行政院國家科學委員會補助專題研究計畫成果報告

依據概念的中英雙語對應句檢索系統

計畫類別：個別型計畫
計畫編號：NSC 90－2411－H－002－043
執行期間：90 年 8 月 1 日至 91 年 9 月 30 日

計畫主持人：高照明
zmgao@ccms.ntu.edu.tw

本成果報告包括以下應繳交之附件：
- □赴國外出差或研習心得報告一份
- □赴大陸地區出差或研習心得報告一份
- ■出席國際學術會議心得報告及發表之論文各一份
- □國際合作研究計畫國外研究報告書一份

執行單位：台灣大學外國語文學系

中　華　民　國　92　年　1　月 15　日

行政院國家科學委員會專題研究計畫成果報告

依據概念的中英雙語對應句檢索系統
Towards a Concept-Based Chinese-English Translation Retrieval System

計畫編號：NSC 90－2411－H－002－043
執行期限：90 年 8 月 1 日至 91 年 9 月 30 日

主持人：高照明 執行單位:台灣大學外國語文學系
計畫參與人員：台灣大學資訊工程研究所研究生 林桂光 梁菁秀
資訊工程系 莊凱祥 資訊管理系 施嘉峻

中文摘要 本計劃利用同義詞及中英平行語料庫來建構一個翻譯檢索系統。此系統將輸入的中文句子先經過分詞，之後計算每個詞在各文件出現的平均次數 TF (term frequency) 及所有文件數中出現該詞的文件數的反比 IDF (inverse document frequency)，除利用 TF * IDF 計算每個詞的權值，並利用同義詞作為 query expansion 從平行語料中擷取與輸入中文的句子概念相近的中文句子及其對應的英文翻譯並依照權植高低依序排列。

關鍵詞： 概念檢索, 翻譯檢索, 雙語對應句檢索, 中英平行語料庫, 電腦輔助翻譯, 電腦輔助語言教學中英檢索

Abstract：We develop a concept-based Chinese-English translation retrieval system which can take a Chinese sentence as input and extract synonymous Chinese sentences and their English translations from parallel Chinese-English corpora. Drawing on a Chinese-English parallel corpus, the system performs word segmentation for Chinese input sentence and employs TF * IDf (term frequency * inverse document frequency) for calculating the weighting of each word. Using term weighting and a Chinese thesaurus for query expansion, the system is capable of retrieving and ranking synonymous or conceptually similar Chinese sentences along with their English translations.

Keywords：conceptual retrieval, translation retrieval, bilingual concordancer, parallel Chinese-English corpora, computer-assisted translation, computer-assisted language learning,

Introduction

Bilingual concordancers are powerful tools for language learning (cf. Barlow (2000), .Nerbonne (2000), John (2001), Wang (2001)). Combined with sentence-aligned bilingual corpora such as the English-French Parliamentary Corpus Hansard, they allow language learners to retrieve sentences containing an input keyword along with their translations. Moreover, they can accept queries in two languages. This greatly facilitates the retrieval of unfamiliar expressions for language learners in reading and writing. When encountering unfamiliar words or phrases in a foreign language, learners can understand the meanings by inputting these words and looking at the translation examples containing the input words. Similarly, they can learn how to express themselves in a foreign language by inputting an expression in his/her native language and inspecting the translation examples. The potential help of bilingual concordancers to language learners is enormous.

Unfortunately, some of the limitations prevent learners from making the best use of them. Due to the scarcity of sentence-aligned parallel corpus, keyword-based pattern match technique that works with short query is not suited to a longer query such as a clause. A typical dilemma language learners face in using a bilingual concordancer is that the longer the expression they input, the less likely it is to find its translation in the bilingual corpus.

In this project, we propose to resolve this problem by using concept-based rather than keyword-based retrieval technique. Our proposed method can take a keyword, phrase, or sentence in the source or target language as input and retrieve the closest translations if no translation equivalents of the input expression are found. Central to this technique is the formulation of a measure for semantic similarity based on thesauruses and the calculation of term weighting of the input query. The proposed method, largely inspired by recent researches in computational linguistics, is implemented as a web-based program employing a Chinese-English parallel corpus and used by students of freshman English for evaluation. Our study shows that such a tool significantly outperforms conventional keyword-based bilingual concordcers. The limitations of this approach and direction of future research is also discussed.

