得到較高的機率。
3. 和原文不一致的翻譯順序，如 (flight eight, 8 字飛行)，在指派機率的模型下，可以得到適當的機率值。

我們檢視對應分析的結果，特別觀察這幾種困難的情況，比較其第一輪和第二輪分析的結果。我們發現確實這些情況到了第二輪使用了新模型後，大部分很明顯的已經扭轉到正確的分析，請參見表 8。

為了評估實驗的效能，我們使用 Och 等人(2000) 的評估方法。將要評估的雙語條目先由人工標示對應，以作為參考答案。例如其中一個雙語條目是 (butter cream biscuit, 奶油夾心餅乾)，人工標示的參考答案如圖 1 所示。標示分為 2 種: S (sure) 和 P (possible)，S 表示確定的對應，P 表示可能的對應，且  $S \subseteq P$ 。

butter	<i>S/P</i>	<i>S/P</i>	.	.	.	.
cream	.	.	<i>P</i>	<i>P</i>	.	.
biscuit	.	.	.	.	<i>S/P</i>	<i>S/P</i>
			奶	油	夾	心
			餅	乾		

圖 1 人工標示參考答案例子

而我們實驗所做的對應的指派方式為 A，再以 (butter cream biscuit, 奶油夾心餅乾) 為例，實驗的對應結果如圖 2 所示。

butter	<i>A</i>	<i>A</i>	.	.	.	.
cream	.	.	<i>A</i>	<i>A</i>	.	.
biscuit	.	.	.	.	<i>A</i>	<i>A</i>
			奶	油	夾	心
			餅	乾		

圖 2 實驗對應結果的例子

以上例子，我們可以得到  $|S|=4$ ， $|P|=6$  (包括標示為 S 的部分)， $|A|=6$ ， $|A \cap S|=4$ ， $|A \cap P|=6$ 。根據 Och 等人(2000)所提出的公式，我們可得到召回率(recall)、準確率(precision)與錯誤率(error rate)如下：

$$recall = \frac{|A \cap S|}{|S|} = \frac{4}{4} = 1$$

$$precision = \frac{|A \cap P|}{|A|} = \frac{6}{6} = 1$$

$$AER(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} = 1 - \frac{4 + 6}{6 + 4} = 0$$

我們從實驗第二輪結果中，隨機抽取 500 個樣本 (包含 2 個英文字及 3 個英文字的樣本各 250 個)，由人工對這些樣本做對應的標示，以作為參考答案。將實驗的結果與人工標示的參考答案比較，我們可以得到以下的召回率(recall)、準確率(precision)與錯誤率(error rate)：

$$recall = \frac{|A \cap S|}{|S|} = \frac{1116}{1248} = 0.894$$

$$precision = \frac{|A \cap P|}{|A|} = \frac{1308}{1517} = 0.862$$

$$AER(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} = 1 - \frac{1116 + 1308}{1517 + 1248} = 0.123$$

在第二輪實驗之後，我們以 EM 演算法繼續第三輪至第五輪的實驗，得到第一至五輪的召回率、準確率與錯誤率如表 9 所示。

表 9. 使用新模型下訓練過程召回率、準確率、錯誤率的收斂情形

	第一輪	第二輪	第三輪	第四輪	第五輪
召回率	0.853	0.894	0.885	0.874	0.873
準確率	0.796	0.862	0.859	0.851	0.849
錯誤率	0.178	0.123	0.129	0.139	0.140

為了了解新的指派機率是否比較有效率，我們也以同樣的語料、同樣的演算法，但把新模型中的指派機率換成 IBM 模型中的孳生機率及位置扭曲機率，進行 IBM Model 的對照組實驗。整體而言，導入新的指派機率，取代孳生與扭曲機率，在執行速度上有很大的改進，在正確率上也略勝一籌。雖然 IBM model 3 對於詞彙孳生部分，考慮了個別詞彙的因素，實驗結果卻顯示，IBM model 3 的錯誤率較高。

