自動答詢系統與其數位典藏應用

許聞廉

中央研究院資訊科學研究所

2006/09/01

大綱

- 問答系統
  - 簡介
  - ASQA問答系統
    - FAQ問答系統
    - 仿真陳述(Factoid)問答系統
- 問答系統技術對數位典藏的重要性
  - 語言理解與智慧型介面
  - NER, SRL與未來的Metadata

最早的問答系統

- Green於1996年開發了一個名為BASEBALL的系統
- 回答與美國大聯盟有關的問題，如各場比賽的得分、參賽隊伍、地點、日期等資訊
- 該系統以IPL語言撰寫，執行在硬體資源相當有限的IBM 7090主機上
- 僅能處理無子句的簡單問題
- 例子：
  - 使用者鍵入：Did the Tigers play the Red Sex in July?
  - 系統搜尋資料庫後答覆：No

問答系統橫跨多個資訊領域

- 問答系統是個整合性的研究，與其相關的研究領域包括：
  - 資訊檢索
  - 自然語言處理
  - 資訊擷取
  - 機器學習
  - 邏輯推理

問答系統分類

- 以領域知識分類
  - 開放領域、限制領域
- 以語言分類
  - 單語：中文、英文、日文…
  - 跨語：中英、英中…
- 以知識媒體分類
  - 資料庫、FAQ、新聞、Web
- 以問題類型分類
  - 仿真陳述(factoid)、清單、定義

大綱

- 問答系統
  - 簡介
  - ASQA問答系統
    - FAQ問答系統
    - 仿真陳述(Factoid)問答系統
- 問答系統技術對數位典藏的重要性
  - 語言理解與智慧型介面
  - NER, SRL與未來的Metadata

ASQA問答系統

- ASQA: Academia Sinica Question Answering System

- ASQA-FAQ: FAQ問答
- ASQA-FQ: 新聞仿真陳述問答
- ASQA-WEB: Web仿真陳述問答
- BeQA: 生物知識問答

ASQA-FAQ

- 找出知識庫中與使用者問句最相近的常問問題(FAQ)，取出相對應的答案後呈現給使用者
- 例如：
  - 請問如何坐公車到中研院？
  - FAQs
    - FAQ1: 問中研院交通
    - FAQ2: 問單位聯絡方式
    - FAQ3: 問單位員工
    - FAQn: ……
  - FAQ1最合適

http://qa.iis.sinica.edu.tw/

i@SL SINICA QA System

Intelligent QA System

中央研究院智慧型答詢系統

Welcome to 中央研究院智慧型答詢系統，請輸入您的問題，例如：請問中研院資訊所所長的電話

歡迎來自中研院網域：140.109.19.151的使用者，中研院開放下載自然輸入法V8.0中研院專用版

您的問題： 請問中研院院長的電話 Go [系統使用說明]

你可以參考目前最熱門的前10大問題：

如何搭公車到中研院? Go

我的回答：

以下是我的回答：
答案網址：http://www.sinica.edu.tw/as/ytlee/index_c.html
若沒有適合的答案，您可以參考關聯性問題，繼續問問題。

關聯性問題

1. 查詢現任院長的聯絡資料 Go
2. 查詢 現任院長 的相關訊息 Go
3. 問中研院聯絡資料 Go
4. 中研院附近餐飲店訂便當的電話 Go
5. 中研院附近餐飲店外送的電話 Go

簡歷
李院長的話
李院長專訪
研究及著作
教育改革審議委員會
演講與專訪
院長給院內同仁的一封信

ASQA-FQ

- 分析使用者問句後，到文件庫中選取相關性高的文件，經過專有名詞分析處理取得候選答案，排序後送出
- 例如：
  - 問題：誰是美國總統
  - 相關文件1：美國/LOC總統**布希**/PER表示……
  - 相關文件2：**胡錦濤**/PER與美國/LOC總統會面……
  - 候選答案過濾排序：
    1. 布希
    2. 胡錦濤

專有名詞辨識
Named Entity Recognition (NER)

- 人名
- 組織名
- 地點

- 王建民拿下今年球季第十五勝，多項紀錄分居美國大聯盟前茅，不但台灣人喜歡談他，連美國媒體，亦興起「王建民熱」。這也沒啥高深道理，只因為王建民是以棒球場上的成績，贏得各方尊敬。

