利用自然語言處理技術自動產生英文克漏詞試題之研究

王俊弘† 劉昭麟† 高照明‡
†政治大學資訊科學系 ‡台灣大學外國語文學系
{g9124, chaolin}@cs.nccu.edu.tw zmgao@ntu.edu.tw

摘要

電腦輔助產生試題系統的研究熱潮正方興未艾，其研究目的在於節省人力以建置大規模的題庫，並進一步支援網路學習、成效評估與適性化測驗。受惠於來自網際網路上充裕的文字資源，吾人可以利用既有的語料來產生涵蓋各種不同主題的克漏詞試題，以增加題庫的多樣性。另一方面，由於電腦輔助產生試題系統減少人為的干預，也得以保持試題隱密性。我們提出一個詞義辨析的演算法，利用詞典與 selectional preference 所提供的資訊，分析試題的答案的詞義，並以 collocation 為基礎的方法篩選誘答選項。實驗結果顯示我們的系統可在每產生 1.6 道試題中，得到 1 道可用的試題。

Key Words

試題編寫工具與方法論、自動化產生試題、電腦輔助語言學習、詞義辨析、自然語言處理、collocations、selectional preferences

1 緒論

電腦輔助產生試題（computer-assisted item generation，CAIG）可提供題庫（item pools）所需求的各種特質，近年來已廣泛地引起國內外研究學者的重視 [2][6]。利用電腦高速運算的能力，電腦輔助產生試題的系統可產生大量且多樣化的試題，以提供評估學生學習能力的試題來源，也因而減輕了確保試題隱密性的問題 [11]。此外，隨著網路資料量的快速成長，我們可以搜尋並篩選網路上的文字資源作為試題的句子，有效率地產生大量的試題。在這篇論文中，我們即利用自然語言處理（natural language processing）的技術，從網路上的文字資源中有效率地產生克漏詞測驗試題（cloze item）。

自然語言處理的技術提供許多可行且有效的方式以產生英文克漏詞測驗試題。其中一種作法，是以句型樣版為基礎（template-base）的方法建立句子 [3]，或採用較複雜的方式以諸多前置條件來建立句子 [2]。另一套截然不同的演算法則是採用現有的語料庫，如 LDC <http://www.ldc.upenn.edu/> 與 OTA <http://ota.ahds.ac.uk/>，或自行建立的語料庫，從中選取合適的句子以產生試題。前者的方法提供了特定句型且語境容易得到控制的測驗試題，但相對地，一些檢查句子是否合理的複雜機制所需的成本要較後者來得高 [16]。因此，我們可以嘗試利用網路上提供大量的文字檔案，並嚴格過濾其中的文字，挑選出高品質的句子以供我們產生克漏詞測驗試題。測驗編撰者便能夠以相當低的代價從這些產生的試題中挑選合適可用的測驗試題。

一些研究學者已致力於應用自然語言處理的技術來構成語意完整的句子，以便產生多選題（multiple-choice）型式的克漏詞測驗試題。（為了簡單起見，以下用「克漏詞試題」或「試題」代表多選題型式的克漏詞測驗試題。）Johns [7] 與 Steven [17] 使用 concordance 與 collocation 的概念從一般化語料庫中產生試題。Coniam [1] 藉由統計標記化語料庫中詞的詞頻（word frequency）以產生特定類型的試題。在我們 2003 年的工作中，以網路為介面的環境來側寫與評估學生英文能力的情況下，則是利用網路上的英文語料為主要的克漏詞試題的來源 [4][18]。

然而，目前為止僅有少數進階的自然語言處理的技術被套用在產生試題上。例如，許多英文詞通常有多種詞義，而測驗編撰者通常希望在試題中測驗某一詞的特定的使用方法。在這種情況下，盲目地使用關鍵詞比對（keyword matching）的方法－如 concordancer，也許會導致我們得到一連串毫無用處的句子，因而提高人員在後續篩選試題的工作量。此外，要組成一道克漏詞試題不僅僅需要一個語意完整的句子。圖 1 顯示了一道克漏詞試題的範例，挖掉一個詞的句子稱為**題幹**（**stem**），被挖掉的詞即是該試題唯一的**答案**（**key**），而其他三個選項稱為**誘答選項**（**distractor**）。

1. My sister is _____, that is, I am going to be an uncle soon.
(A) supposing (B) assigning
(C) expecting (D) scheduling

圖1　一道英文克漏詞試題的範例

給定一個句子，我們仍需要誘答選項來組成一道試題。誘答選項的選擇影響到試題的**難易度**（**item facility**）與試題的**鑑別度**（**item discrimination**），是一項重要的工作 [12]。因此，誘答選項的選擇也需要更謹慎的策略，若只考慮以詞頻作為挑選誘答選項的依據 [1][4]，顯然不符合實際的需求。為了消弭這類型的缺失，我們使用了詞義辨析（word sense disambiguation）的技術，從語料庫中挑選含有指定詞義的答案的句子，並利用統計 collocation 與 selectional preference [10] 的技術來挑選誘答選項。實驗評估的結果顯示我們的方法能夠建立令人滿意的品質的試題，我們並實際運用系統產生的試題在大一程度的英文課程隨堂測驗之中。

我們在第 2 節概述產生克漏詞試題的流程，並在第 3 節解釋語料庫的準備方法，旁及詞典的介紹與使用。在第 4 節中我們詳述將詞義辨析應用到系統中以選擇句子，並且在第 5 節探究將 collocation 與 selectional preference 的應用套在產生誘答選項的策略上。對於系統的評估與相關的應用將在第 6 節提出。

