以網際網路內容為基礎之問答系統 “Why” 問句研究

沈天佐　林川傑　陳信希

國立台灣大學資訊工程學系

{tzshen,cjlin}@nlg.csie.ntu.edu.tw; hh_chen@csie.ntu.edu.tw

摘要

以 “Why” 開頭的問句，問題的答案是 “原因”。“原因” 有不同的型態，可能是一個片語、一個子句、一個句子，甚至跨越句子的範圍。目前的問答系統特別針對 “Why 問句” 研究的並不多，本文探討如何從文件中擷取出 “Why 問句” 的答案，文件的來源設定在網際網路。我們運用搜尋引擎取得相關文件，以描述因果關係的句型來擷取答案。由於句型本身可能會有歧義性，某個句型的出現並不代表一定是問句的答案，本文也針對這項議題進一步分析。我們並將所發展的問答系統，與另外兩個以網際網路為基礎的問答系統—AnswerBus 和 LCC，作了效能的評估。在以 50 個問句的測試中，我們的系統、AnswerBus 和 LCC 的 MRR 值分別為 0.623、0.429 和 0.229，顯示我們的系統的效能優於這兩個系統。

1. 緒論

問答系統接受使用者的自然語言問句，從一堆文件集中，找出問句的答案。透過問答系統，使用者可以直接得到答案，而不必自己瀏覽資訊檢索系統所傳回的一堆相關文件尋找答案。TREC (Text Retrieval Conference) 自 1999 年開始舉辦問答系統的效能評比 (Voorhees, 1999)，帶動近年來問答系統的研究風潮。TREC 評比的重點隨著研究成果的進展，每年都進行調整。以 2002 年為例，評比的重點在於參賽者的系統是否能夠準確地定出答案的範圍，而不是以一個固定長度的文字片段當作答案。

完整的問答系統分為兩步驟，第一個步驟是從所有文件中找出與問句相關的

文件，此即「資訊檢索」的部分。如何將自然語言問句轉換為適合資訊檢索系統的查詢字串，是個研究課題。第二個步驟是從相關文件中找出問句的答案，此稱為「答案擷取」，這個部分是問答系統主要研究重點。進行「答案擷取」，問答系統必須針對問句進行分析，以取得答案的類型。常見的「答案擷取」方法是利用 “Named Entity Tagging” 的技術，再加上 “問句與上下文相似度的計算”。從簡單的關鍵字比對，到較複雜的語意一致性判斷，都是可能的上下文與問句相似度計算方法 (Harabagiu *et al*., 2000a; Moldovan and Rus, 2001)。

以網際網路為基礎的問答系統研究，主要是利用網路上常見的搜尋引擎進行資訊檢索，以取得相關文件，再利用與 TREC 問答系統類似的技術來擷取答案。這種類型的問答系統，必須考量即時性，避免太複雜技術帶來的負擔。目前的研究有 Radev *et al*. (2001)、Radev *et al*. (2002)、Zheng (2002)、Lin (2002)。另外，網頁文件的一些特性，例如 HTML 標記、超鏈結、風格差異、內容正確性等，也是在研究上必須考量的議題。

目前大部分問答系統擷取答案方法，主要針對答案類型為 Named Entities。對於答案較複雜，沒有固定形式的問句類型，如 “Why … ?” 和 “How does *S V*?”，則較少有深入的探討與分析。Girju 與 Moldovan (2002) 曾經探討過回答 “cause-effect questions”，研究因果關係在文中的表達方法。不過這篇文章的重點擺在 <NP1 VERB NP2> 這種 pattern 上，其中的動詞必須是個 “causative verb”，例如：“cause”、“lead to”、“make”等。由於這些動詞未必一定代表因果關係，如 “make” 有時的意義為 “製造”，所以研究重點在於如何由 VERB、NP1 和 NP2 來判斷是否描述因果關係。

在閱讀測驗問答系統 (reading comprehension) 的研究上, Anand *et al*. (2000) 和 Riloff and Thelen (2000) 也有相關研究。系統針對一篇文章，找到問句的答案。TREC 問答系統與這類問答系統主要的不同點是答案來源為多篇相關文件，答案可能重複出現多次，有較多機會找到答案，但雜訊也會比較多。閱讀測驗問答系統則相反，答案可能只出現在文章中一次，所以需要較複雜的方法來找到不

是那麼明顯的答案，但另一方面雜訊會比較少。

第 2 節說明實作系統的架構，以及各個子系統。第 3 節引用 Penn Treebank 語料庫，分析擷取答案 patterns 的準確率。第 4 節為本系統的效能評估，並與另外兩個以網際網路為基礎的問答系統比較。第 5 節是結論與未來研究方向。

2. 系統概觀

2.1 資訊檢索系統

本文所提的問答系統架構如圖 1，只針對單一的問句類型 (也就是以 “why”開頭的問句) 進行處理，所以並未包含問句分析子系統，同時我們選擇 Google 來找出與問句相關的網頁文件。首先將問句轉為查詢字串，去掉問句中的停用詞 (stop words，包括疑問詞、介系詞、連接詞、代名詞、助動詞、某些副詞……等) 與標點符號，剩下來的字以空白相連接，為交給 Google 的原始查詢字串。由於 Google 採 AND 的方式來解讀關鍵字，一定要含所有關鍵字的文件才會被取回，所以有可能取回的文件篇數很少。

