電腦輔助閱讀測驗自動出題

楊媛茜 [1] 楊捷扉 [1] 張嘉銘 [2] 張俊盛 [1,2]
[1] 國立清華大學資訊系統與應用研究所
[2] 國立清華大學資訊工程研究所
{g936724,g936731,g936305}@oz.nthu.edu.tw, jschang@cs.nthu.edu.tw

摘要

電腦輔助自動出題爲自然語言處理領域近年來剛起步的研究，有相當大的應用潛力，很可能提供數位學習所急需的自動化工具。我們提出了一個閱讀測驗自動出題的做法。這個句法分析爲本的做法，涉及對閱讀的文章進行詞性分析、基本片語分析、完整剖析分析、N-gram 統計分析等等，再根據多重的策略，擷取具備特定句法樣式的句子，形成題目、答案、誘答項目。我們也根據這個做法，製作了電腦輔助閱讀測驗線上自動出題系統雛形，MARCT (Machine-Assisted Reading Comprehension Test)[1]，並對 MARCT 的各種出題策略設計適合的評估方式。實驗的結果顯示，MARCT 可以產生適當的題目，只要稍加修改，即可運用於練習或考試。

關鍵詞： 電腦輔助自動出題，關鍵詞抽取(Key Term Extraction)，問題類型(Question Type)，誘答項目(Distractor)。

1. 簡介

閱讀測驗目的在於了解學習者對某文章的掌握程度，因此題目內容即由文章內文而來，且多爲選擇題。此種學習策略在英文閱讀能力的培養上是極爲普遍地被使用。然而，在出題的過程中，由於必須先了解文意，找出文章的重要概念，而後設計題目，並且放入易使學習者誤判爲正確的答案選項（誘答項目）供爲選擇，步驟繁複，整個必須耗費極大的人力與時間。若是能夠利用電腦自動出題，並保持出題品質，勢必能有效提升閱讀測驗出題的效率，以強化自主學習、學習評量。

人工或自動化的出題，都需要有一套做法，才能達到考試合理的效度和信賴度。若是隨意選取文章的句子作爲考題的主題，且擷取其中任意一個字詞或片語，作爲標準答案，再配以任意的誘答項目，並不能達到我們預期的目標。另外考試題目也不宜集中在單一的題目類型，使得受測者有預期心態，無法測驗出真正的語言能力。

在語言學習與語言測驗學理上，題目應該涉及閱讀文章的重要觀念與詞彙，以求測驗出受測者掌握全文重點的能力，而誘答項目應該和標準答案很相似，以避免一知半解的學生只要比較選項，就可以猜測到正確答案。例如，在圖 1 之有關土豬（aardvarks）之文章之閱讀測驗，就應該採用如例 1、2 有關土豬覓食、出沒的習性的題目。而正確答案爲白蟻（termites）時，誘答項目就應該是其他動物的常見食物，如魚、小哺乳類動物、植物等。

[1] MARCT home page: http://140.114.75.11/MARCT/index.htm

（1） Aardvarks specialized in eating _____ as long as 35 million years ago.

(A) fish

(B) termites

(C) small manuals

(D) plants

（2） Adult aardvarks are usually _____, coming together only for mating.

(A) territorial

(B) nocturnal

(C) social

(D) solitary

直覺上，利用基於詞彙的全面與局部分佈統計，可以得知 termite**s** 在本文中出現 5 次，比起遠遠高於 termite**s** 在一般文章中平均出現的次數。而第二次出現的"Aardvarks specialized in eating termite**s** as long as 35 million years ago. " 似乎比較文意完整，結構簡單。只要將 termites 挖空，就成了很適合的選擇題。而 termites 自然就是標準答案。就誘答項目而言，參考類似 WordNet 或 Roget 's Thesaurus 一類的同義辭典，並考慮配合主題文脈（topical context）(如，談論動物的文章)，以及局部文脈（local context）(作爲"eat"的受詞) 相合，似乎不難計算出像 fish、small manuals、plants 的誘答項目。這些和主題與局部的文脈相合的誘答項目，自然容易引誘一知半解的受測者，誤入陷阱，做出錯誤的選擇。

另外，擷取二至三句，轉換其肯定與否定語氣，做成是非型的選擇題，也是一個常見的題目型態。例如針對圖一的文章，我們可以產生例 3 的題目：

（3） Which of the following statement about the aardvark is TRUE?