Sentence Alignment

The origin of a bilingual concordancer may in part be attributed to a technical report by Martin Kay at Xerox in 1980, which was not published until 1997 (Kay (1997). In that seminal paper, Kay proposed a computer-aided translation system which can access previous translation examples created by professional translators. Similar idea was reintroduced by Harris (1988), who suggested using “bitexts” (i.e. translation equivalents coupled together unit by unit) as a computer-aided tool for professional translators. It was not until the mid 1990s that Kay's idea was finally realized (cf. Dagan and Church (1997)) Unfortunately, this kind of powerful tool is not widely used in Taiwan. Part of the reason is that it presupposes the correct identification of bilingual correspondences of the source and target language at the sentence level. In computational linguistics, this procedure is called sentence alignment and has been actively researched in recent years. Approaches to sentence alignment can be categorized into statistics-based (Brown et al. (1991), Church and Gale (1991, 1993), Kay and Roscheisen (1993), dictionary-based, and hybrid method (Kumano and Hirakawa (1994), Haruno and Yamazaki (1996)), none of which performs well in Chinese-English bilingual corpora according the experiments reported in Gao (1997).

A simple method to circumvent the problem of sentence alignment is to look for bilingual texts where there are clearly identifiable markers for text alignment. The internal structure of Bible makes it a good resource for aligning texts at paragraph level, because each paragraph (more precisely passage) in Bible is marked by a numerical indicator. The difficult task of text alignment is thus simplified to matching numerical indicators in the English Bible and the Chinese Bible.

Another important Chinese-English bilingual resource is Sinorama Magazine. Articles in the Sinorama Magazine are written in Chinese and translated into English, Spanish, and Japanese. The English translations are all checked by native speakers of English. The Chinese-English bilingual resources provide important resources for Chinese learners of English who wish to learn English expressions related to Chinese culture or events that takes place in Taiwan.

Sentence alignment in the Sinorama bilingual magazine is complicated by several factors. As with most translations, additions and omissions of sentences or even paragraphs are quite common in the Sinorama magazine. Another difficulty is caused by the abundance of nonliteral translations in the Sinorama magazine. These suggest that a hybrid approach combining statistical and dictionary information is a better choice. We first use Church and Gale's algorithm to derive the initial sentence correspondences in the Chinese-English bilingual corpus. The method is based on the assumption that a longer sentence in the source text tends to be translated in into a longer sentence in the target text. Gale and Church calculate the probability of various sentence alignment such as a 1:1 and 1:2 based on a small manually sentence-aligned corpus and then employ dynamic programming to derive the most likely sentence correspondences. As reported in McEnery and Oakes (1996) and Gao (1997), this algorithm is subject to genre and length. Although the accuracy of this method is not high, it can derive initial sentence correspondences, which can be subsequently refined by a dictionary-based approach. This hybrid approach has the merit of limiting the dictionary-based matching to the neighboring sentences proposed by the length-based statistical approach, thus reducing the time and search space of the matching process.

Algorithms of Developing a Concept-based Translation Retrieval System

A bilingual concordancer can accept a query word in the source or target language and present example sentences containing the input word juxtaposed with their translations. As expressions might take the form of a phrase or a sentence, it is more desirable to design a bilingual concordancer which can take a sentence as input. But given the small size of Chinese-English bilingual corpora available, the chance of finding examples identical to the input sentence is slight. To make the tool useful, it should have the capability of extracting and ranking synonymous examples in terms of their relevance to the input query. With this powerful tool, translators and language learners can use the intelligent bilingual concordancer to find the closest match to the input sentence and its translation.

The central question of building such a program is how to compute the similarity between a sentence and an input sentence. The following procedure is proposed for a concept-based Chinese-English translation retrieval system.

1. Language Identification: A program is written to check if the input is Chinese or English.
2. Tokenization: As there is no delimiter between words in Chinese, a word segmentation program is required to process the Chinese input as well as the Chinese corpus.
3. Lemmatization: Through lemmatization, English words with inflection (whether regular or irregular) will be converted into their lemmas (i.e. basic forms).
4. Indexing: All the words in the Chinese and English texts are indexed so that sentences containing a particular word can be retrieved in no time.
5. Term Weighting: Every Chinese and English word is assigned a weighting based on TF * IDF. In general, an important term in a document occurs several times. We can thus use the average frequency of a term in all document as an indicator of its importance. Because TF assigns high value to function words, we need IDF to offset the undesirable high value of function words induced by IDF. Inverse Document Frequency (IDF) assumes that the importance of a word is in inverse proportion to the number of documents in which it occurs. In other words, the more documents in which a word occurs, the lower IDF value it has. IDF captures the generalization that words of low indexing value such as function words occur in most documents.
6. Query Expansion Using a Thesaurus: We build a Chinese thesaurus based on an online dictionary of the Ministry of Education (http://140.111.1.22/mandr/clc/dict/). The synonyms of the input Chinese words are used for query expansion. The weighting of these expanded query terms are arbitrarily given slightly lower value than that of the original input words on

the ground that synonyms are seldom replaceable in all contexts.