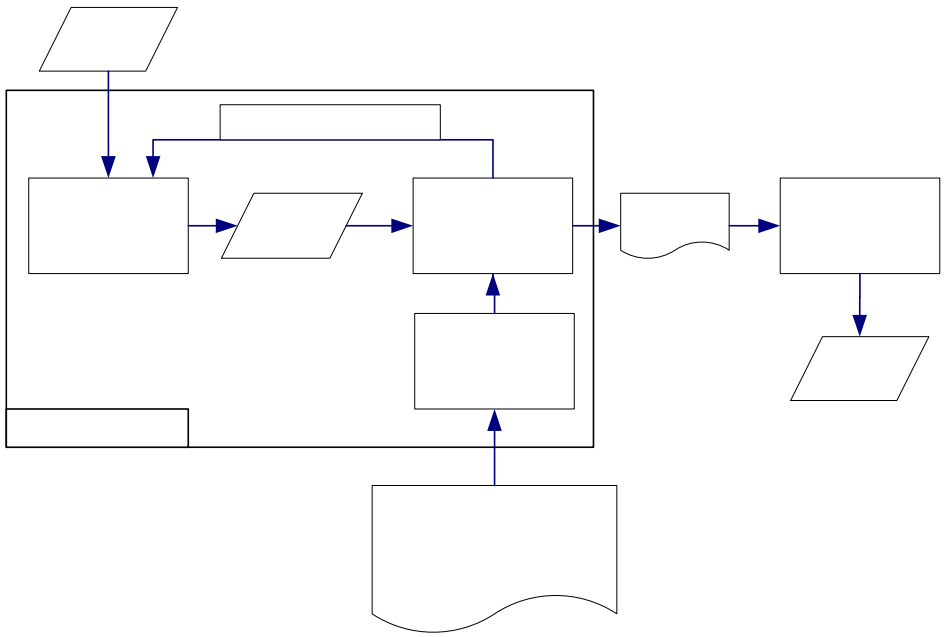

圖 1. 問答系統架構

當 Google 所找回的相關文件數量不足時，我們會修改查詢字串，再進行一次查詢以補足不足的部分。查詢字串修改的方法是刪除查詢字串中的某個關鍵字，產生新的查詢字串。我們選擇“權重”較小的先刪除，設定權重如下：

> 專有名詞 > 名詞詞組的 Head > 動詞詞組的 Head > 名詞詞組的其他字 > 動詞詞組的其他字 > 不在名詞詞組或動詞詞組的其他字

權重越大的關鍵字，與文件主題的關係越密切。

當文件中含有某些特殊字如“reason”時，可能表示此文件中提到某個事件的原因。因此，若在進行資訊檢索時，能夠將這些特殊字加到原有的查詢字串中，所檢索到的排名較前面的網頁文件，就會是那些既含有問句中的關鍵字 (表示和問句所問的主題相關)，而且內容又描述了某種因果關係的網頁文件。同時 Google 在檢索文件時，也考慮了各關鍵字在文中的接近程度。當各關鍵字在文件中越接近，文件的排名會越前面。可以協助尋找因果關係的特殊字，包括 “reason”、“why” 和 “because” 等。

2.2 答案擷取系統

要在文件中找尋表達因果關係的資訊，目前已知有四種情形：

一、 利用因果 patterns 來判斷文件中描述因果關係的部分。

二、 以整篇文章來解釋原因和理由。網際網路上較常看到這種情形，作者在問句處提供一個指向答案的超連結。

三、 原因和結果出現在前後文，兩者間並無明顯關連詞出現。

四、 某些動詞隱含因果關係，如 Girju and Moldovan (2002) 所做的研究，以及在 WordNet 中也有動詞間 causation 關係的資訊。

以下針對各情形詳細說明：

一、 利用因果 patterns 來判斷文件中描述因果關係的部分

在文法及修辭學上，有不少句型可用來描述兩件事之間的因果關係。我們試著利用這樣的句型來找出“原因”的部份。這樣的句型包括：

[EVENT] because [REASON].
[REASON], therefore [EVENT].
[EVENT] in order to [REASON].