2　系統架構

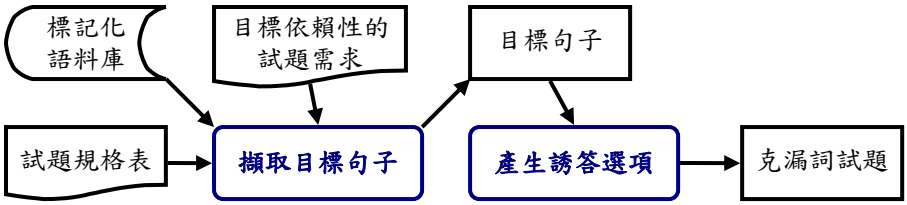

圖2　系統架構圖

圖 2 是自動產生克漏詞試題的系統架構圖。產生克漏詞試題涉及兩個主要的步驟，分別實作於兩個子系統中。**擷取目標句子**（**target sentence retriever**）子系統在測驗編撰者的要求與目標依賴性的試題需求的雙重限制下，從**標記化語料庫**（**tagged corpus**）中擷取克漏詞試題所需的句子。透過**試題規格表**（**item specification**）的介面，測驗編撰者可輸入試題的答案，並指定答案的詞性與詞義。圖 3 顯示試題規格表的介面。我們的系統將嘗試依試題規格表的要求以產生所需的試題量。**目標依賴性的試題需求**（**target-dependent item requirements**）具體地指定所有針對特殊測驗目標所設計的試題，必需遵循的一般化原則。舉例而言，在台灣大學入學考試（College Entrance Examinations）中，一道克漏詞試題所包含的詞數大致介於 7 個到 28 個詞之間 [18]，而測驗編撰者可依循這樣的傳統來建立測驗用的試題。此外，我們的系統允許測驗編撰者也可以在不指定答案的情形下，要求我們的系統產生特定數量且特定詞性的試題。

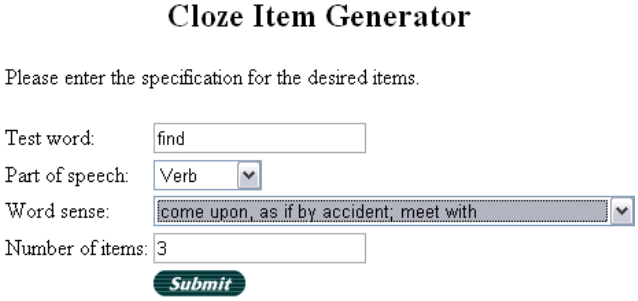

圖3　試題規格表的介面

在取得目標句子（target sentence）後，下一個步驟是由**產生誘答選項**（**distractor generator**）子系統考慮詞頻排名、collocation 與 selectional preference 等參考條件來篩選誘答選項。如果無法找到足夠的誘答選項（一般情形下是 3 個）以滿足答案與題幹的限制，系統會放棄這個目標句子而重新啟始整個產生克漏詞試題的程序。

3 語料來源與詞典

在搜集語料的工作上，我們利用網路爬梳器（web crawler）從 Taiwan Journal <http://taiwanjournal.nat.gov.tw>、Taiwan Review <http://publish.gio.gov.tw/FCR/fcr.html> 與 China Post <http://www.chinapost.com.tw> 定期抓取最新的文章。這些線上期刊與新聞報導提供高品質且更新速度快的文字資源，拼詞錯誤率極為罕見。目前在我們的語料庫中，共有 163,719 個句子，其中包含了 3,077,474 個詞次（word token）與 31,732 個詞型（word type）。

一份 HTML 格式的網頁文件含有各式各樣的多媒體內容。我們需要從混合了標題、主體文字、圖片與影音檔案的網頁內將主體文字的部份擷取出來，而擷取出的文字段落需經由斷句的步驟以切裁成個別的句子，作為試題題幹的來源。我們使用 Reynar 開發的 MXTERMINATOR 工具以完成斷句的工作。MXTERMINATOR 實驗於 Brown 與 Wall Street Journal 等著名的語料庫中有大約 97.5%的斷句正確率 [15]。我們接著對每個句子斷詞，以利於我們接下來對個別的詞加註有用的標記。

為語料庫中的詞標記各種資訊可提高產生克漏詞試題的效率。我們利用 Ratnaparkhi 的 MXPOST 工具為語料庫中的詞標記詞性，MXPOST 遵循 Pen Treebank 詞性集的標準 [13]。在標記詞性後，我們依每個詞的詞性標記其應有的詞根（lemma）。舉例而言，若 *classified* 的詞性是 *VBN*，我們標記其詞根為 *classify*；若詞性是 *JJ*，則標記成 *classified*。另外，我們也使用 Lin 的 MINIPAR [8] 標記句子中的慣用語。MINIPAR 能夠偵測 *arrive at* 與 *in order to* 等不可分的慣用語。這對於偵測可分的慣用語而言是不足夠的，我們也將極積尋求較佳的替代方案。

然而，使用 MINIPAR 最主要的目的在於它提供了局部語法剖析（partial parse）的功能，我們將之應用於產生克漏詞試題的系統中。一個句子中的詞與詞之間若有語法上的關係將會被 MINIPAR 所偵測，利用這些關係可輔助我們施行詞義辨析。為求簡便，對於詞 *w* 而言，句子中其他詞與 *w* 有語法上的關係，稱之為 *w* 的**信號詞**（**signal word**）或簡稱為**信號**（**signal**）。

由於歷屆大學入學考試英文科的克漏詞試題大都是以動詞、名詞、形容詞與副詞為測試標的 [18]，我們目前也著重於產生以這四種詞性作為答案的克漏詞試題，因此只對句子中屬於這些詞性的詞施行詞義辨析。對於動詞而言，其信號詞包括它的主詞、受詞與修飾它的副詞；對於名詞而言，其信號詞包括修飾它的形容詞或是將之視為主詞或受詞的動詞；舉例而言，在 *Jimmy builds a grand building* 一句中，*build* 與 *grand* 都是名詞 *building* 的信號詞；對於形容詞與副詞而言，其信號詞包括它們所修飾的詞與修飾它們的詞。