表 10. 使用 IBM Model 3 訓練過程的召回率、準確率、錯誤率的收斂情形

	第一輪	第二輪	第三輪
召回率	0.853	0.883	0.867
準確率	0.796	0.848	0.838
錯誤率	0.178	0.136	0.149

## 5.2 其他平滑化方法

由訓練語料得到各項翻譯的機率後，我們可以用這些機率，在語料庫中繼續對應擷取片語，或是進行跨語言檢索中的查詢片語的翻譯。此時，我們可能會因為資料不足，而遇到訓練資料以外的情況，例如在估計語料庫中的某一片語的詞彙對應：

(flight attendant 空服員)

如果我們的訓練語料，沒有 (flight, 空) 的詞彙配對，就無法正確的分析 (flight attendant 空服員) 的對應。當然此時我們可以應用 flight 對 \$any\$ 的機率。但是 \$any\$ 的機率是平均的分配，無法做個別的狀況的考慮。有少數訓練外的翻譯的狀況，是和中文縮寫有關。另外有些中文字容易孳生很多的同義或近義，也會造成很多可能但在訓練外的狀況。這些訓練外的情況，和訓練資料有部分重疊，其機率並不低，我們應該透過比較複雜的機率估計平滑化的做法，給予適當的估計值。

中文的使用有縮寫的現象，如 (flight, 航空) 的部分翻譯 (flight, 航) 與 (flight, 空) 在訓練外出現的可能性不低，而非部份翻譯的 (flight, 員) 則趨近於 0。同樣的，在 (attendant, 服務員) 配對例子中，部分翻譯 (attendant, 服員) \ (attendant, 服) \ (attendant, 服務) 的可能性也顯然高於 (attendant, \$any\$) 的平均值。另外翻譯有部份重疊的情況，也應給予較高的平滑機率。例如訓練內的詞彙配對有 (preservation, 保留)，而 (preservation, 保持) 與 (preservation, 保護) 雖然沒有出現在訓練語料中，其可能性仍然高於其他完全沒有重疊的翻譯配對。

如果沒有這樣的考慮，對於 (flight attendant 空服員) 的對應，詞彙翻譯機率就會全然都使用  $Pr(\$any\$|flight)$  和  $Pr(\$any\$|flight)$  的機率值，無法區隔可能與不可能的配對。如此將流於由指派機率  $Pr(A|2,3)$  來決定一切。在這種狀況下，由於兩字到三字片語的最高可能對應為 (0, 12, 3)，我們很可能得到以下的不完全正確的對應分析：

(flight, 空服), (attendant, 員)

若能考慮翻譯部分符合的條件，給予 (flight, 空) 與 (attendant, 服員) 較高的平滑機率估計值，則比較容易得到正確的對應分析，如

(flight, 空), (attendant, 服員)

目前我們正實驗以英文到中文單字以及英文到中文雙字的兩組詞彙翻譯機率，來合成機

率估計值。實驗的目標在於讓部分字相符的對應，可以透過英文字對中文單字、中文雙字詞彙翻譯機率的線性組合模型，得比較合理的估計值。

## 6. 結論與未來的研究方向

雖然統計式機器翻譯的研究，已經有十多年的歷史，在本研究中我們發現仍然有很大的改進空間，特別是在片語的對應與翻譯方面。我們提出新的統計式片語模型來進行片語的對應，並可應用於查詢關鍵詞的翻譯，以提高跨語言檢索的效率。我們在實驗中，初步驗證新的模型，確實可以在計算效率與對應效果上，有所改進。