  王建民的15勝，是亞洲投手在大聯盟單季排名第三，落後於日本野茂英雄的16勝、南韓朴贊浩的18勝。但王建民是大聯盟二年級生，野茂英雄闖蕩多年後，如今棲身於小聯盟；朴贊浩雖仍在大聯盟，不過戰績沒那麼輝煌了，反倒是王建民大有機會改寫亞洲前輩的成績。說王建民能為亞洲投手在大聯盟寫歷史，絕不為過。

語意角色標記
Semantic Role Labeling (SRL)

- Chinese PropBank:

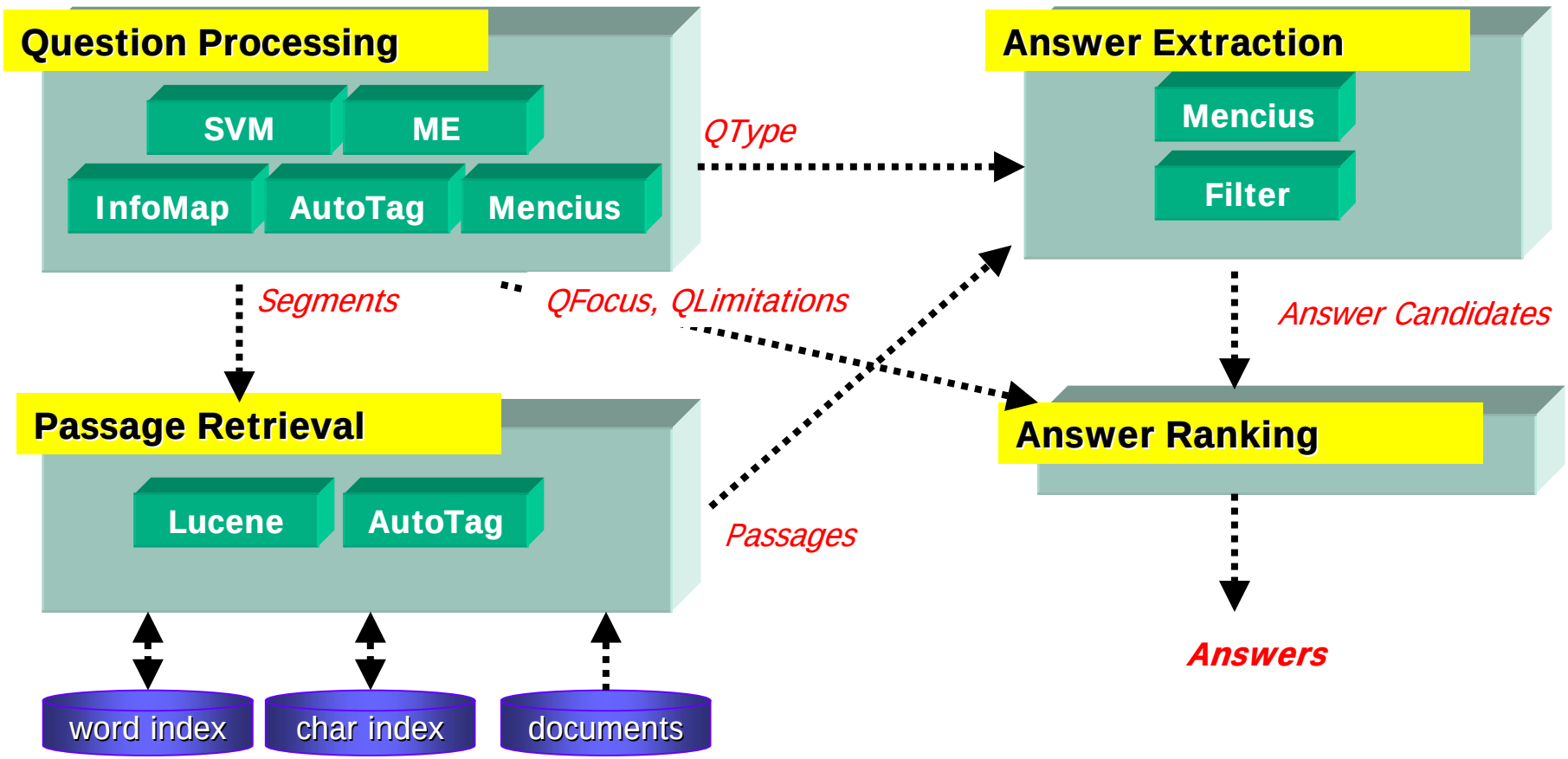
[arg0 我][argm-adv 已經][rel 打][arg1 電話][arg2 給斯恩特]