(A) The name aardvark comes from a word meaning "earth pig."

(B) Because the aardvark shares some similarities with the South American anteater, the two are related.

(C) BOTH

(D) NO ANSWER

其他常見的題目型態，包括數字型（如例 4）與清單項目型的選擇題（如例 5）：

（4） Aardvarks give birth to _____ at a time.

(A) one offspring

(B) two offsprings

(C) three offsprings

(D) four offsprings

（5） **_____ use abandoned aardvark holes as shelter.**

(A) Bats

(B) ground squirrels

(C) hares

(D) ALL OF THE ABOVE

The name aardvark comes from a word meaning "earth pig." Although the aardvark, endemic to Africa, shares some similarities with the South American anteater, the two are not related. The last survivor of a group of primitive ungulates, the aardvark could more accurately be called a near-ungulate that has developed powerful claws.

The aardvark has a short neck connected to a massive, dull brownish-gray, almost hairless body that has a strongly arched back. The legs are short, the hind legs longer than the front ones. The head is elongated and ends in a long, narrow snout, with nostrils that can be closed. The long, tubular ears are normally held upright but can be folded and closed. The short but muscular tail is cone-shaped and tapers to a point. The thick claws on the forefeet are used as digging tools.

Aardvarks are found in all regions, from dry savanna to rain forest, where there are sufficient termites for food, access to water and sandy or clay soil. If the soil is too hard, aardvarks, despite being speedy, powerful diggers, will move to areas where the digging is easier.

Aardvarks are nocturnal, usually waiting until dark before they emerge from their burrows, although after a cold night, they may occasionally sun themselves. They leave a distinctive track from dragging their tails during which their travels average one to three miles but can range up to 18 miles a night. Aardvarks are seldom seen, but their presence in an area serves many other animals. Bats, ground squirrels, hares, cats, civets, hyenas, jackals, porcupines, warthogs, monitor lizards, owls and other birds use abandoned aardvark holes as shelter. The burrows vary from simple chambers with one entrance, to a complicated maze of galleries with 20 or more entrances. Aardvarks keep their burrows clean they deposit their dung in a hole away from the entrance, carefully covering it with earth.

As it is nocturnal and has poor eyesight, the aardvark is cautious upon leaving its burrow. It comes to the entrance and stands there motionless for several minutes. Then it suddenly leaps out in powerful jumps. At about 30 feet out it stops, raises up on its legs, perks up its ears and turns its head in all directions. If there are no sounds, it makes a few more leaps and finally moves at a slow trot to look for food.

Aardvarks specialized in eating termites as long as 35 million years ago. At night they go from one termite mound to another, dismantling the hills with their powerful claws. Insects are trapped by the aardvark's long protractile tongue, which is covered with a thick, sticky saliva. Sometimes the aardvark will press its snout against an opening in a mound and suck up the termites. Aardvarks, with their keen sense of smell, also hunt for the long columns of termites that move outside the mounds at night.

Aardvarks give birth to one offspring at a time. The pinkish, hairless newborn stays inside the burrow for about 2 weeks and then begins to follow its mother in her search for food. The young first eats solid food at 3 months of age and is suckled until 4 months.

At about 6 months the young male becomes independent and goes off on its own, while the young female stays with the mother until after the next baby is born. The young female may then dig its own burrow a few yards away from its mother but still joins her to forage for termites.

The adult aardvark's principal enemies are human (who sometimes kill it for meat), lions, hyenas and leopards; pythons also take the young. When in danger the aardvark takes to the nearest hole, or rapidly excavates one, pushing the dirt backwards with its feet and moving the dirt away with its tail. But if cornered, it defends itself by sitting up, using its tail, shoulders and foreclaws- or it will lie on its back and strike with all four feet.