7. Scoring Function: All sentences in the bilingual texts which contain a matched word with the input sentence are retrieved. The similarity of the input sentence and the retrieved sentences is computed based on the sum of the value of TF * IDF of the matched terms. The sentence with the highest score is considered most similar to the input query.

The algorithm above proposes a simple measure to calculate semantic similarity, which is used in building an intelligent web-based Chinese and English bilingual concordancer. To test its performance, we create a bilingual database by extracting bilingual sentences from a bilingual dictionary. The interface and the output of the program are shown in Figure 2 – 4. As can be seen from Figure 2 and 3, the program works reasonably well and is a valuable and convenient tool to search for appropriate English expressions.

Limitations of the our Approach

The proposed method relies on paraphrases whose component words can be substituted with synonyms. It cannot handle synonymous relation between *不及格* ‘not pass’ and *被當* ‘flunk’, which are phrasal synonyms that cannot be decomposed using synonyms at word level. Collocations also provide difficulty for our approach. For instance, *低落* ‘down’ is not synonymous with 差 ‘not good’. But combined with *心情* ‘mood’, the collocation *心情低落* is synonymous with *心情差*. Our method also falls short in cases where identical component words result in different meaning because of different word order or syntax. Cases like these suggest that a much more sophisticated model is needed.

References

Dagan, I., Church, W, and Gale, W. (1993) “Robust Bilingual Word Alignment for Machine Aided Translation.” In Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, pp. 1-8, Ohio.

Dagan, I. (1996) “Bilingual Word Alignment and Lexicon Construction.” Tutorial paper given at the International Conference on Computational Linguistics, Copenhagen.

Fung, P. and Church, K. (1994) “K-vec: A New Approach for Aligning Parallel Texts.” Proceedings of the International Conference of Computational Linguistics, pp.1096-1102, Kyoto.

Fung, P. and McKeown, K. (1994) “Aligning Noisy Parallel Corpora Across Language Groups: Word Pair Feature Matching by Dynamic Time Warping.” Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation, pp. 81-88.

Fung, P. and KcKeown, K. (1997) “A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora Across Language Groups.” Machine Translation, Vol. 12, Nos. 1-2., pp. 53-87.

Gao, Zhao-Ming. (1997) Automatic Extraction of Translation Equivalents from A Parallel Chinese-English Corpora. Ph.D. Thesis. Department of Language Engineering. University of Manchester Institute of Science and Technology.

Hutchins, J. (1998) The Origins of the Translator‘s Workstation. Machine Translation, Vol 13, No. 4, pp. 287-307.

Haruno, M. and Yamazaki, T. (1996) “High-Precision Bilingual Text Alignment Using Statistical and Dictionary Information.” Proceedings of Annual Conference of the Association for Computational Linguistics, pp. 131 -138.

Jones, D. and Somers, H. (1995) “Bilingual Vocabulary Estimation from Noisy Parallel Corpora Using Variable Bag Estimation.” In JADT III GiornateInternazionali di Analsi Statistica dei Dati Testuali, pp. 255-262, Rome.

Kay, M. and Röscheisen, M. (1993) “Text-Translation Alignment.” Computational Linguistics, Vol. 19, No 1, pp 121-142.

Kay, M. (1997) “The Proper Place of Men

and Machines in Language Translation", Machine Translation, Vol. 12, No. (1/2), pp.3-23, 1997

Ker, S.-J. and Chang, J.-S. (1997) "A Class-based Approach to Word Alignment." Computational Linguistics, Vol. 23., No. 2, pp. 313-343.

Kumano, A. and Hirakawa, H. (1994) "Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistic Information." in Proceedings of International Conference on Computational Linguistics, pp. 76-81, Kyoto.

Macklovitch, M. and Hannan, M. (1998) "Line 'Em Up: Advances in Alignment Technology and their Impact on Translation Support Tools", Machine Translation, Vol. 13, No. 1, pp. 41-57.