若產生試題的過程中需要詞典定義詞的資訊時，系統將訴諸於機器可讀取的電子詞典。當我們需要詞的詞義、同義詞與例句等資訊施行詞義辨析時，我們藉由 WordNet <http://www.cogsci.princeton.edu/wn/> 的輔助；當我們需要動詞、名詞、形容詞與副詞的類別（class）以統計 selectional preference 與 collocation 的程度時，我們仰賴 HowNet <http://www.keenage.com\verb> 的定義。

4 擷取目標句子

在圖 2 中，**擷取目標句子**子系統從語料庫中擷取高品質的句子。一個被視為候選目標句子（candidate target sentence）的句子必需包含指定的答案與詞性。藉由先前利用 MXPOST 對詞所標記的詞性，我們可以輕易地達成上述的要求。在有了候選目標句子後，試題產生器（item generator）需要決定答案的詞義是否符合需求。我們施行詞義辨析的演算法是建構在 selectional preference 的觀念上。

4.1 廣義的 Selectional Preference

利用 selectional preference 的輔助以施行詞義辨析的著眼點在於，在一般情形下，句子中某一特定詞的詞義，會受到句子中其他詞的種類所限制。Selectional preference 與詞義辨析之間的密切性可用一個簡單的例子加以說明，當名詞 *chair* 出現在句子 *Susan interrupted the chair* 中時，應是指一個人而並非傢俱 [10][14]。因此我們可以觀察句子中與某個多義詞（polysemous word）有語法關係的信號詞來猜測多義詞在該句子中所扮演的詞義。

我們可以仰賴 HowNet 的定義，將動詞對其主詞及受詞的偏好性，延伸到克漏詞試題的答案（詞性可能是動詞、名詞、形容詞或副詞）對其信號詞的偏好性。在統計 selectional preference 的強度時，以「詞對類別」的方式統計，令 w 與 π 分別代表詞與類別（定義在 HowNet 中），以 $f_v(w,\pi)$ 表示 w 與 π 共同參

與語法關係 v 的頻率，且 π 為 w 的信號詞的類別，並以 $f_v(w)$ 表示 w 參與語法關係 v 的頻率，不計其信號詞的類別。我們將 w 與 π 的 selectional preference 的強度以式子 (1) 表示：

$$A_v(w,\pi) = f_v(w,\pi) / f_v(w) \tag{1}$$

當語料庫中出現 w 參與 v 的情形下，不論 w 的信號詞為何，$f_v(w)$ 皆累加 1 次。令 s 為 w 在 v 關係下的信號詞，並以 $\Pi(s) = \{\pi_1, \pi_2, \ldots, \pi_y\}$ 表示 s 的類別集合，當語料庫中出現 w 與 s 共同參與 v 且 $\pi \in \Pi(s)$ 時，則 $f_v(w,\pi)$ 的計數加上 $1/y$ 。表 1 顯示 3 個英文的動詞 *eat*、*tell* 與 *find* 與其受詞 HUMAN、FOOD 兩個類別的統計資料。由表 1 可知，動詞 *eat* 對其受詞的偏好性，明顯地傾向類別 FOOD，與動詞 *tell* 恰好形成對比。

表1　Selectional preference 的部份統計資料

v = 動詞對受詞		動詞		
		eat	*tell*	*find*
類別	HUMAN	0.047	0.487	0.108
	FOOD	0.441	0.005	0.057

4.2　詞義辨析

我們藉由 4.1 節介紹的廣義的 selectional preference 與詞典 WordNet 的輔助，辨析候選目標句子中答案的詞義。為了避免造成混淆，本小節將以「關鍵詞」代表候選目標句子中欲施行詞義辨析的詞。若是關鍵詞在 WordNet 中僅具有一種詞義，詞義辨析演算法將會指派其唯一的詞義給予關鍵詞。反之，若是關鍵詞擁有多種詞義，以一個候選目標句子 *They say film makers don't spend enough time developing a good story* 為例說明詞義辨析的演算法。我們欲對句子中的動詞 *spend* 作詞義辨析，在 WordNet 的定義下，*spend* 有兩種詞義：

1. (99) spend, pass – (pass (time) in a specific way; “How are you spending your summer vacation?”)
2. (36) spend, expend, drop – (pay out; “I spend all my money in two days.”)

第一個詞義為 *pass (time) in a specific way*，第二個詞義是 *pay out*。WordNet 對每個詞義所包含的資訊包括 (Ⅰ) **標頭詞**（**head words**），由一個或數個詞組成，這些標頭詞共享該詞義；(Ⅱ) 該詞義所專屬的**例句**，展示該詞義的獨特用法。在之後關於詞義辨析的作法的討論中，每當提及關鍵詞的詞義的標頭詞時，我們不考慮關鍵詞為其本身某個詞義的標頭詞。因此，動詞 *spend* 的第一個詞義僅有 1 個標頭詞：*pass*，而第二個詞義有 2 個標頭詞：*extend* 與 *drop*。

一個對候選目標句子中動詞 *spend* 施行詞義辨析的直覺的方法，是以 *spend* 的標頭詞取代其在句子中的地位。正確詞義的標頭詞與其他詞義的標頭詞相較之下，套用在候選目標句子中應有較合理的語意。反之，若是語料庫中極少出現標頭詞取代後的句子，該標頭詞所屬的詞義就不太可能是關鍵詞的詞義。這項直覺引領我們計算一個詞義在標頭詞部份所應得的分數，也就是我們所表示的 Ω_t 。