▲圖一 在非洲野生動物基金會網站 http://www.awf.org/ 刊載的有關土豬的文章

基於這些多重的策略，我們提出一個新的自動出題的方法。爲了評估該做法的可行性，我們也製作了一個叫做 MARCT 的系統，初步的實驗，發現在不損及題目品質之前提下，藉由電腦輔助，的確可以減少出題的人力負擔，並加快出題速度。我們以有關於動物生態的文章，探索閱讀測驗的問題與自動化的課題。閱讀文章引用自非洲野生物種基金會(AWF)網站[2]，共五十篇的非洲野生動物介紹。藉由關鍵詞抽取、閱讀測驗題型分析兩種策略，MARCT 能產生出三種基本測驗題型的題目，也能產生有意義的誘答項目。

2. 相關研究

電腦自動出題爲一項剛起步的研究，相關文獻中，僅有 Mitkov and Ha(2003)提出半自動的閱讀測驗的出題方法，並以測驗理論的方法評估出題的好壞，經分析比較，發現電腦輔助之下，出題的時間較短，而題目的品質較佳。Mitkov and Ha 的做法，主要採取單一，以關鍵詞抽取爲主的策略。在這個做法下，電腦首先分析文章，擷取其中最重要的關鍵詞作爲題目的正確答案，再進而選取含有關鍵詞的特定型態的句子轉化爲問句，最後產生容易產生混淆的誘答項目。

王俊弘、劉昭麟和高照明(2004)藉由自然語言的詞義辨析技術，加上搭配詞(collocation)概念生成誘答選項機制，產生自動英文克漏詞試題系統。

我們觀察語言教學專家，所製作的閱讀測驗題，發現有許多種出題的型態，並不限於 Mitkov and Ha 所提出的題目型態。爲突破一些自動出題的限制，因此我們參酌此類的題目，採用多重策略來進行自動或半自動出題的研究。

3. MARCT 自動出題的作法

我們主要採取兩種策略：第一種策略爲"Computer-Aided Generation of Multiple-Choice Tests"該篇文章中所提到的方法，即是先決定關鍵詞爲何，然後利用抽取出來的關鍵詞以及所在的句子出題；第二種策略爲分析閱讀測驗。

3.1 關鍵詞抽取

在這一部份，我們做了三項分析，分別爲名詞(片語)頻率分析、AWF 單字頻率與 BNC 單字頻率比較、以及利用最長共同子序列(Longest Common Subsequence)在兩兩句子之間嘗試找出可能的類似句型。

3.1.1 名詞(片語)頻率分析

我們採取和 Mitkov and Ha 相同的關鍵詞抽取策略，因此以抽出名詞片語做爲出題的切入點。利用基本片語分析將 50 篇文章中標示爲名詞片語的單字取出，然後計算其出現頻率。我們觀察到，在 50 篇文章內高頻的名詞片語內以虛字居多，但以虛字作爲關鍵詞來出題並沒有任何意義。此外，高頻字並不一定就是文章中的關鍵詞，而低頻字有時反而是文章中較特殊且重要的資訊。這可能是由於我們與 Mitkov and Ha 的訓練資料文類不同，因此並不適用以名詞片語的出現頻率來決定文章的關鍵詞。

---

[2] African Wildlife Foundation (AWF) home page: http://www.awf.org/

3.1.2 AWF Uni-gram 頻率與 BNC Uni-gram 頻率分析

接下來，我們採取統計 N-gram 的出現頻率這個方式，利用 BNC(British National Corpus)與 AWF 50 篇文章單字頻率的比較來取得關鍵詞(Constantin Orasan 2001)。BNC 是包含超過一億個單字的語料庫，其資料包羅萬象，可以代表一般性文章最可能出現的單字頻率情形，利用 BNC 這樣的特性與我們的 50 篇文章中的單字頻率作比較，可以突顯出，有哪些單字在撰寫動物介紹的文章中是經常出現的。

我們根據經驗法則，將這 50 篇文章的單字出現頻率比在 BNC 中單字出現頻率排名多上 1000 名以上者抽取出。由我們的統計結果得知，的確有許多單字是在動物描述性文章中經常出現的字眼，比如 female、animal 以及 specie 等等，但卻不是我們所要、可用以出題的關鍵詞。

由以上兩種關鍵詞抽取策略，我們歸納出一個重要的結論：利用統計 N-gram 出現頻率的方式，僅能得到在特定文類中經常出現的字彙，卻無法從這些字彙判定可以出題的關鍵詞。因此，我們採取不同的關鍵詞抽取方式。

3.1.3 最長共同子序列(Longest Common Subsequence)

在這個實驗中，我們嘗試利用最長共同子序列(LCS)來找出兩兩句子間可能的共同表面式樣(surface pattern)，然後利用這樣共通的式樣來找出句子內的關鍵詞。結果發現，最大的問題在於我們的訓練資料太少，利用 LCS 得到的共同式樣有限，且相似度不高，無法作爲出題的類型。我們的最初想法是，若是兩兩句子有部分一樣，那麼「不一樣的部份」很有可能是該文章關鍵詞所在。但是分析結果顯示，訓練資料過少致使我們無法找出高相似度的共同表面式樣，「不一樣的部份」也就顯得沒有意義，失去出題的價值。因此，我們改採第二出題策略。