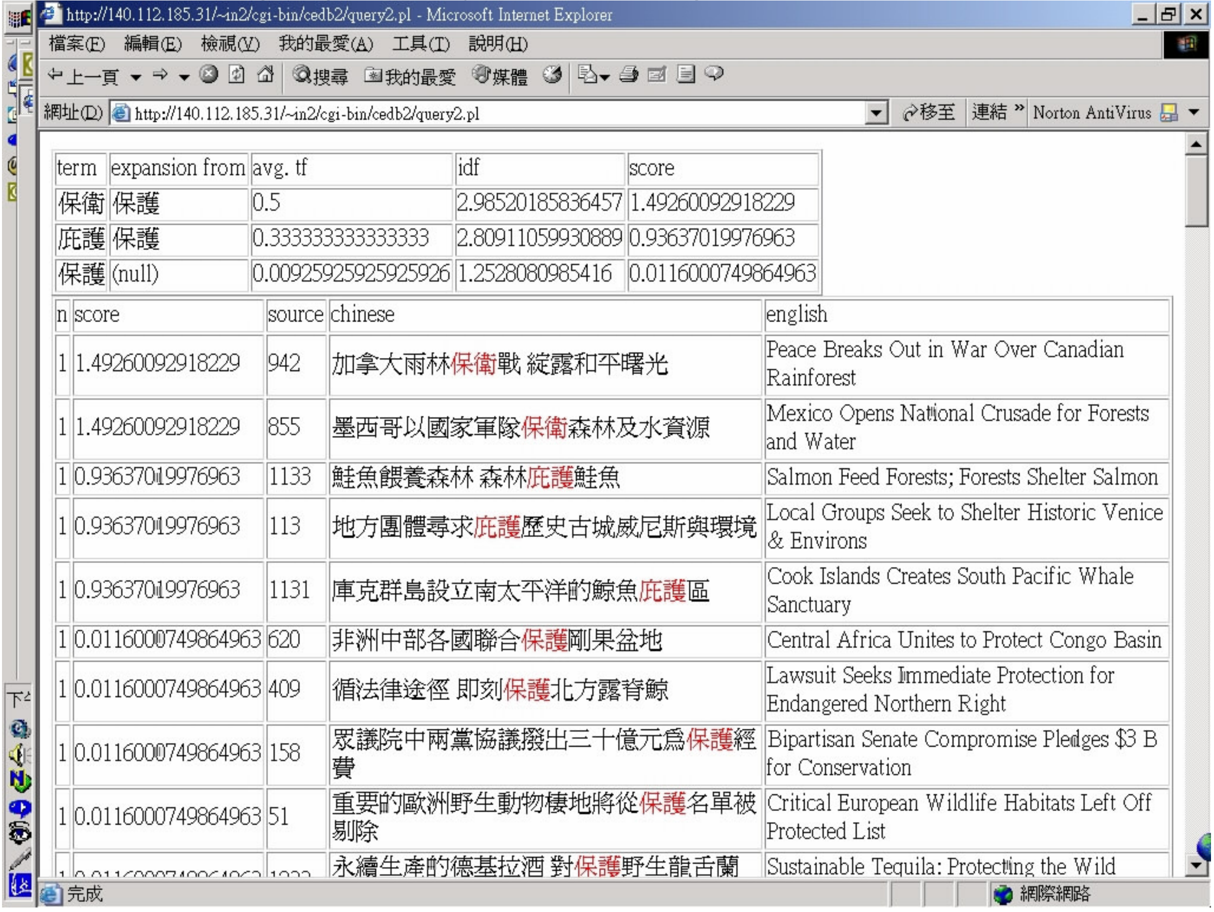
Melamed, D. (1995) "Automatic Construction of Clean Broad-Coverage Translation Lexicons." In Proceedings of 2nd Conference of the Association for Machine Translation in the Americas, Montreal.

Somers, H. and Ward, A. (1996) "Some More Experiments in Bilingual Text Alignments." In Oflazer, K. and Somers, H.(eds) Proceedings of the Second International Conference on New Methods in Language Methods in Language Processing, pp. 66-78, Ankara.

Utsuro, T. et al. (1994) "Bilingual Text Matching Using Bilingual Dictionary and Statistics." in Proceedings of International Conference on Computational Linguistics, pp. 1076-1082, Kyoto.

Wu, D. and Xia, X. (1995) "Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon." Machine Translation, Vol. 9, pp. 285-313.