其中 [EVENT] 代表結果事件，[REASON] 表示其發生原因。這些句型所得到的 patterns 不但可以用來判斷因果關係的資訊，也可以用來決定“原因”部份的邊界。

二、 以整篇文章來解釋原因和理由

在某些以教育為主題，或是提供常見問答集 (FAQ) 的網站中，就可看到這類以整個段落或整篇文章來解釋或回答一個問題的網頁。例如圖 2 即為“Why is the sky blue?”答案的網頁，其中答案的描述長達一整篇文章。

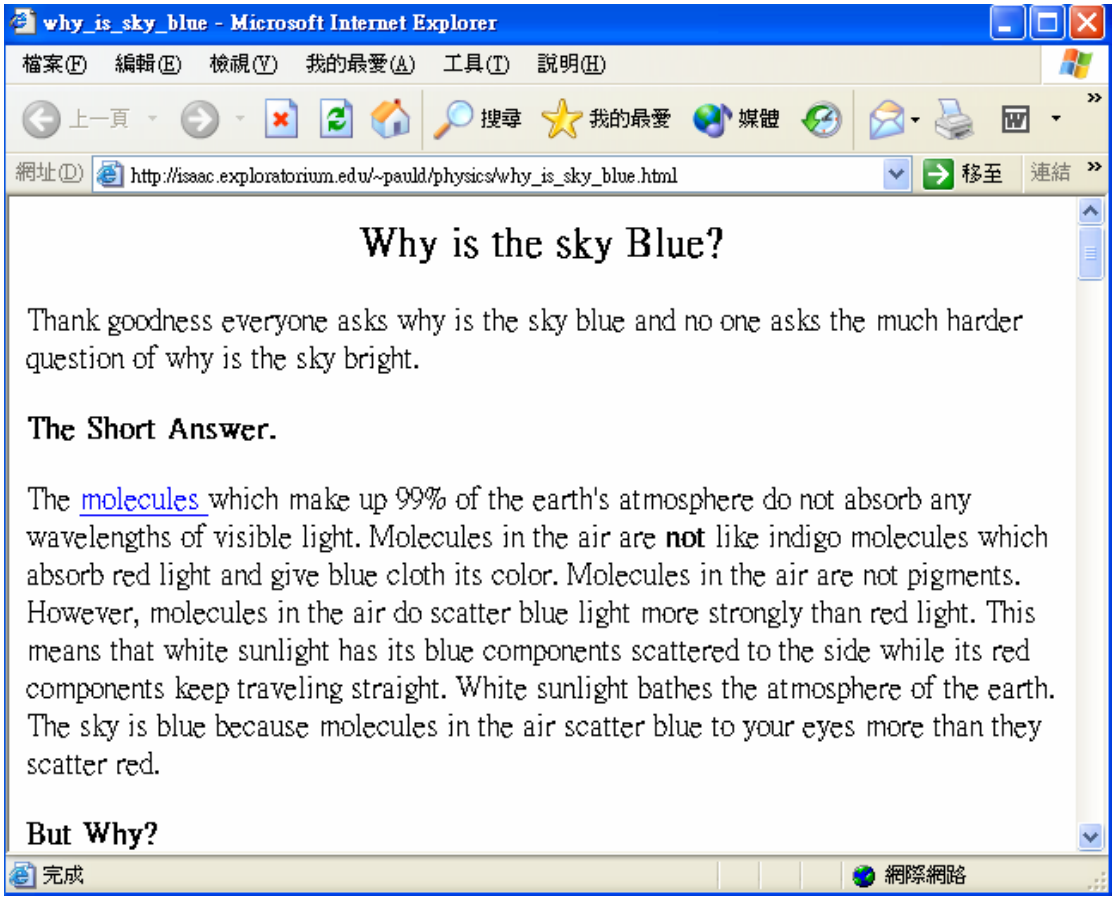

圖 2. 常見問題集答案網頁的範例

因為是人工建置的，這種情形所取得的答案無疑會是正確答案。我們所需要做的工作是，尋找此類的網頁並找到何者是它的原始問句。如果原始問句與使用者所問的問題是一樣的，則以此段文字 (或網頁連結) 提供給使用者做為答案。這方面的研究比較接近 FAQ Finding。

三、 原因和結果出現在前後文，兩者間並無明顯關連詞出現

在下面這個例子中：

問： Why can't ostriches fly?
答： The flightless birds … include ostriches, … . These birds have only small or rudimentary wings.

答句中的 “These birds have only small or rudimentary wings.” 就與前一話沒有直接的因果關連詞，但是人類仍可以知道這句和前句有著因果關係。

四、 動詞隱含因果關係

在 Girju 與 Moldovan 的研究中，某些動詞帶有因果次序的訊息，例如 “provoke”、“induce”。WordNet 中則提供了動詞之間 “cause to” 的關係，例如『“kill” cause to “die”』。然而這樣的動詞並不多，並不是描述因果關係的最主要方式。

由以上的說明，我們可以發現，方法一不但較為簡單，適用性又廣，非常適合於網際網路環境的問答系統建構。因此我們將重點放在因果 patterns 的建置，以及比對出文件敘述因果關係部分的方法。另外，為了處理第三種情形，我們設計了一個 pattern 稱為 “final pattern”，將在第 2.2.1 節中介紹。動詞隱含因果關係則可做為未來加強系統回答能力的有用資訊。

2.2.1 因果關係 patterns

在文法及修辭學上，有不少句型可用來描述兩件事之間的因果關係。我們試著利用這樣的句型來找出 “原因” 的部份。這樣的句型包括：

[EVENT] because [REASON].
[REASON], therefore [EVENT].
[EVENT] in order to [REASON].