3.2 閱讀測驗題型分析

以單字頻率分析嘗試找出關鍵詞的方式，的確可以找到在描述動物的文章中重要的名詞(片語)，但是再由此過濾出可茲以出題的名詞(片語)則有相當程度上的限制，原因是我們並沒有界定關鍵詞出現的範圍，以至於無法精確地抽取；其次，關鍵詞所在的句子是轉爲問句的關鍵句，即使關鍵詞抽取出來，但是句子的複雜度以及句子本身可以成爲哪一種問句卻無法掌握。基於上述原因，我們限定關鍵詞出現的範圍，並藉以確立閱讀測驗的題型。

我們從網路上蒐集一千多句閱讀測驗的測驗題目，計算 trigram 找出最常出現的組合爲何，再與 BNC 的 trigram 做比較，結果如表一：

▼表一 BNC trigram 與閱讀測驗題目用字(RC)trigram 之出現頻率排名比較
(僅列出排名前八部分)

Word1	**Word2**	**Word3**	**RC Rank**	**BNC Rank**
which	of	the	1	10041
of	the	following	2	2148
what	is	the	8	895
in	which	of	17	453955
how	many	species	19	1184740
many	species	of	20	78415
what	was	the	21	2164
the	world	‘s	23	3410377

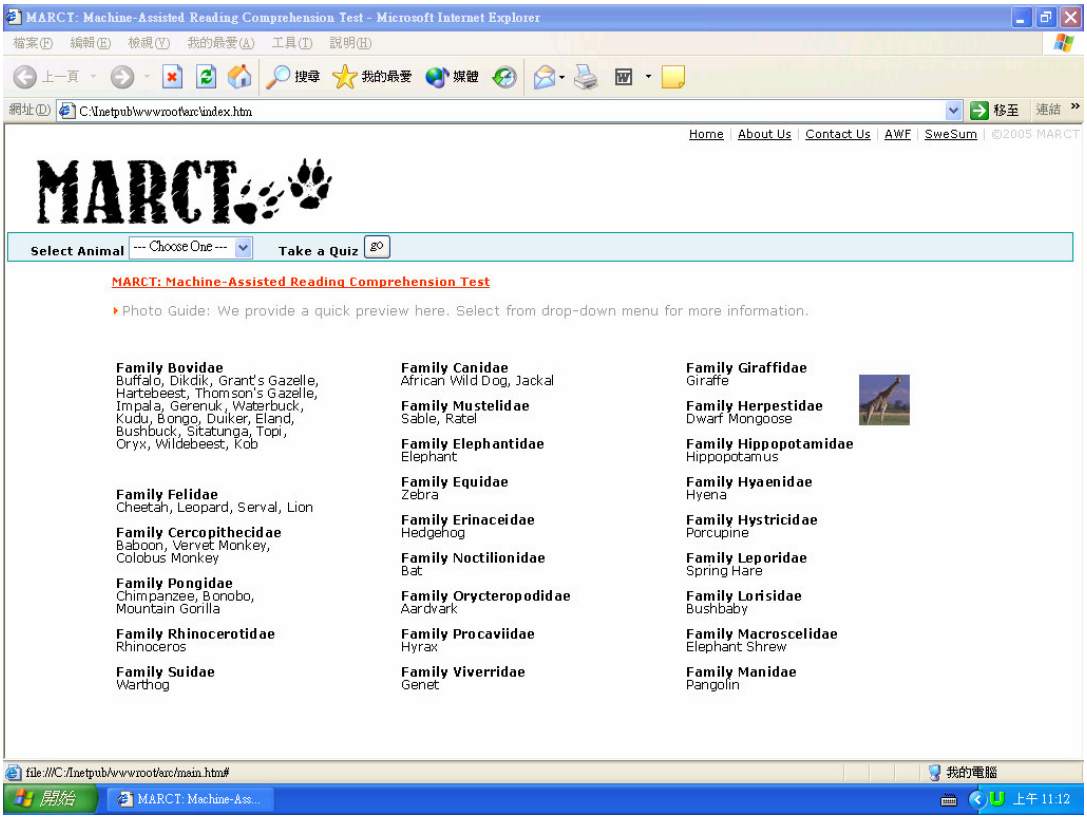
由表一我們可以看到，”which of the”是在閱讀測驗題目中最常出現的用字，和排名第二的”of the following”結合似乎也是合乎文法且有意義的，因此我們推論，在動物性描述文章中，最常出現的問句形式可能是”Which of the following” + Other Statement。我們決定根據這個推論，把它應用在兩種題型上：是非題(True-False Question)以及清單項目題(List Question)。此外根據經驗法則，文章內出現的數字通常是出題重點所在，因此我們也會進行數字類型的出題。總結以上的閱讀測驗題型分析結果，MARCT 將會自動產生是非題、清單項目題以及數字數量題。至於做法細節，將會在第四章實作階段做詳細描述。