此外，我們可以比較不同詞義的例句與候選目標句子中，*spend* 的上下文（context）的相似程度，在這裡所指的上下文，與 *spend* 的信號詞有密切關係。再次以句子 *They say film makers don't spend enough time developing a good story* 為例，我們可以比較 *spend* 在候選目標句子中主詞（makers）與受詞（time）的類別，是否與 *spend* 在例句中主詞與受詞的類別相同。若 *spend* 的第一個詞義的例句提供一個與 *spend* 在候選目標句子中較近似的上下文，則第一個詞義獲得較高的分數，反之則由第二個詞義獲得較高分。這項直覺引領我們得到詞義在例句部份所應得的分數，也就是我們將在下面介紹的 Ω_s 。

假設關鍵詞 w 在 WordNet 的定義下有 n 個詞義，令 $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ 是關鍵詞 w 的詞義集合。假設關鍵詞 w 的詞義 θ_i 在 WordNet 中有 m_i 個標頭詞。（注意我們並不考慮 w 本身為 θ_i 的標頭詞。）我們使用集合 $\Lambda_i = \{\lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,m_i}\}$ 代表關鍵詞 w 的詞義 θ_i 的標頭詞集合。

當我們使用 MINIPAR 對一關鍵詞所屬的候選目標句子 T 作語法分析時，可以得到關於關鍵詞的信號詞的資訊。假設 w 在句子 T 中有 $\mu(T)$個信號詞。我們令集合 $\Psi(T,w) = \{\psi_{1,T}, \psi_{2,T}, \ldots, \psi_{\mu(T),T}\}$ 代表 T 中 w 的信號詞集合。同時，以 $\nu_{k,T}$ 代表 T 中 w 與 $\psi_{k,T}$ 的語法關係，並且以 $\Gamma(T,w) = \{\nu_{1,T}, \nu_{2,T}, \ldots, \nu_{\mu(T),T}\}$ 代表 w 與 $\psi_{k,T}$ 之間的語法關係的集合。

就關鍵詞 w 的第 i 個詞義 θ_i 而言，其第 j 個標頭詞 $\lambda_{i,j}$ 所獲得的分數，是利用式子 (1) 計算 $\lambda_{i,j}$ 與 T 中每個信號詞 $\psi_{k,T}$ 的 selectional preference 的強度，再求其平均值。

$$\frac{1}{\mu(T)}\sum_{k=1}^{\mu(T)} A_{v_{k,T}}(\lambda_{i,j},\psi_{k,T})$$

因此詞義 θ_i 所獲得的分數，即是 θ_i 所有標頭詞所獲得的分數的平均值。

$$\Omega_t(\theta_i \mid w,T) = \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{\mu(T)}\sum_{k=1}^{\mu(T)} A_{v_{k,T}}(\lambda_{i,j},\psi_{k,T}) \tag{2}$$

我們用式子 (2) 計算候選目標句子 T 中關鍵詞 w 的詞義 θ_i 在標頭詞部份所獲得的分數。注意 selectional preference 的強度 $A_{v_{k,T}}(\lambda_{i,j},\psi_{k,T})$ 與標頭詞分數 Ω_t 兩者的數值皆落於 0 到 1 的範圍之內。

既然 WordNet 對許多詞義有提供例句，我們可以比較詞義的例句與候選目標句子之間的語境相似程度去判別候選目標句子中的關鍵詞的詞義。我們利用關鍵詞的信號詞決定句子的語境。令 T 與 S 分別為 w 所屬的候選目標句子與 w 的詞義 θ_i 的例句。我們將以下列 3 個步驟，計算詞義 θ_i 在例句部份所獲得的分數 Ω_s。如果一個詞義有多個例句，我們將用式子 (3) 對 θ_i 的每個例句計算分數，最後並對個別分數的總合值作平均。

計算 $\Omega_s(\theta_i \mid w,T)$ 的步驟

1. 分別求得關鍵詞 w 在 T 與 S 中的信號詞集合 $\Psi(T,w)$ 與 $\Psi(S,w)$，以及它們與 w 的語法關係集合 $\Gamma(T,w)$ 與 $\Gamma(S,w)$。

$$\Psi(T,w) = \{\psi_{1,T},\psi_{2,T},\ldots,\psi_{\mu(T),T}\}$$
$$\Psi(S,w) = \{\psi_{1,S},\psi_{2,S},\ldots,\psi_{\mu(S),S}\}$$
$$\Gamma(T,w) = \{v_{1,T},v_{2,T},\ldots,v_{\mu(T),T}\}$$
$$\Gamma(S,w) = \{v_{1,S},v_{2,S},\ldots,v_{\mu(S),S}\}$$

2. 我們尋找 T 中 w 的信號詞 $\psi_{j,T}$ 與 S 中 w 的信號詞 $\psi_{k,S}$ 使得 $v_{j,T} = v_{k,S}$，假設 $\psi_{j,T}$ 在 HowNet 中有 $n_{j,T}$ 個類別，以集合 $\Pi(\psi_{j,T}) = \{\pi_{j,T,1},\pi_{j,T,2},...,\pi_{j,T,n_{j,T}}\}$ 表示，而 $\psi_{k,S}$ 在 HowNet 中有 $n_{k,S}$ 個類別，以集合 $\Pi(\psi_{k,S}) = \{\pi_{k,S,1},\pi_{k,S,2},...,\pi_{k,S,n_{k,S}}\}$ 表示。對於 $\Pi(\psi_{j,T})$ 內的每個類別 $\pi_{j,T,l}$，逐一比對 $\Pi(\psi_{k,S})$ 中是否存在類別 $\pi_{k,s,m}$ 使得 $\pi_{j,T,l} = \pi_{k,S,m}$。每比對一組相同的類別，將累計分數 $M(\theta_i,T)$ 加上 $1/n_{j,T}$。