其中 [EVENT] 代表結果事件，[REASON] 表示其發生原因。這些句型所得到的 patterns 不但可以用來判斷因果關係的資訊，也可以用來決定“原因”部份的邊界。

我們從一些相關書籍及研究成果中蒐集到許多表示因果關係的句型。當一個“why”問句被提出後，我們的問答系統會先利用上述 patterns 找出所有包含因果關係的句子，並且評估 patterns 中對應 [EVENT] 的部份與問句的相似度。當有個文句符合其中某一條因果 pattern，且對應 [EVENT] 部份高度相似時，系統就可以抽取出文句中對應 [REASON] 部份，做為回應給使用者的答案。

然而在實際實驗時，我們發現有些應用上的問題。首先，有的 patterns 是描述句子之間 (而不是句子之內) 的關係的。舉例來說：

> Molecules in the air scatter blue to your eyes more than they scatter red. Therefore, the sky is blue.

上面段落上，“the sky is blue”的原因位於前一個句子。因此，這些 patterns 被修正為以兩句或三句話做為比對單位來決定 [EVENT] 及 [REASON] 的位置。像在上面的例子中，pattern 即為 “[REASON]. Therefore, [EVENT].”。

此外，有些 patterns 並不僅只代表因果關係，同時也有其他含意的用法，因此會有歧義性存在。舉例來說，“since”這個字就有“由於”和“自從”兩種不同意義。在下面兩句話中：

(1) Since their enemies had been destroyed, they sent back their army.
(2) Since that day, the flowers she had planted had spread all over the hill.

第一句在 “since”後面所接的文字表達了一件事情的原因，而在第二句中“since”則是點出某個事件的起始時間，而不是原因。

為了能夠正確地使用 patterns，我們必須更進一步地瞭解各 patterns 應用上的準確性，且做可能的修正。因果 patterns 準確率的預估方法將於第 3 節中敘述。一旦有了準確率的資訊後，尋找答案時就由準確率較高的 pattern 開始比對起，以最先符合的 pattern 來考慮是否可能為正確答案。

有時，在文字的表現上，[EVENT] 和 [REASON] 之間並沒有很強烈的字面訊息。人類是由上下文以及人類具有的知識得知它們的因果關係。唯一的字面線索是 [EVENT] 和 [REASON] 僅出現在前後文。為了也能捕捉到這種情形，我們加了一個 pattern 為 “[REASON]. [EVENT]. [REASON].”，稱之為 “final pattern”。設定其擁有最低的準確率，成為最後一個被比對的 pattern。

2.2.2 答案擷取步驟

以下為因果 patterns 比對的步驟：

一、 先使用一個詞性標記系統對問句進行詞性標記。我們使用的是 QTAG 3.1[1]。

二、 去除掉問句中的 “why”，餘下的字做為比較相似度時的關鍵字，每個字的權重依其詞性而定。名詞、動詞的權重為 5，形容詞、副詞、數詞、符號或公式的權重為 4，連接詞、冠詞、介系詞及不定詞中的 “to” 的權重為 1，其餘詞性的關鍵字權重為 2。

三、 利用 Porter Stemmer 對所有的關鍵字進行字根還原，由字根還原得到的字稱為 “字根關鍵字”。這些字的權重為原始關鍵字權重的一半。

四、 當文件中的句子符合某因果 pattern 時，切出文句中對應 [EVENT] 的部分。

五、 計算與問句之間的相似度。針對問句中的每一個關鍵字，如果出現在 [EVENT] 部份，則加此關鍵字的權重於相似度的分數中。若是僅為字根關鍵字，則加上字根關鍵字的權重。若未出現則不加分。[EVENT] 和問句的相似度即為所有關鍵字所貢獻之分數總和。

六、 最後，此句子可能為正確答案的分數為：(所符合 pattern 的權重)×([EVENT] 與問句的相似度)。其中 pattern 的權重定義為：0.5 + (0.5×pattern 準確率)。在此定義 “final pattern” 的準確率為 0。

問答系統依照上面的流程，對每個相關文件中的句子做比對並計算可能成為答案的分數。最後依照步驟六所得分數排序，將分數高的答案回覆給使用者。

[1] http://web.bham.ac.uk/O.Mason/software/tagger/

3. 句型歧義性分析

為了觀察各 patterns 的正確性，需要一個較大規模的測試集來測試。測試集中需包含 patterns 的出現，並標示出“原因”所在的位置。然而目前並沒有這樣的測試集存在。底下我們利用兩種方法來求得各因果 patterns 的準確率。

3.1 Penn Treebank 之 PRP 標記

Penn Treebank (Marcus, Santorini, and Marcinkiewicz, 1993) 裡有一個功能標記是“PRP”，用來標示該詞組帶有“目的”或是“理由”的角色。例如，底下這句話的剖析樹中：

> Chevron had to shut down a crude-oil pipeline in the Bay area to check for leaks.
> ((S (NP-SBJ-1 Chevron) (VP had (S (NP-SBJ *-1) (VP to (VP shut (PRT down) (NP (NP a crude-oil pipeline) (PP-LOC in (NP the Bay area) (S-***PRP*** (NP-SBJ *-1) (VP to (VP check (PP-CLR for (NP leaks))))))) .))