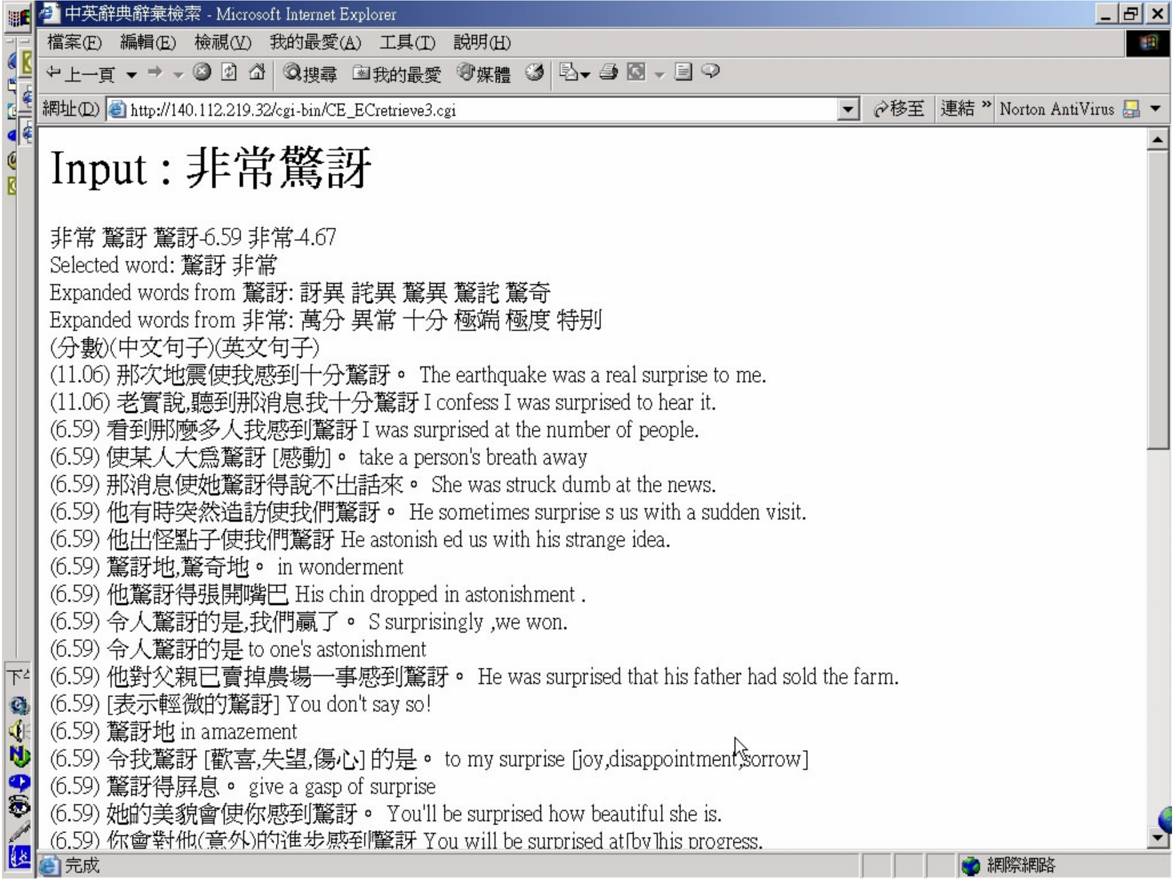
Figure 1. Term Weighting and Query Expansion Using a Chinese Thesaurus


http://140.112.185.31/~in2/cgi-bin/cedb2/query2.pl - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛 媒體
網址(D) http://140.112.185.31/~in2/cgi-bin/cedb2/query2.pl
移至 連結 Norton AntiVirus
完成
網際網路

term	expansion from	avg. tf	idf	score
保衛	保護	0.5	2.98520185836457	1.49260092918229
庇護	保護	0.333333333333333	2.80911059930889	0.93637019976963
保護	(null)	0.00925925925925926	1.2528080985416	0.0116000749864963

n	score	source	chinese	english
1	1.49260092918229	942	加拿大雨林保衛戰 綻露和平曙光	Peace Breaks Out in War Over Canadian Rainforest
1	1.49260092918229	855	墨西哥以國家軍隊保衛森林及水資源	Mexico Opens National Crusade for Forests and Water
1	0.93637019976963	1133	鮭魚餵養森林 森林庇護鮭魚	Salmon Feed Forests; Forests Shelter Salmon
1	0.93637019976963	113	地方團體尋求庇護歷史古城威尼斯與環境	Local Groups Seek to Shelter Historic Venice & Environs
1	0.93637019976963	1131	庫克群島設立南太平洋的鯨魚庇護區	Cook Islands Creates South Pacific Whale Sanctuary
1	0.0116000749864963	620	非洲中部各國聯合保護剛果盆地	Central Africa Unites to Protect Congo Basin
1	0.0116000749864963	409	循法律途徑 即刻保護北方露脊鯨	Lawsuit Seeks Immediate Protection for Endangered Northern Right
1	0.0116000749864963	158	眾議院中兩黨協議撥出三十億元爲保護經費	Bipartisan Senate Compromise Pledges $3 B for Conservation
1	0.0116000749864963	51	重要的歐洲野生動物棲地將從保護名單被剔除	Critical European Wildlife Habitats Left Off Protected List
1	[illegible]	[illegible]	永續生產的德基拉酒 對保護野生龍舌蘭	Sustainable Tequila: Protecting the Wild