$M(\theta_i,T) = 0;$
mark all $v_{j,T} \in \Gamma(T,w)$ as unmatched;
for($j = 0; j < \mu(T); j++$)
  for($k = 0; k < \mu(S); k++$)
    if (($v_{j,T}$ unmatched) and ($v_{j,T} = v_{k,S}$))
    {
      mark $v_{j,T}$ as matched;
      for($l = 0; l < n_{j,T}; l++$)
        for($m = 0; m < n_{k,S}; m++$)
          if ($\pi_{j,T,l} = \pi_{k,S,m}$) $M(\theta_i,T) = M(\theta_i,T) + 1/n_{j,T}$
    }

3. 式子 (3) 度量關鍵詞在候選目標句子的信號詞與在例句中的信號詞，當與關鍵詞有相同語法關係的情形下，所得到的平均分數。

$$\Omega_s(\theta_i \mid w,T) = \frac{M(\theta_i,T)}{\mu(T)} \tag{3}$$

候選目標句子 T 中的關鍵詞 w 的詞義 θ_i 所獲得的分數，是由式子 (2) 所計算的 $\Omega_t(\theta_i \mid w,T)$（標頭詞所獲得的分數）與式子所 (3) 計算的 $\Omega_s(\theta_i \mid w,T)$（例句所獲得的分數）加總而得，若至少存在一個詞義的分數超過我們在式子 (4) 所選定的門檻值（threshold），候選目標句子 T 中的關鍵詞 w 就會被指派給分數最高的詞義。反之，若 $\Omega_t(\theta_i \mid w,T)$ 與 $\Omega_s(\theta_i \mid w,T)$ 的加總值太小，表示演算法的結果可信度不高，系統將不會作出任何草率的決定，並挑選其他候選目標句子，針對其中的關鍵詞重新起始詞義辨析的運作。我們將在 6.1 節闡明並討論選用不同的門檻值對詞義辨析的正確率所造成的影響。

$$\arg\max_{\theta_i \in \Theta_i} \Omega_t(\theta_i \mid w,T) + \Omega_s(\theta_i \mid w,T) \tag{4}$$

5 產生誘答選項

克漏詞試題中的誘答選項影響了學生幸運猜中答案的可能性。若試題中含有明顯可看出不可能是答案的誘答選項，學生也許能夠在不知道答案的情況下得知正確的答案。因此，我們需要選擇可以訴諸於填補這漏洞的誘答選項，並必需避免同一試題中有多重答案的情形發生。

有一些可想到的方法與可供選擇的方案是容易參考且實作的。答案的反義詞是一項選擇，但一般情形下學生將會忽視之。而誘答選項的詞性必需與答案一致，否則學生將很容套用基本的文法知識僅依詞性去選擇答案，而不需知道整個句子的語意。我們也可以考慮文化背景的影響。華文語系背景的學生較易受到具有相同中文譯詞的英文詞彙所干擾。雖然學習策略在大部份時間是有作用的，學生也許會發現要分辨有相似中文詞義的英文詞彙是困難的。因此，文化背景依賴性（culture-dependent）的策略，可將具有與答案相似中文詞義的英文詞作為誘答選項的考慮，但需準備具公信力的中英文雙語詞典與中文同義詞詞典。

為了有系統地產生誘答選項，我們使用詞的詞頻排名作為初步篩選誘答選項之用[12][18]。假設我們產生一道答案的詞性為 ρ 的試題，且詞性 ρ 在辭典中有 n 個詞次，而答案的詞頻排名在 n 個詞次中排第 m 個。我們從 n 個詞次中，在 $[m-n/10, m+n/10]$ 的詞頻排名範圍內隨機挑選詞作為候選誘答選項。我們限制誘答選項的詞頻排名需與答案相近。接著檢驗這些候選誘答選項與題幹的**合適度**（**fitness**），以從中篩選誘答選項。合適度是由候選誘答選項的類別與題幹中其他詞的類別的 collocation 值而決定，詞的類別同樣定義在 HowNet 中。

在第 3 節曾提及，我們已對語料庫中的詞標記上它們的信號詞。句中的某一詞若擁有較多的信號詞，通常該詞對句子的語意貢獻較多，所以也應當在選擇誘答選項時佔有較重要的地位。既然我們並非真正檢視整個目標句子的語意，一個相對上挑選誘答選項較安全的方法，是選擇很少與重要的詞一起出現在句子中的詞。

令 $T = \{t_1, t_2, \ldots, t_q\}$ 代表題幹的詞的集合（亦即不包括答案在內）。我們從 T 中依下列兩個條件過濾出**重要詞**：(Ⅰ) 詞性為動詞、名詞、形容詞或副詞之一者，且 (Ⅱ) 該詞擁有兩個（含）以上的信號詞或被某一子句（clause）所修飾。令 $T' \subset T$ 為句子 T 中重要詞的集合，$T' = \{t_1', t_2', \ldots, t_{q'}'\}$，我們計算候選誘答選項 κ 與每個重要詞的 pointwise mutual information，並求其平均值。假設 $C = \{S_1, S_2, \ldots, S_{\aleph}\}$ 為語料庫中所有句子的集合，$\Pi(\kappa)$ 與 $\Pi(t_i')$ 分別為候選誘答選項 κ 與重要詞 t_i' 的類別集合，我們定義 $\Pr(\Pi(\kappa))$ 為語料庫中存在 S_η 包含一詞 w 且 $\Pi(w)$ 與 $\Pi(\kappa)$ 的交集不為空集合的比例；同理 $\Pr(\Pi(t_i'))$ 為語料庫中存在 S_η 包含一詞 χ 且 $\Pi(\chi)$ 與 $\Pi(t_i')$ 的交集不為空集合的比例。

$$\Pr(\Pi(\kappa)) = \frac{1}{\aleph}\{\sum_{S_\eta} 1 \mid S_\eta \text{ contains } w \text{ and } \Pi(w) \cap \Pi(\kappa) \neq \phi\}$$

$$\Pr(\Pi(t_i')) = \frac{1}{\aleph}\{\sum_{S_\eta} 1 \mid S_\eta \text{ contains } \chi \text{ and } \Pi(\chi) \cap \Pi(t_i') \neq \phi\}$$