4. 實驗與實作

由統計既有人工自動出題之 N-gram 發現，是非題、清單項目題以及數字數量題爲常出的題型。在實作部分，我們利用上述方法，建置一個名爲 MARCT (Machine-Assisted Reading Comprehension Test)的系統，以下分別介紹系統功能以及技術細節。

4.1 MARCT (Machine-Assisted Reading Comprehension Test)

MARCT 能針對閱讀文章自動出題，讓使用者測試對文章內容了解程度。其中，閱讀內容資料來自非洲野生物種基金會(AWF)，對非洲各種動物有完整的介紹。系統提供每一篇閱讀文章至少兩題以上的閱讀測驗，圖二爲系統首頁：



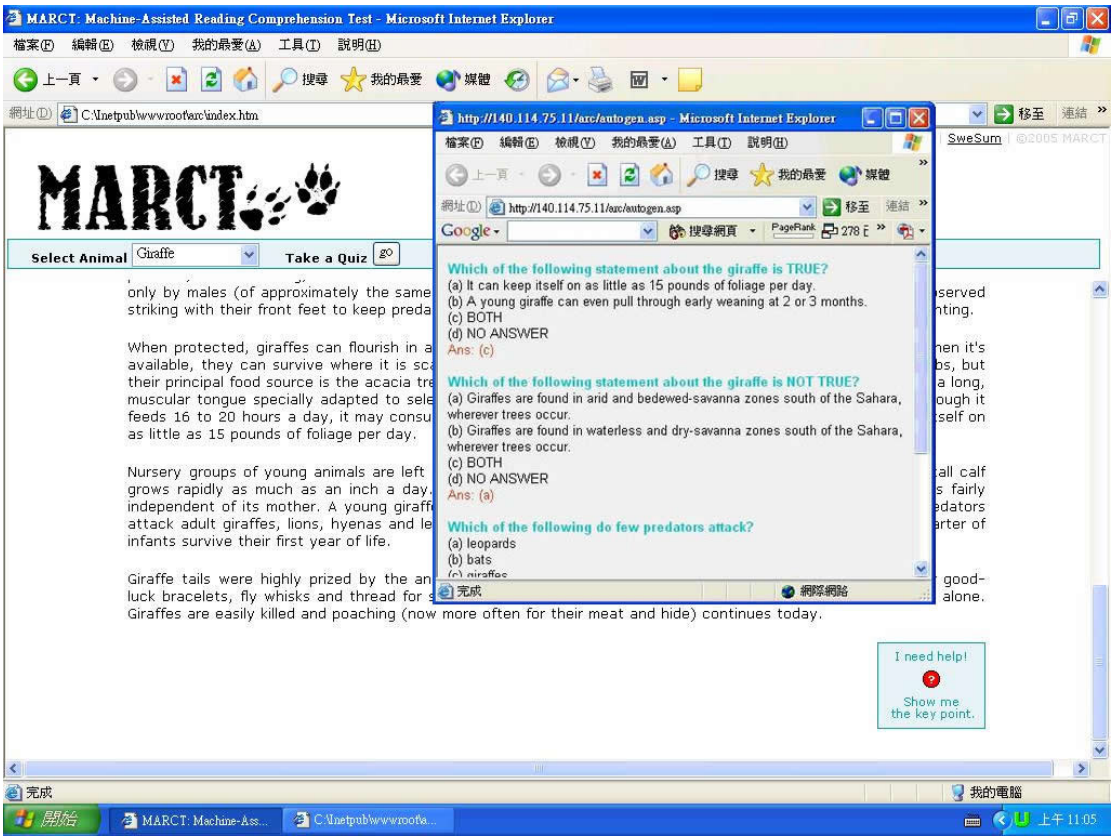
▲ 圖二 系統首頁

粗體字部份，爲動物主要分類類別，隸屬其下者爲該類別之動物名稱，點選下拉式選單選擇動物，即可觀看文章內容。

當閱讀上有困難、無法掌握重點時，只要點選便可看到原文利用 SweSum 系統[3]產生的摘要，以及計算 BNC、AWF 文章內單字頻率後，將兩者排名相差 1000 名所得到的可能文章關鍵字來標示出重點句子及文章關鍵字的結果。