我們認為未來統計式雙語對應與機器翻譯，還有很大的改進空間與應用可以發揮。幾個可能的研究方向包括：

1. 目前指派機率雖然在以辭典中的片語，加以訓練。其中的指派機率函數，假設和詞彙本身無關，只考慮片語的總字數，與詞彙位置。雖然如此，實驗證明還是能夠充分的反映詞彙的翻譯字數與位置的安排。然而不同詞性與語法結構的片語的翻譯的指派方式，有很大的差異。目前的做法，只考慮字數，未能考慮片語與詞彙的詞性。我們預計導入片語的句法的訊息，對於不同的名詞、動詞、形容詞片語，訓練不同的指派機率模型，可提供更精確的對應與翻譯的效果。在詞彙翻譯機率方面，以辭典中的片語的用字與翻譯，加以訓練，並不能反映詞彙正常的使用與翻譯的情形。我們預計以大型的語料庫，如光華雜誌漢英語料庫，來訓練詞彙翻譯機率，將可以得到精確的詞彙翻譯機率。
2. 應用片語翻譯模型於雙語語料庫的片語對應。學者大多認為統計式機器翻譯最有應用潛力的地方，在於雙語詞典的編輯與機率翻譯詞典的發展。詞彙對應的發現，不限於單字詞，而應及於多字的片語 [Kupiec 1993]。我們提出的新的統計式片語對應與翻譯的模型，可以在平行語料庫中擷取雙語的片語，提供建立語料庫相關的翻譯詞典，作為翻譯與術語管理的基本工具。利用新發展出來的模型，我們預計提出一套逐句進行的片語對應做法。以新的統計式片語翻譯模型為中心，我們將擷取光華雜誌英中平行語料庫，進行互譯片語的擷取實驗。預計可以提出之統計式片語對應與翻譯模型，可獲得一般雙語辭典無法找到的片語與翻譯，如(graduate institute, 研究所), (cross-strait affair, 兩岸事務), (exclusive interview, 專訪)等。
3. 應用片語翻譯模型於跨語言檢索中的查詢詞翻譯。目前的跨語言檢索的研究，顯示通常會有一半以上的查詢片語，無法在雙語詞典中查到適當的翻譯。對於詞典沒有收錄的片語必須逐字翻譯，通常一字多譯，而逐字的翻譯只有部分和查詢主題相關。統計式的片語翻譯模型，可以由片語中透過對應分解出來的詞彙翻譯機率，比較傾向於文脈中的翻譯 (translation in context)，可以比雙語辭典提供更有效的翻譯。如 graduate 在 graduate student 的文脈下，可以對應到「研究」的翻譯，而雙語詞典的 graduate 詞條普遍的只有列出「畢業」的翻譯。由片語翻譯模型所提供的單字翻譯也比較豐富，如 nuclear 可以有「原子」、「核子」、「核」、「原」等等翻譯。透過

雜訊通道模型，結合翻譯機率函數與文件庫所訓練出來的 N-gram 語言模型，可以產生的和文件庫相關的翻譯，提升跨語言檢索的效果。

## 致謝

本文之研究受到國科會編號 NSC 89-2420-H-007-001 計畫之補助。

## 參考文獻

- BDC 1992 “The BDC Chinese-English electronic dictionary” (version 2.0), *Behavior Design Corporation*, Taiwan.
- Brown, P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Mercer R. L., and Roosin P. S. 1988 “A Statistical Approach to Language Translation”, *In Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, Hungary, pp. 71-76.
- Brown, P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., and Roosin P. S. 1990 “A Statistical Approach to Machine Translation”, *Computational Linguistics*, 16/2, pp. 79-85.
- Brown, P. F., Della Pietra S. A., Della Pietra V. J., and Mercer R. L. 1993 “The Mathematics of Statistical Machine Translation: Parameter Estimation”, *Computational Linguistics*, 19/2, pp. 263-311.
- Chang, J. S. *et al.* 2001. “Nathu IR System at NTCIR-2”. *In Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization*, pp. (5) 49-52, National Institute of Informatics, Japan.
- Chang, J. S., Ker S. J., and Chen M. H. 1998 “Taxonomy and Lexical Semantics – From the Perspective of Machine Readable Dictionary”, *In Proceedings of the third Conference of the Association for Machine Translation in the Americas (AMTA)*, pp. 199-212.
- Chen, H.H., G.W. Bian and W.C. Lin. 1999. “Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval”. *In Proceedings of the 37<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*, pp 215-222.
- Dagan, I., Church K. W. and Gale W. A. 1993 “Robust Bilingual Word Alignment or Machine Aided Translation”, *In Proceedings of the Workshop on Very Large Corpora Academic and Industrial Perspectives*, pp. 1-8.
- Fung, P. and McKeown K. 1994 “Aligning Noisy Parallel Corpora across Language Groups: Word Pair Feature Matching by Dynamic Time Warping”, *In Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA)*, pp. 81-88, Columbia, Maryland, USA.
- Gale, W. A. and Church K. W. 1991 “Identifying Word Correspondences in Parallel Texts”, *In Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, pp. 152-157.
- Gey, F C and A. Chen. 1997. “Phrase Discovery for English and Cross-Language Retrieval at TREC-6”. *In Proceedings of the 6<sup>th</sup> Text Retrieval Evaluation Conference*, pp 637-648.