除此之外，定義 $\Pr(\Pi(\kappa),\Pi(t_i'))$ 為語料庫中存在 S_η 同時包含 w 與 χ 且 $\Pi(w)$ 與 $\Pi(\kappa)$ 的交集、$\Pi(\chi)$ 與 $\Pi(t_i')$ 的交集皆不為空集合所佔的比例。

$$\Pr(\Pi(\kappa),\Pi(t_i')) = \frac{1}{\aleph}\{\sum_{S_\eta} 1 \mid S_\eta \text{ contains } w, \chi \text{ where } w \neq \chi \text{ and } \Pi(w) \cap \Pi(\kappa) \neq \phi \text{ and } \Pi(\chi) \cap \Pi(t_i') \neq \phi\}$$

候選誘答選項 κ 與題幹的適合度的定義如式子 (5)。

$$f(\kappa) = \frac{-1}{q'} \sum_{t_i' \in T} \log \frac{\Pr(\Pi(\kappa),\Pi(t_i'))}{\Pr(\Pi(\kappa))\Pr(\Pi(t_i'))} \tag{5}$$

若 $f(\kappa)$ 的值高於 0.3，則可被接受成為誘答選項。為了讓較低的 collocation 得到較高的分數，我們將整個式子加上一個負號。設定門檻值為 0.3 的原因是基於式子 (5) 對 220 筆訓練資料統計後得知，這份訓練資料搜集自 1992 年到 2003 台灣的大學入學考試英文科的克漏詞試題。

6 評估與應用

6.1 詞義辨析

詞義辨析在自然語言處理的研究中是一項被廣泛探索與討論的課題 [10]。不同的演算法在異質的環境下使用相異的評估方法，使得詞義辨析的正確率落於 40%到 90%的大範圍內 [14][19]。主觀地比較不同演算法之間的優劣，若不依賴像 SENSEVAL 一個共同比較的基礎，並非一項簡單的工作，因此在本論文中只回報我們的實驗結果。

表2　詞義辨析正確率

關鍵詞的詞性	基準	門檻值 = 0.4	門檻值 = 0.7
動詞	38.0%(19/50)	57.1%(16/28)	68.4%(13/19)
名詞	34.0%(17/50)	63.3%(19/30)	71.4%(15/21)
形容詞	26.7%(8/30)	55.6%(10/18)	60.0%(6/10)
副詞	36.7%(11/30)	52.4%(11/21)	58.3%(7/12)

實驗材料是從語料庫中選取 160 個句子，針對每個句子選定其中一個關鍵詞作詞義辨析。這些關鍵詞包含了 50 個動詞、50 個名詞、30 個形容詞與 30 個副詞，這 160 個關鍵詞在 WordNet 的定義中，詞義數量介於 2 個（如名詞 *verification* 與形容詞 *frightened*）到 19 個（如動詞 *have*）之間－亦即測試用的關鍵詞皆屬於多義詞，每個關鍵詞平均有 4.85 個詞義。我們將這 160 個關鍵詞交由系統施行詞義辨析，並由人工判斷詞義辨析的正確率，正確率顯示於表 2。

基準（**baseline**）一欄所顯示的正確率，是當我們總是選擇該關鍵詞最常見的詞義，而一個關鍵詞最常見詞義則是仰賴 WordNet 所提供的資訊。其右兩欄顯示我們套用式子(4)，設定不同的門檻值（分別是 0.4 與 0.7）所得到的正確率。如同我們在 4.2 節所提到的，當門檻值提高時，將會留下較多的關鍵詞無法做詞義辨析（因主觀認定詞義辨析的可信度不足以採信），因此門檻值的選擇直接影響了詞義辨析的正確率。不令人意外地，選用較高的門檻值會得到較高的正確率，但同時卻增加了退回率（rejection rate）。所幸語料庫可以不斷擴充以容納更多的句子，我們可專注於提高詞義辨析的正確率，在退回率上稍作犧牲。

我們注意到 WordNet 並非對每個詞的詞義都提供例句，當一個詞義沒有任何例句時，這個詞義將得不到任何例句方面所提供的分數，也就是 Ω_s 得到 0 分。在我們目前相當倚重 WordNet 提供的例句的情形下，所造成必然的結果是：系統不傾向將一個詞分派給沒有例句的詞義。這點在我們目前的設計中是一項明顯的缺失，但事實上，這個問題在對於產生試題的系統並非嚴重且無解的。一個多義詞常用或重要的詞義通常會配有一個以上的例句，所以詞義辨析的問題並不常發生。若我們想要完全避免這個問題，可以客製化 WordNet，使得常見的詞的所有詞義皆有專屬的例句。