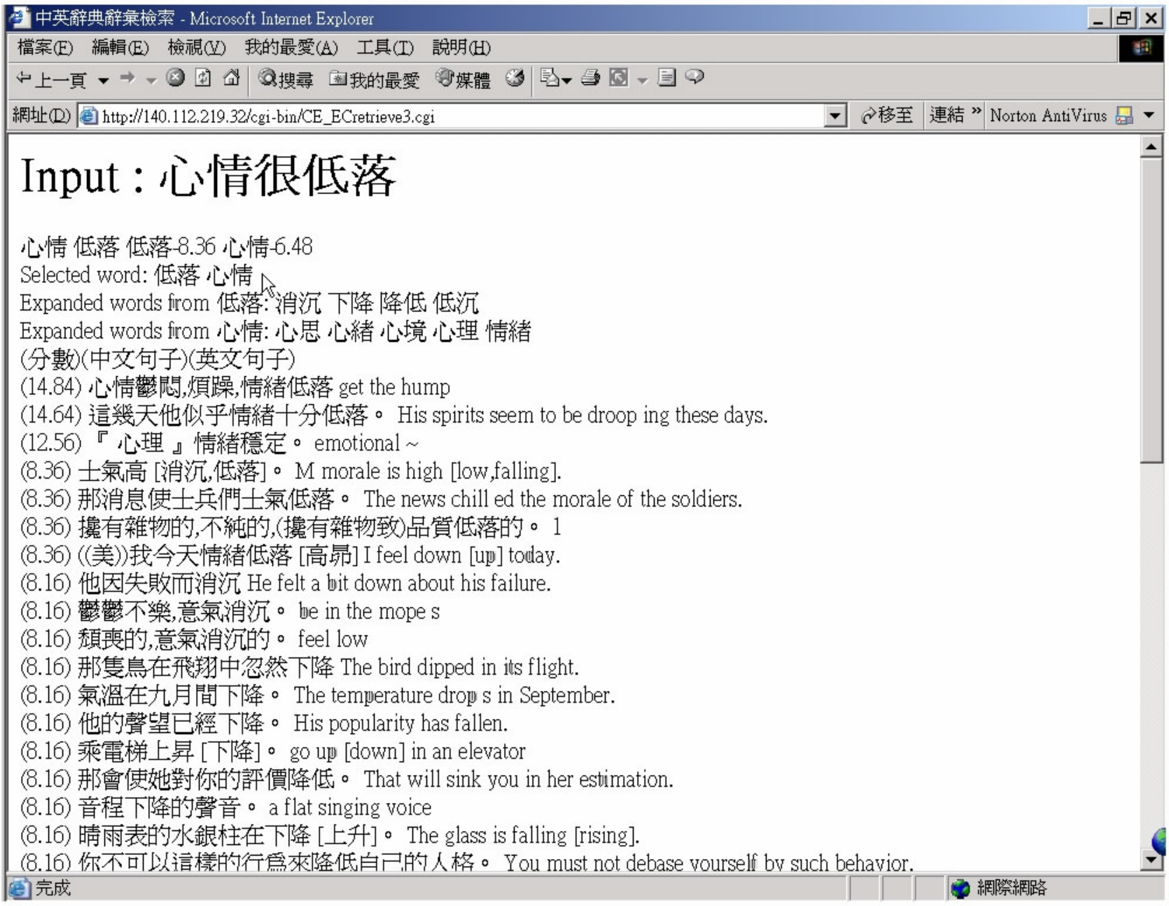
Figure 2.