在此句中，“to check for leaks”就是“shut down a crude-oil pipeline”的原因。

Penn Treebank 中有 9,613 個含有 PRP 標記的句子。有些句子含有一以上的 PRP 標記，故 PRP 出現總數為 10,720 次。

然而有些同樣表達因果關係的情形，卻沒有被標上 PRP 標記。Penn Treebank 只標示帶有原因及目的角色的附屬子句或是介系詞片語，若是整個句子代表理由時 (例如 “Because he was young.” “Therefore, he will not attend.”)，就不會帶有 PRP 標記。如此一來，不是每個與因果 patterns 相符的句子都會標上 PRP 標記。

因此，我們改利用 PRP 標記這項資訊來估算 patterns 關鍵字的準確率。Patterns 關鍵字就是 patterns 中 [EVENT] 和 [REASON] 之外的部份。之後以 patterns 中關鍵字的準確率來做為 patterns 的準確率。

抽取出 Penn Treebank 中所有被標上 PRP 標記的詞組，統計出現在詞組開頭的因果 patterns 關鍵字個數。再計算每個 patterns 關鍵字出現在整個 Penn Treebank 中的次數，就可得出當一個 pattern 關鍵字出現時，它會被標為 PRP 的比例是多

少。我們以此比例視為 patterns 關鍵字的準確率。統計結果如表 1 所示。

由表 1 的統計資料中，我們可以看到某些 patterns 關鍵字準確率很高，像是“because”、“in order to”等等。然而除了包含“because”出現次數很高以外，其他 patterns 關鍵字的出現次數並不高。相反的，“to”、“for”以及“since”這幾個在 PRP 詞組中常出現的 patterns 關鍵字，卻因為準確率不夠高而無法直接用來判斷因果關係。這是因為這些 patterns 關鍵字有歧義性的原因，以下我們針對這三個 patterns 關鍵字再做進一步的分析。

表 1. 各因果 patterns 準確率

Patterns 關鍵字	PRP 個數	總次數	準確率
‘cause	9	9	1
because	3750	3861	0.971
because of	641	661	0.97
in order to/for/that	108	116	0.931
so as to	6	7	0.857
as a result of	61	85	0.718
on account of	5	7	0.714
as a result	39	86	0.453
so that	180	416	0.432
so as	4	10	0.4
due to	40	110	0.364
cause	82	249	0.329
since	310	1169	0.265
why	133	824	0.161
to	3318	55272	0.06
for	731	18075	0.044
so	173	8768	0.02
as	46	10481	0.004
that	16	36897	0.0004
for \w+ing	66	823	0.08

(1) “to + 原形動詞”

出現“to”而表示因果關係的情形中，通常是先以完整句子描述事件，再接以“to”開頭的不定詞子句說明原因，因此“to”之後必定接原形動

詞。我們觀察 Penn Treebank 中所有 “w_1 to w_2” 或是 “To w_2” 的情形，若是限制 w_2 為原形動詞而 w_1 不為動詞時才判定為因果關係，準確率可由 6%提升至 15.1%。

(2) “for”

以 “for” 來表達因果關係時，用法如同 “because”，是用以連接描述原因和結果的兩個句子的。因此 “for” 多半出現在句首，或是在句中但以逗號與前句隔開。我們於是將 pattern 改為 “For …” (限定在句首) 以及 “…, for …” (在句中前接逗號)。然而 Penn Treebank 中並不會將所有表示因果關係的情形都標上 PRP (此點將在第 3.2 節中討論)，因此我們改以人工判斷的方式，隨機取出的 25 句符合本 patterns 的句子觀察，準確率是 7/25=28%。

(3) “since”

為了排除以 “since” 描述時間起始點的情形，只要 “since” 之後接有年份、月份，或是“year”、“day” 等等之類代表時間的關鍵字，以及 “ever since”、“since then” 的話，都不視為因果關係。如此一來，準確率提升至 38.4%。如果更進一步限制 “since” 只能出現在句首或是逗號之後，準確率可達 64%，但是這樣只能判斷出一半以 “since” 為起首的 PRP 詞組。為了彌補召回率的不足，除了 “since” 在句首或逗號後並做時間詞判斷的 patterns 外，我們也保留了完全不做判斷的 “since” pattern (準確率 26.5%)。

3.2 人工觀察

因為 Penn Treebank 只標示帶有原因及目的角色的附屬子句或是介系詞片語，有些 patterns 關鍵字就不會被標上 PRP，像是 “therefore”。要得到這些 patterns 關鍵字的準確率，我們改以人工的方式來進行。我們隨機至 Penn Treebank 中抽出至多 25 句出現 patterns 關鍵字的句子，再以人工判定是否為因果關係。所得到的觀察結果列在表 2。