6.2 產生克漏詞試題

圖 4 顯示圖 3 給定的條件的輸出。當系統依需求而產生一些試題後，測驗編撰者可以從中挑選最佳的數道試題以供實測。

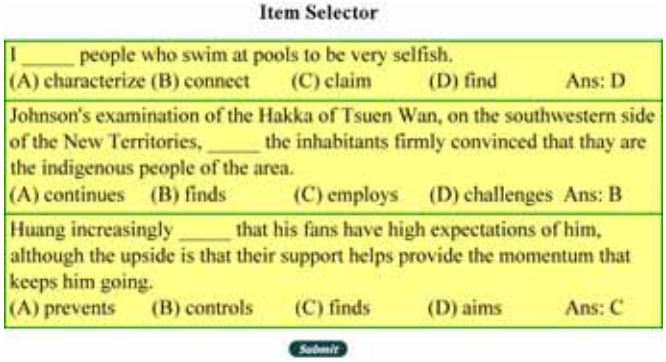

圖4　依圖 3 的規格所產生的試題

在評估階段我們要求試題產生器（item generator）產生 200 道試題。為了模擬真實情況下測驗題型的分布，我們對於以動詞、名詞、形容詞與副詞為答案的試題，個別分配不同的題數。參考先前的研究成果 [18]，我們從 1992 年到 2003 年台灣的大學入學考試的英文克漏詞試題中，歸納出答案的詞性不外乎 4 種，所佔的比例分別為：動詞 35%、名詞 30%、形容詞 20%與副詞 14%。因此，我們依相似的比例選用 77 題動詞、62 題名詞、35 題形容詞與 26 題副詞做為答案以評估系統在擷取目標句子上的效能。

在評估的過程裡，我們檢查產生的試題的答案，其詞義與詞性是否能夠符合需求。我們使用式子 (4) 並設定門檻值為 0.7，對試題中的答案作詞義辨析。實驗結果顯示在表 3，事實上，結果與表 2 的差距並不大。因為在標記詞性的正確率相當高的情況下，表 3 的實驗結果主要仍受到詞義辨析的正確率的影響。不論對於何種詞性的答案而言，在每產生小於 2 道試題之中，就有 1 道試題的答案能符合所要求的詞義與詞性。

此外，我們亦依照 [18] 對四種詞性的試題所歸納的比例，由系統產生 200 道克漏詞試題以檢驗誘答選項的品質，並判斷這些產生的試題是否能確保 4 個選項中僅有 1 個是正確選項（即答案本身）。由於試題是由語料庫中任意篩選出來的句子，所以我們是由人工判斷試題是否有唯一的答案，而以四個選項只有一個可以作為答案的題目來計算正確率。表 4 顯示，我們的系統在大多數情形下，能夠符合多選題的基本要求。

由於詞義辨析的工作相當具有挑戰性，而試題產生器能回傳大量的候選試題供測驗編撰者挑選，我們認為這套系統足以實際用於教師準備測驗卷的輔助工具。

表3　擷取目標句子的正確率

試題類型	答案的詞性	試題數量	目標句子的正確率
克漏詞試題	動詞	77	66.2%
	名詞	62	69.4%
	形容詞	35	60.0%
	副詞	26	61.5%
		總結	65.5%

表4　產生誘答選項的正確率

試題類型	答案的詞性	試題數量	誘答選項的正確率
克漏詞試題	動詞	64	90.6%
	名詞	57	94.7%
	形容詞	46	93.5%
	副詞	33	84.8%
		總結	91.5%

6.3 更多的應用

我們已將自動產生試題的系統實際應用於政治大學大一英文課程的隨堂測驗，並整合試題產生器到網路英文學習系統的環境中 [4]。在這個系統中，我們有兩項主要的子系統：測驗編撰系統與線上測驗系統。使用測驗編撰系統，測驗編撰者可以從圖 4 的介面中挑選試題，並將試題存放到測驗卷中，之後可依個人需求編輯測驗卷中的試題，包題幹、答案、誘答選項與正確選項等，測驗編撰者可對試題獲得最大的控制權。當測驗卷編輯完成後，輸入測驗卷的標題即可製成一份線上測驗卷。使用線上測驗系統，學生可透過網路進行線上測驗，並能夠立即獲得系統回報的成績（如果測驗編撰者有開放此功能）。學生的作答情形會記錄在學生模型（student modeling）中，系統將利用這些資料分析試的題難易度與試題的鑑別度。

為了支援不同題型的克漏詞測驗，我們也開發了產生慣用語（idiom）試題與片語（phrase）試題的系統。圖 5 描繪了這部份功能的輸出情形。更進一步地，我們的系統提供學生英文聽寫能力的測驗 [5]。在可見的未來內，我們計劃擴展我們的系統以支援全方面英文學習的需求（聽、說、讀、寫），並適性學生的能力以加強我們系統 [9]。

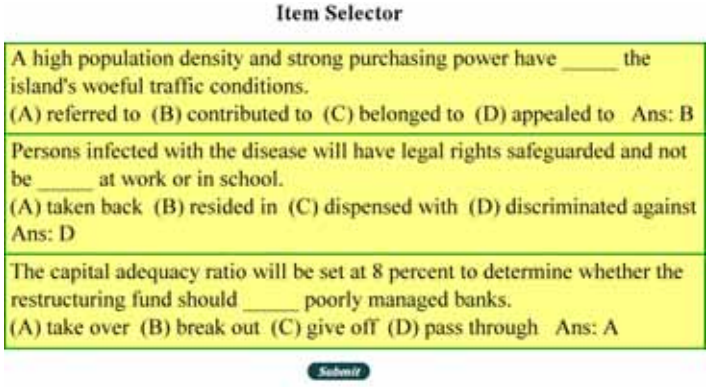

圖5　英文片語試題的範例

7 結論

在本論文中，我們提出了以自然語言處理技術為基礎的方法，可依特定需求而產生克漏詞試題，對於測驗編撰者在編寫測驗卷上有相當的助益。藉由將詞義辨析的演算法加入產生試題的流程中，使得產生的克漏詞試題能夠包含具有指定詞義的答案。詞義辨析本身並不是件容易的工作，在自然語言處理的相關研究中已研究了數年。對於一個目標句子中的答案而言，雖然我們的方法並不能對其做到最佳的詞義辨析，但是在產生試題的工作上已提供一個重要的幫助。畢竟，語言學家與心理學家普遍認為，詞義辨析仍需要來自上下文甚至是整個段落的資訊，而並非僅僅是單一個句子，而在單一句子做詞義辨析並非不可能，只是有較高的難度 [10]。