- Ide, N. and J Veronis. 1998. "Special Issue on Word Sense Disambiguation", editors, *Computational Linguistics*, 24/1.
- Isabelle, P. 1987 "Machine Translation at the TAUM Group", In M. King, editor, *Machine Translation Today: The State of the Art, Proceedings of the Third Lugano Tutorial*, pp. 247-277.
- Kando, Noriko, Kenro Aihara, Koji Eguchi and Hiroyuki Kato. 2001. *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization*, National Institute of Informatics, Japan.
- Kay, M. and Röscheisen M. 1988 "Text-Translation Alignment", *Technical Report P90-00143*, Xerox Palo Alto Research Center.
- Ker, S. J. and Chang J. S. 1997 "A Class-base Approach to Word Alignment", *Computational Linguistics*, 23/2, pp. 313-343.
- Knight, K. and J Graehl. 1997. "Machine Transliteration", In *Proceedings of the 35<sup>th</sup> Annual Meeting of the Association for Computational Linguistics and 8<sup>th</sup> Conference of ACL European Chapter*, pp. 128-135.
- Kupiec, Julian. 1993 "An Algorithm for finding noun phrase correspondence in bilingual corpus", In *ACL 31*, 23/2, pp. 17-22.
- Kwok, K L. 2001. NTCIR-2 Chinese, "Cross-Language Retrieval Experiments Using PIRCS". In *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization*, pp. (5) 14-20, National Institute of Informatics, Japan.
- Longman Group 1992 *Longman English-Chinese Dictionary of Contemporary English*, Published by Longman Group (Far East) Ltd., Hong Kong.
- McCarley, J. Scott. 1999. "Should we Translate the Documents or the Queries in Cross-Language Information Retrieval?" In *Proceedings of the 37<sup>th</sup> Annual Meeting of the Association for Computation Linguistics*, pp 208-214.
- Melamed, I. D. 1996 "Automatic Construction of Clean Broad-Coverage Translation Lexicons", In *Proceedings of the second Conference of the Association for Machine Translation in the Americas (AMTA)*, pp. 125-134.
- Nagao, M. 1986 *Machine Translation: How Far Can it Go?* Oxford University Press, Oxford.
- Oard, D W and J. Wang. 1999. "Effect of Term Segmentation on Chinese/English Cross-Language Information Retrieval". In *Proceedings of the Symposium on String and Processing and Information Retrieval*. <http://www.glue.umd.edu/~oard/research.html>
- Och, Franz Josef and Hermann Ney. 2000. "Improved Statistical Alignment Models". In *Proceedings of the 38<sup>th</sup> Annual Meeting of the Association for Computation Linguistics*.
- Pirkola, A. 1998. "The Effect of Query Structure and Dictionary Setups in Dictionary-based Cross-Language Retrieval". In *Proceedings of the 21<sup>st</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 55-63.
- Smadja, F., McKeown K., and Hatzivassiloglou V. 1996 "Translating Collocations for Bilingual Lexicons: A Statistical Approach", *Computational Linguistics*, 22/1, pp. 1-38.

- Utsuro, T., Ikeda H., Yamane M., Matsumoto M., and Nagao M. 1994 “Bilingual Text Matching Using Bilingual Dictionary and Statistics”, *In Proceedings of the 15th International Conference on Computational Linguistics*, pp. 1076-1082.
- Wu, D. and Xia X. 1994 “Learning an English-Chinese Lexicon from a Parallel Corpus”, *In Proceedings of the first Conference of the Association for Machine Translation in the Americas (AMTA)*, pp. 206-213.
- Yamada, K. and K. Knight. 2001. “A Syntax-Based Statistical Translation Model”. *In Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