在表 2 中，“due to” 會加上 “不接原形動詞” 這一項條件，是因為我們觀察發現，當 “due to” 意義為 “預定要” 的時候，其後會接動詞的原形，而這就不是描述因果關係的句子了。

表 2. 由人工觀察之 patterns 準確率 (%)

Patterns	**人工判定表因果關係個數**	**準確率**
therefore	25/25	100
Thus	20/25	80
hence	21/21	100
So … 或 …, so …	15/25	60
due to (不接原形動詞)	25/25	100
as a result	25/25	100

此外，以 “so” 起始的子句會被標上 PRP 者，都是等同於 “so that” 的情形。為了計估 “so” 當連接詞、表示 “所以” 的比例，我們也以人工的方式觀察了 “so” 出現在句首或逗號之後的情況。這樣的 patterns 準確率可由 2% 提升到 60%。準確率仍然不夠高的原因是因為 “so” 的用法實在太廣，無法單以字面就能決定 “so” 的真正角色。

3.3 因果 patterns 的比對

由 Penn Treebank 及人工判斷得到各準確率之後，我們就可以依照 patterns 的準確率排序。要尋找文件中含有因果關係的句子時，會由準確率較高的 pattern 開始比對起。如此一來，如果同一段文字中出現兩個以上符合因果 patterns 的部份，將會優先判斷準確率較高的部份是否為可能答案。

由於有些 patterns 的準確率來自小量測試集的人工評估，我們於是將由 Penn Treebank 評估所得高準確率 patterns 的優先順序往前挪。先比對在 Penn Treebank 中準確率大於 80%的 patterns，再依照其餘 patterns 的準確率由高到低分別比對。

4. 實驗與討論

我們根據第 2 節及第 3 節所得到的 patterns 及其準確率，實作了一個針對 “why

問句”回答的問答系統。本節描述我們如何進行效能評估的實驗，並且也與另外兩個以網際網路為基礎的問答系統 AnswerBus (http://www.answerbus.com/) 及 LCC (Language Computer Company, http://www.languagecomputer.com/) 做比較。

4.1 實驗資料

在 TREC QA-Tracks 歷屆的題目中，只有 8 題是屬於 “why 問句”。為了擴大實驗規模，我們至 AskJeeves (http://www.ask.com/) 網站蒐集之前使用者曾經提過的問題。在十萬多個問句中，僅有 87 句是 “why 問句”。另外還有 50 題是由 AnswerBus (http://www.answerbus.com/) 網站中 “Sample questions from Excite” 網頁內容所整理出來的，總共得到 145 個問句可進行實驗。

受限於人力的不足，我們先以其中的 50 題來進行系統的效能評估。先去掉這 145 題中意義重複的問句，以及某些沒有標準答案或是詢問建議、需結合使用者背景資料才能回答的題目，例如 “Why is my monitor only showing 16 colors ?”、“Why should I go to college?” 等。之後隨機選取 42 題，連同來自 TREC 的 8 題共 50 題 “why 問句” 來進行評估。

檢索相關文件以備尋找答案時，我們會在由問句所建構成的原始查詢中，分別加入 “reason”、“why” 和 “because” 這三個特殊字，成為新的查詢，分別利用 Google 找出最相關的前 10 篇共 30 篇文件備查。之後再利用原始查詢檢索出不與這 30 篇重複的 200 篇相關文件。因果 patterns 比對及答案擷取就在這 230 篇文件中進行。

4.2 答案評估

AnswerBus 和 LCC 在給使用者答案時，都是以句子為單位回覆。為了要和這兩個系統比較，我們的系統也以完整的句子做為給答單位。但如果比對成功的 pattern 會跨過句子邊界，則系統會將所有此 pattern 所涵蓋的句子都抓出來做為一個答案。

測試時，將第 4.1 節選出的 50 題問句分別向這三個系統提出。由每個系統

的回答中，各題都挑出前五名的答案以進行評估。AnswerBus 和 LCC 常常只回覆了 5 個以下的答案，AnswerBus 平均回答 4 個，LCC 平均回答 4.88 個。我們的系統則是一定提供前五名比對到的答案。

標定各答案是否正確是由人工來進行。我們將答案打散，讓評估者無從得知各答案是由哪個系統所回答的。每一題都給三個評估者評估，以多數人的意見決定是否為正確答案。得到的評估結果列在表 3 之中。

表 3. 問答系統回答 "why 問句" 的效能評估

系統	正解在第一名	正解在前五名	MRR
AnswerBus	15	31	0.429
LCC	8	20	0.229
我們的系統	26	39	0.623

表 3 中第四欄的 MRR (Mean Reciprocal of Rank) 是在 TREC QA 比賽中所用的評比標準 (Voorhees, 1999)。其計算方法為，針對一問句，若系統所給出第一名的答案即為正確答案的話，得一分。若第二名的答案才正確的話，得 1/2 分。若第三名才正確的話，得 1/3。也就是以正解所在的最高名次的倒數為得分，最後的 MRR 值為每一題所得分數的平均。