[3] SweSum website：http://swesum.nada.kth.se/index-eng.html

按下"go"按鈕，開始自動出題。



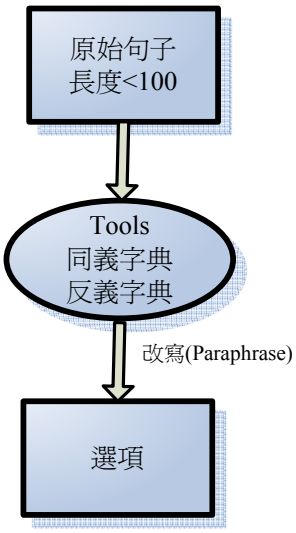
▲ 圖三 自動出題

4.2 系統製作

我們將測試資料切成句子，作爲自動出題的基本單位。以下分別介紹系統細節：

4.2.1 是非題

分析既有人工自動出題，我們可歸納出此類題型問句型式爲" Which of the following statement is (not) true?"；在選項設計上，爲考慮選項敘述需精簡，因此取整句長度不超過 100 位元之句子作爲出題對象。



▲ 圖四 選項生成示意圖

圖四中，我們利用 WordNet 所有單字及詞性，製成各單字各詞性之同義字典，以及各單字動詞、名詞、形容詞及副詞之反義字典。爲簡化實驗，我們不考慮歧義辨析，只抽出各單字排名第一位的同義字/反義字收入同義字典及反義字典裡。接下來，我們做字與字之間的替換，並除去因沒有同義/反義字而無法出題的句子，以及一些句子中具承接前後文語意的連接詞，例如 however。測驗題產生如下：

Which of the following statement about the aardvark is TRUE?

(a) If there are no sounds, it makes a few more leaps and finally moves at a boring trot to look for food.

(b) The heavy claws on the forefeet are used as digging tools.

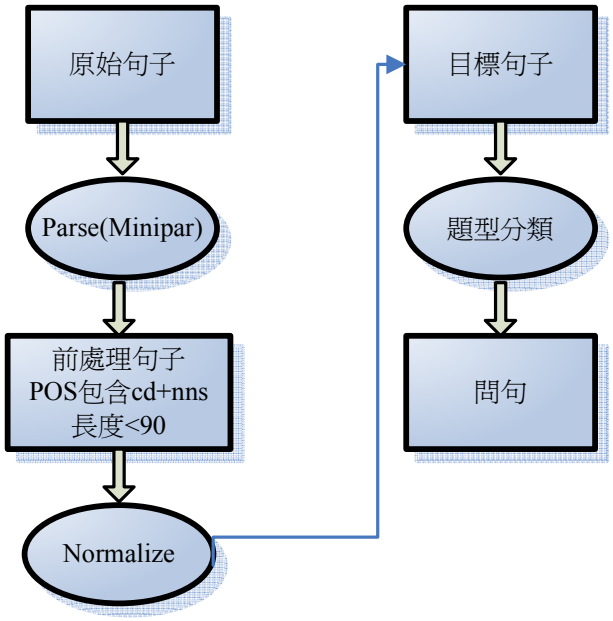
(c) BOTH

(d) NO ANSWER

Ans: (c)

4.2.2 數字數量題

由於使用句子產生問句必須先了解句子之句法結構，因此我們先利用 Minipar 產生剖析樹，利用詞性過濾出可利用該類型出題之句構，並且爲簡化實驗，僅取長度小於 90 位元之句子作爲出題對象。



▲ 圖五 問句生成示意圖

在題型分類上，是以根據抽出來的量詞，來歸類屬於哪種題型，分類如表二：

▼ **表二 題型分類**

Question Type	**量詞概念**
When/ How long	時間
How many	個體
How tall	長度
How large	空間
What is the weight of	重量

句子中的主詞、動詞與其他資訊，我們利用 Minipar 產生出之第一個出現的主詞、最靠近主詞的動詞、保留包含數字資訊且最靠近主詞、動詞之介係詞片語以及其他資訊作爲問句出題之依據，問句類型分爲三大類：

1. 疑問詞+ be 動詞+主詞+除去主詞動詞介係詞片語的其它原句資訊+?
2. 疑問詞+助動詞+主詞+原型動詞+其它原句資訊+?
3. What is the weight of +主詞+?