中英辭典辭彙檢索 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛 媒體
網址(D) http://140.112.219.32/cgi-bin/CE_ECretrieve3.cgi
移至 連結 Norton AntiVirus
完成
網際網路

Input : 非常驚訝

非常 驚訝 驚訝-6.59 非常-4.67
Selected word: 驚訝 非常
Expanded words from 驚訝: 訝異 詫異 驚異 驚詫 驚奇
Expanded words from 非常: 萬分 異常 十分 極端 極度 特別
(分數)(中文句子)(英文句子)
(11.06) 那次地震使我感到十分驚訝。 The earthquake was a real surprise to me.
(11.06) 老實說,聽到那消息我十分驚訝 I confess I was surprised to hear it.
(6.59) 看到那麼多人我感到驚訝 I was surprised at the number of people.
(6.59) 使某人大爲驚訝 [感動]。 take a person's breath away
(6.59) 那消息使她驚訝得說不出話來。 She was struck dumb at the news.
(6.59) 他有時突然造訪使我們驚訝。 He sometimes surprise s us with a sudden visit.
(6.59) 他出怪點子使我們驚訝 He astonish ed us with his strange idea.
(6.59) 驚訝地,驚奇地。 in wonderment
(6.59) 他驚訝得張開嘴巴 His chin dropped in astonishment .
(6.59) 令人驚訝的是,我們贏了。 S surprisingly ,we won.
(6.59) 令人驚訝的是 to one's astonishment
(6.59) 他對父親已賣掉農場一事感到驚訝。 He was surprised that his father had sold the farm.
(6.59) [表示輕微的驚訝] You don't say so!
(6.59) 驚訝地 in amazement
(6.59) 令我驚訝 [歡喜,失望,傷心] 的是。 to my surprise [joy,disappointment,sorrow]
(6.59) 驚訝得屏息。 give a gasp of surprise
(6.59) 她的美貌會使你感到驚訝。 You'll be surprised how beautiful she is.
(6.59) 你會對他(意外)的進步感到驚訝 You will be surprised at[by]his progress.