由表 3 我們可以看到，我們的系統利用因果 patterns 的幫助，系統效能優於其他兩個線上系統。

4.3 分析

4.3.1 增加特殊字查詢相關文件的幫助

如第 2.1 節所提，在檢索相關文件時，系統會在查詢中加入 "reason"、"why" 和 "because" 等特殊字，以期找回的相關文件中能含有因果關係的文句。但是這個動作的幫助有多少？

首先我們觀察加不加入特殊字，對於檢索所得相關文件的影響。分別以 145 個問句的原始查詢字串與加入特殊字查詢字串做檢索，原始查詢 (包括刪去查詢字以求足量相關文件的動作) 取前 200 名。特殊字查詢字串所得到的 145×3×10 =

4,350 篇中，有 1,216 篇完全未出現在以原始字串查詢的結果中，顯示特殊字確實可以幫助抓到更多可能含有因果關係描述的文件。

而在系統評估時，我們針對各正確答案的來源文件做了統計，結果在表 4。其中 Rn、Wn、Bn 分別表示加入 “reason”、“why”、“because” 查詢所得的第 n 篇相關文件，Nn 則表示利用原始查詢所得、但排去已由特殊字查詢檢索出文件的第 n 篇相關文件。

表 4. 答案與文件來源的關係

範圍	總數	正解	範圍	總數	正解
R1-R10	28	13	N91-N100	3	1
W1-W10	47	20	N101-N110	8	3
B1-B10	45	25	N111-N120	6	1
N1-N10	14	5	N121-N130	7	2
N11-N20	3	1	N131-N140	8	3
N21-N30	3	2	N141-N150	12	4
N31-N40	8	4	N151-N160	7	2
N41-N50	11	5	N161-N170	8	3
N51-N60	9	5	N171-N180	2	0
N61-N70	6	4	N181-N190	3	2
N71-N80	6	2	N191-N200	1	1
N81-N90	5	2	總計	250	110

針對實驗的 50 個問句，我們的系統提出了 250 個可能答案，有 110 個被評估為正確。由表 4 可知，有一半以上 (13+20+25=58) 的正確答案來自加入特殊字查詢所得的相關文件中。顯示這利用特殊字所查得的 30 篇相關文件，提供因果關係的資訊遠多於原始查詢的相關文件。

此外，有趣的是，是否找到正確答案，與排去特殊字的原始查詢相關文件的名次不很相關，這和問答系統研究中的一個性質相吻合：正確答案不一定出現在所謂「最相關」的文件中。

4.3.2　各因果 patterns 的答題正確率

表 5 中列出了各 patterns 提供答案的個數，以及被評估為正確答案的個數 (以關

鍵字做為分類統計)。由表 5 中可以看到，有許多 patterns 在這次評估中並未被用到，“because” 和 “because of” 則佔了一半以上。這表示 “because” 確實是一個很常用的句型，它們在我們整理出的 patterns 排名中又很前面，所以容易先被比對到。

然而在表 5 中 “because” 各 patterns 答題正確率卻只有 50%上下。經過觀察，發現這並不是 patterns 的錯誤。許多 patterns 比對的錯誤都來自 [EVENT] 部份與問句比對這階段。由於我們的系統在計算相似度時是以關鍵字比對為主，會發生比對錯誤的情形。比方說，問句為 “Why is the sky blue?”，而有一個句子是 “Blue ocean is beautiful because…”，這時問句和 [EVENT] 部份有不小的相似度，造成答案抽取的錯誤。

同樣的情形也會造成正確答案未被找到的狀況。當含有正確答案且符合因果 patterns 的句子中，[EVENT] 部份使用了與問句語意相同但字面差異很大的說法時，這個句子就無法比對成功，答案也就無法被找到。由此可知，短文句間的相似度比對及語意比對是影響問答系統效能的重要因素。

其次的錯誤就來自於 patterns 關鍵字本身的歧義性，如同我們在第 3 節中所討論的一樣。在擷取答案的過程中，仍會找到並非因果關係描述的文句。

“Why [EVENT]? [REASON].” 這個 pattern 有另一個錯誤情形。在文章中，會以 “Why…?” 提詞的敘述方式，其後可能會以三四句甚至一整段文字來解釋這個原因。而我們的步驟至多只抽取往後一個句子，因此會因答案不完整而被判斷錯誤。

表 5. 答案與 patterns 之關係

Pattern 關鍵字	**總數**	**正解**	**正確率**
because of	24	12	50.0%
because	102	56	54.9%
‘cause	0	0	-
In order to	0	0	-
so as to	0	0	-
as a result of	2	0	0.0%
as a result	0	0	-
therefore	8	3	37.5%
hence	3	3	100.0%
due to	8	5	62.5%
thus	2	1	50.0%
on account of	0	0	-
that is why	2	2	100.0%
for this reason	0	0	-
Why ?	31	13	41.9%
reason that	7	1	14.2%
so as	0	0	-
so that	1	1	100.0%
since（經判別）	6	3	50.0%
So/,so	14	3	21.4%
For/,for	2	1	50.0%
to-V（經判別）	3	1	33.3%
since（未判別）	0	0	-
to-V（未判別）	8	1	12.5%
for	11	2	18.2%
so	0	0	-
as	2	1	50.0%
“final pattern”	13	0	0.0%
總和	250	110	44.0%