結果產生如下：

When/ How long does the young eat first solid food and is suckled until 4 months?

(a) at 1 months of age

(b) at 2 months of age

(c) at 3 months of age

(d) at 4 months of age

Ans: (c)

4.2.3 清單項目題

我們抽取出句子單字詞性包含(,+NN+CC+NN)或是(,+NNS+CC+NNS)之形式，作爲出題之對象。

在問句產生中，我們分析既有人工自動出題，得到此題型題目爲"Which of the following +be 動詞+主詞+?" 或是"Which of the following do/does +主詞+一般動詞+?"；主詞與動詞的抽取則依據數字數量題之規則。

選項的抽取上，正確答案即由原句子而來，誘答選項則利用 Google Sets，將三個正確選項作爲關鍵字查詢，扣除與正確答案相同部份，其餘取出作爲誘答選項。

出題結果產生如下：

Which of the following are aardvark principal enemies?

(a) kangaroos

(b) leopards

(c) lions

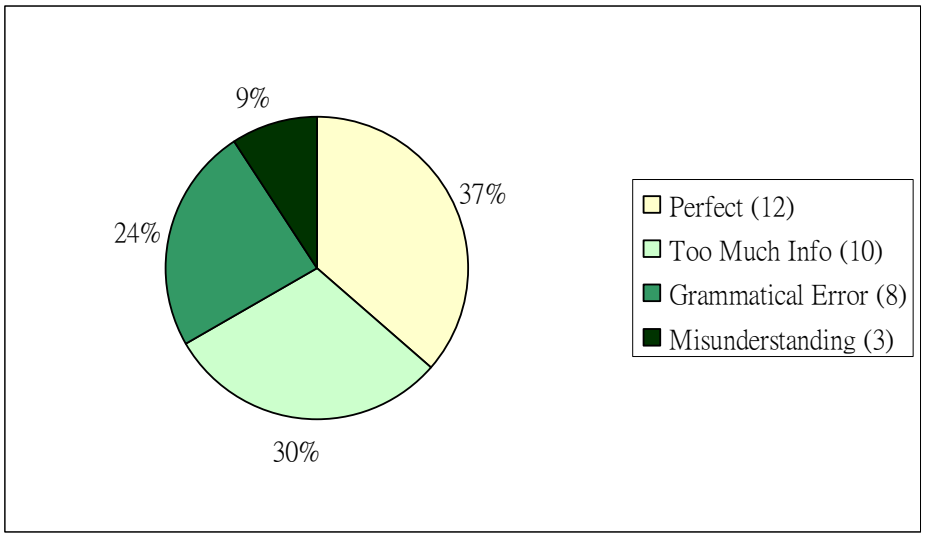
(d) hyenas

Ans: (b) (c) (d)