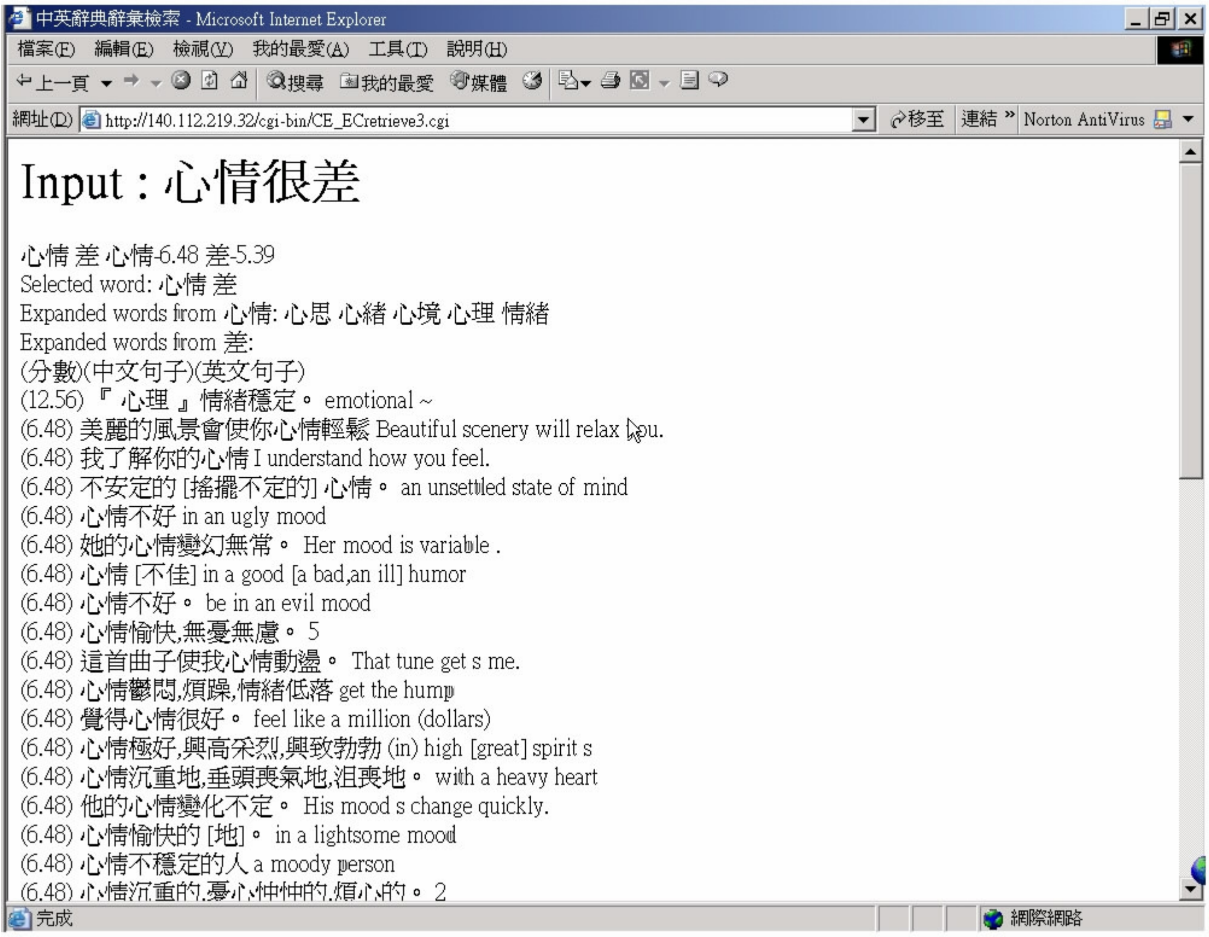
Figure 3


中英辭典辭彙檢索 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛 媒體
網址(D) http://140.112.219.32/cgi-bin/CE_ECretrieve3.cgi
移至 連結 Norton AntiVirus
完成
網際網路

Input : 心情很低落

心情 低落 低落-8.36 心情-6.48
Selected word: 低落 心情
Expanded words from 低落: 消沉 下降 降低 低沉
Expanded words from 心情: 心思 心緒 心境 心理 情緒
(分數)(中文句子)(英文句子)
(14.84) 心情鬱悶,煩躁,情緒低落 get the hump
(14.64) 這幾天他似乎情緒十分低落。 His spirits seem to be droop ing these days.
(12.56) 『 心理 』情緒穩定。 emotional ~
(8.36) 士氣高 [消沉,低落]。 M morale is high [low,falling].
(8.36) 那消息使士兵們士氣低落。 The news chill ed the morale of the soldiers.
(8.36) 攙有雜物的,不純的,(攙有雜物致)品質低落的。 1
(8.36) ((美))我今天情緒低落 [高昂] I feel down [up] today.
(8.16) 他因失敗而消沉 He felt a bit down about his failure.
(8.16) 鬱鬱不樂,意氣消沉。 be in the mope s
(8.16) 頹喪的,意氣消沉的。 feel low
(8.16) 那隻鳥在飛翔中忽然下降 The bird dipped in its flight.
(8.16) 氣溫在九月間下降。 The temperature drop s in September.
(8.16) 他的聲望已經下降。 His popularity has fallen.
(8.16) 乘電梯上昇 [下降]。 go up [down] in an elevator
(8.16) 那會使她對你的評價降低。 That will sink you in her estimation.
(8.16) 音程下降的聲音。 a flat singing voice
(8.16) 晴雨表的水銀柱在下降 [上升]。 The glass is falling [rising].
(8.16) 你不可以這樣的行爲來降低自己的人格。 You must not debase yourself by such behavior.

Figure 4.



中英辭典辭彙檢索 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛 媒體
網址(D) http://140.112.219.32/cgi-bin/CE_ECretrieve3.cgi
移至 連結 Norton AntiVirus
完成
網際網路

Input : 心情很差

心情 差 心情-6.48 差-5.39
Selected word: 心情 差
Expanded words from 心情: 心思 心緒 心境 心理 情緒
Expanded words from 差:
(分數)(中文句子)(英文句子)
(12.56) 『 心理 』情緒穩定。 emotional ~
(6.48) 美麗的風景會使你心情輕鬆 Beautiful scenery will relax you.
(6.48) 我了解你的心情 I understand how you feel.
(6.48) 不安定的 [搖擺不定的] 心情。 an unsettled state of mind
(6.48) 心情不好 in an ugly mood
(6.48) 她的心情變幻無常。 Her mood is variable .
(6.48) 心情 [不佳] in a good [a bad,an ill] humor
(6.48) 心情不好。 be in an evil mood
(6.48) 心情愉快,無憂無慮。 5
(6.48) 這首曲子使我心情動盪。 That tune get s me.
(6.48) 心情鬱悶,煩躁,情緒低落 get the hump
(6.48) 覺得心情很好。 feel like a million (dollars)
(6.48) 心情極好,興高采烈,興致勃勃 (in) high [great] spirit s
(6.48) 心情沉重地,垂頭喪氣地,沮喪地。 with a heavy heart
(6.48) 他的心情變化不定。 His mood s change quickly.
(6.48) 心情愉快的 [地]。 in a lightsome mood
(6.48) 心情不穩定的人 a moody person
(6.48) 心情沉重的,憂心忡忡的,煩心的。 2