5. 結論與未來工作

本論文建構了一個以網際網路為基礎的問答系統，自動回答 “why” 類型的問句。我們使用了搜尋引擎檢索出相關的網頁文件以用來尋找可能答案。接著利用描述因果關係的 patterns，評估 patterns 中 [EVENT] 部份與問句本身的相似度。最後以問句相似度和符合之 pattern 權重的乘積做為這個可能答案的分數，將分數較高的答案優先回覆給使用者。

設定 patterns 權重時，我們以 Penn Treebank 及人工評估的方式，得到各 patterns 關鍵字的準確率，準確率越高的 pattern 有越高的權重。

進行效能評估時，我們以另兩個以網際網路為基礎的問答系統 (AnswerBus 和 LCC) 來與我們的系統做比較，發現我們系統的效能優於另外兩個線上系統。以 TREC 中 QA 評比的 MRR 值來評估，AnswerBus、LCC 和我們系統的 MRR 值分別為 0.429、0.229 和 0.623。

在未來的工作中，[EVENT] 與問句相似度的比較會是一個重要的研究議題。除了關鍵字與字根比對外，還可試著加入語法或語意上的比較，或者使用 WordNet 來進行關鍵字的擴充，甚至是處理代名詞指涉問題等來加強相似度比較的正確性。不過如果是基於網際網路的問答系統，必須考慮反應時間的長短，所以也不適宜使用太複雜的相似度比較方法。

而如何修改 patterns，或增加比對上的限制，藉以提升 patterns 的準確率，是未來研究的另一個重點。此外，當 patterns 涵蓋兩句以上的段落時，如何確定答案的邊界就是一個值得研究的課題。

參考文獻

Anand, Pranav; Breck, Eric; Brown, Brianne; Light, Marc; Mann, Gideon; Riloff, Ellen; Rooth, Mats and Thelen, Michael (2000) “Fun with Reading Comprehension,” Final report, Reading Comprehension group, Johns Hopkins Center for Language and Speech Processing Summer Workshop 2000. Johns Hopkins University, Baltimore MD. [Online] Available URL:

http://www.clsp.jhu.edu/ws2000/groups/reading/ WS00_readcomp_final_rpt.pdf

Girju, Roxana and Moldovan, Dan (2002) "Mining Answers for Causation Questions," *Proceedings of the American Association for Artificial Intelligence (AAAI) - Spring Symposium*, Stanford University, California, USA, March 2002.

Harabagiu, Sanda; Moldovan, Dan; Pasca, Marius; Mihalcea, Rada; Surdeanu, Mihai; Bunescu, Razvan; Girju, Roxana; Rus, Vasile and Morarescu, Paul (2000a) "FALCON: Boosting Knowledge for Answer Engines," *Proceedings of the Text REtrieval Conference (TREC-9)*, Gaithersburg, Maryland, USA, November 2000, pp. 479-488.

Lin, Jimmy (2002) "The Web as a Resource for Question Answering: Perspective and Challenges," *Proceedings of the third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain, May 2002.

Marcus, Mitchell P.; Santorini, Beatrice and Marcinkiewicz, Mary Ann (1993) "Building a large annotated corpus of English: the Penn Treebank," *Computational Linguistics*, Vol. 19, No. 2, June 1993. pp. 313-330.

Moldovan, D. and Rus, V. (2001) "Logic Form Transformation of WordNet and its Applicability to Question Answering," *Proceedings of the ACL 2001 Conference*, Toulouse France, July 2001, pp. 394-401.

Radev, Dragomir R.; Qi, Hong; Zheng, Zhiping; Blair-Goldensohn, Sasha; Zhang, Zhu; Fan, Weiguo and Prager, John (2001) "Mining the Web for Answer to Natural Language Questions," *Proceedings of the ACM CIKM-2001*: *Tenth International Conference on Information and Knowledge Management*, Atlanta, Georgia, USA, November 2001.

Radev, Dragomir; Fan, Weiguo; Qi, Hong; Wu, Harris and Grewal, Amardeep (2002) "Probabilistic Question Answering on the Web," *Proceedings of the eleventh International World Wide Web Conference (WWW 2002)*, Honolulu, Hawaii, USA, May 2002.

Riloff, Ellen and Thelen, Michael (2000) "A Rule-based Question Answering System for Reading Comprehension Tests," *Proceedings of the ANLP/NAACL-2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, Seattle, Washington, USA, May 2000.

Voorhees, E. (1999) "The TREC-8 Question Answering Track Evaluation," *Proceedings of the Text REtrieval Conference (TREC-8)*, Gaithersburg, Maryland, USA, November 1999, pp. 23-37.

Zheng, Zhiping (2002) "AnswerBus Question Answer System," *Proceedings of Human Language Technology Conference (HLT 2002)*, San Diego, California, USA, March 2002.