我們也提出一個新的策略來挑選克漏詞試題中的誘答選項。利用 collocation 為基礎的方法與詞頻統計資料，我們能挑選與答案具有相似挑戰性的誘答選項，並確保產生的試題中，有 90%以上的試題，答案是唯一或最佳選項。

既然測驗編撰者可以要求我們的系統傳回大量的克漏詞試題，並從中挑選數道最合其意的試題以編入測驗卷中，我們並不需要建立一個完美的電腦輔助試題編寫系統。目前我們的系統能夠做到每產生 1.6 道試題中，就有 1 道能夠用於實測的試題。然而，我們意圖考慮較深入的語言特徵來改進詞義辨析的正確率，以增進我們的系統的效能，並可望在不久的將來更新我們的實驗結果。

我們的研究仍可朝三個大方向持續發展。其一是句子的分析從語法的層面深入到語意的層面，檢驗句子是否含有完整的語義，藉以提高試題的品質；其次是從學生作答的情形中，歸納出控制試題難易度的因子，期許系統能大致猜得試題的難度；最後，我們希望能由英文教師的觀點，檢視系統的效能，比較教師在使用本系統後，是否在出題效率上有顯著的提升。

致謝

我們感謝多位不具名的評審對本文原稿的指正和建議，雖然因為篇幅和一些其他的限制使得我們暫且不能完全依照評審的建議加入新的材料，不過我們會在未來參照評審的珍貴建議加強論文的內涵。本研究承蒙國家科學委員會資助之研究案 91-2411-H-002-080 和 92-2213-E-004-004 的部分補助，謹此致謝。

參考文獻

[1] D. Coniam, A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests, *Computer Assisted Language Instruction Consortium*, **16** (2–4), 1997, 15–33.

[2] P. Deane, K. Sheehan, Automatic item generation via frame semantics, Education Testing Service: http://www.ets.org/research/dload/ncme03-deane.pdf (2003).

[3] I. Dennis, S. Handley, P. Bradon, J. Evans, S. Nestead, Approaches to modeling item generative tests, in: *Item Generation for Test Development* [2] 53–72, 2002, 53–72.

[4] Z.-M. Gao, C.-L. Liu, A Web-based assessment and profiling system for college English, *Proc. of the 11th Int'l Conf. on Computer Assisted Instruction*, 2004, CD–ROM.

[5] S.-M. Huang, C.-L. Liu, Z.-M. Gao, Toward computer assisted learning for English dictation, *Proc. of the 2003 Joint Conf. on Artificial Intelligence, Fuzzy Systems, and Grey Systems*, 2003, CD–ROM.

[6] S. H. Irvine, P. C. Kyllonen (Eds.), *Item generation for test development* (Lawrence Erlbaum Associates, 2002).

[7] T. Johns, http://web.bham.ac.uk/johnstf/timcall.htm.

[8] D. Lin, Dependency-based evaluation of MINIPAR, *Proc. of the Workshop on the Evaluation of Parsing Systems in the 1st Int'l Conf. on Language Resources and Evaluation*, 1998,.

[9] C.-L. Liu, Using mutual information for adaptive student assessments, *Proc. of the 4th IEEE Int'l Conf. on Advanced Learning Technologies*, 2004, to appear.

[10] C. D. Manning, H. Schütze, *Foundations of statistical natural language processing* (MIT Press, 1999).

[11] A. Oranje, Automatic item generation applied to the national assessment of educational progress: Exploring a multilevel structural equation model for categorized variables, Education Testing Service: http://www.ets.org/research/dload/ncme03-andreas.pdf (2003).

[12] C. J. Poel, S. D. Weatherly, A cloze look at placement testing, *Shiken: JALT (Japanese Assoc. for Language Teaching) Testing & Evaluation SIG Newsletter*, **1** (1), 1997, 4–10.

[13] A. Ratnaparkhi, A maximum entropy part-ofspeech tagger, *Proc. of the Conf. on Empirical Methods in NLP*, 1996, 133–142.

[14] P. Resnik, Selectional preference and sense disambiguation, *Proc. of the Applied NLP Workshop on Tagging Text with Lexical Semantics: Why, What and How*, 1997, 52–57.

[15] J. C. Reynar, A. Ratnaparkhi, A maximum entropy approach to identifying sentence boundaries, *Proc. of the Conf. on Applied NLP*, 1997, 16–19.

[16] K. M. Sheehan, P. Deane, I. Kostin, A partially automated system for generating passage-based multiple-choice verbal reasoning items, paper presented at the Nat'l Council on Measurement in Education Annual Meeting (2003).

[17] V. Steven, Classroom concordancing: vocabulary materials derived from relevant authentic text, *English for Specific Purposes*, **10** (1), 1991, 35–46.

[18] C.-H. Wang, C.-L. Liu, Z.-M. Gao, Toward computer assisted item generation for English vocabulary tests, *Proc. of the 2003 Joint Conf. on Artificial Intelligence, Fuzzy Systems, and Grey Systems*, 2003, CD–ROM.

[19] Y. Wilks, M. Stevenson, Combining independent knowledge sources for word sense disambiguation, Proc. of the Conf. on Recent Advances in NLP, 1997, 1–7.