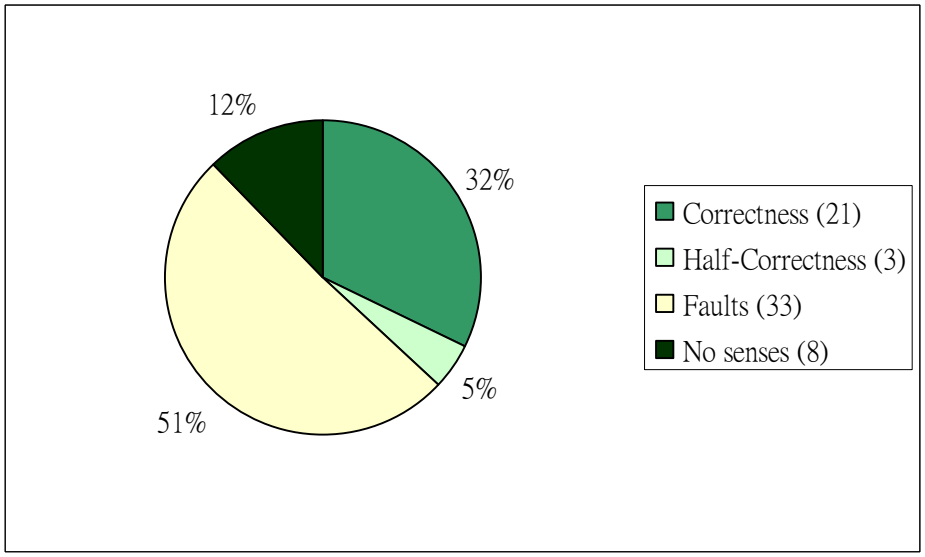
5. 系統效能評估與討論

在實作中我們利用先確立題型作爲出題策略，讓每篇文章產生出至少兩題之測驗題，增進出題之速度與效率，然而，由於實驗之簡化以及尚未考慮之因素，因此，仍有許多不足及可改進之處。

是非題在改寫的過程中由於未考慮歧義辨析，造成抽出的同義詞或反義詞的語意不夠精確，以致影響使用者閱讀選項時的語意順暢度；此外，作詞義辨析時便有詞性標錯的問題，連帶改寫出現錯誤；而改寫僅做最基本單字對單字的層面，讀者依然可由文中找出蛛絲馬跡，降低此題型的鑑別度。



數字數量題共轉換 33 句包含數字之句子，結果如圖六所示。在這部份的實驗，數字資訊包含在介系詞片語，但在其他領域實際情形可能並非如此；此外，許多量詞並非只對應到一種問題類型，比如 month，可能是意指一段時間或者是一個時間點，造成疑問詞的使用不同；由 Minipar 抽取動詞時有些取錯誤，例如無法抽出完整的動詞片語；當一個句子本身包含過多資訊時，也可能增加出題的困難度；而 Minipar 偶爾不正確的標記(準確率爲八成)，也連帶我們的產生結果受到影響。



清單項目題則轉換 65 句可做爲該題型之句子形成測驗題，結果如圖七所示。造成許多錯誤的原因包括：Minipar 找主詞與動詞的結果，經常產生錯誤；而當句子過於複雜，比如句子中含有兩個以上的動詞或是有動詞轉爲形容詞型態的句子，也容易導致出題結果失敗；另外，雖然大部分的句子可以適用我們的策略來出題，但仍有部分句子出現語意上的錯誤；更重要的一點，在於我們出題基本假設是題目生成可以透過句子當中資訊做處理，但在實際處理情形卻並非如此，往往使句子原意無法呈現，而且題型上也無法全自動出題。因此在出題上，若有其他的工具能輔助支持會更好。

6. 結論與未來研究方向

MARCT 系統爲自動出題的初步嘗試，仍有許多尚待改進與突破之處。但 MARCT 的確也達成了某些我們寄予自動出題的期許─的確能節省出題的時間以及增進出題的效率。然而，自動出題這個領域仍然有許多尚待克服的限制與有待突破的困境。例如在題目的處理上，我們還是只能依賴句法結構來做爲出題的根據，也就是只能在句法這個層面做發揮，仍然無法涉及人工出題能掌握的、更深層的語意題型，這一直是目前這個研究領域中無法突破的瓶頸，也可以說是在整個自然語言處理的發展中，一直努力挑戰的部份。

參考文獻

[1] African Wildlife Foundation (AWF) home page. World Wide Web page. http://www.awf.org/

[2] SweSum - Automatic Text Summarizer home page. World Wide Web page. http://swesum.nada.kth.se/index-eng.html

[3] British National Corpus (BNC) home page. World Wide Web page. http://www.natcorp.ox.ac.uk/

[4] Constantin Orasan, Patterns in scientific abstracts, in Proceedings of Corpus Linguistics (2001)

[5] 王俊弘，劉昭麟，高照明。電腦輔助英文字彙出題系統之研究（Toward Computer Assisted Item Generation for English Vocabulary Tests）。（2002）

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

[7] 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 (2003): http://www.ets.org/research/dload/ncme03-andreas.pdf

[8] Ruslan Mitkov, Le An Ha, Computer-Aided Generation of Multiple-Choice Tests. (2003): http://clg.wlv.ac.uk/papers/ruslan-NAACL-03.pdf

[9] 王俊弘，劉昭麟，高照明。利用自然語言處理技術自動產生英文克漏詞試題之研究。（2004）