[arg1 這些算盤][arg0 產業界自己][rel打]的[argm-ext最精]

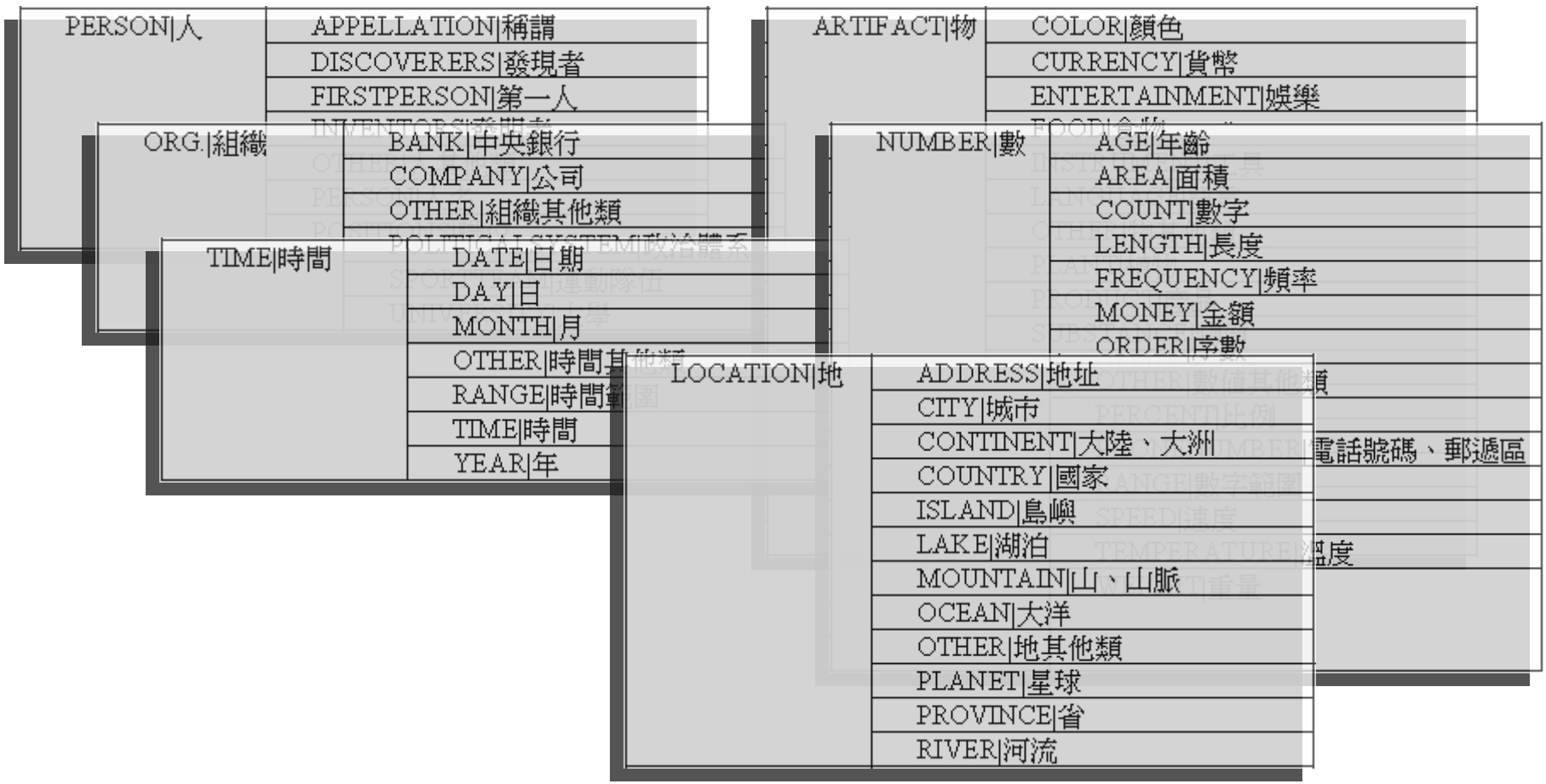
[arg1 鮑薩]被[arg0 泰森的鐵拳][rel打]得[arg2 爬不起來]

[arg0 他][argm-tmp 晚上]則到體育場[rel打][arg1 籃球]

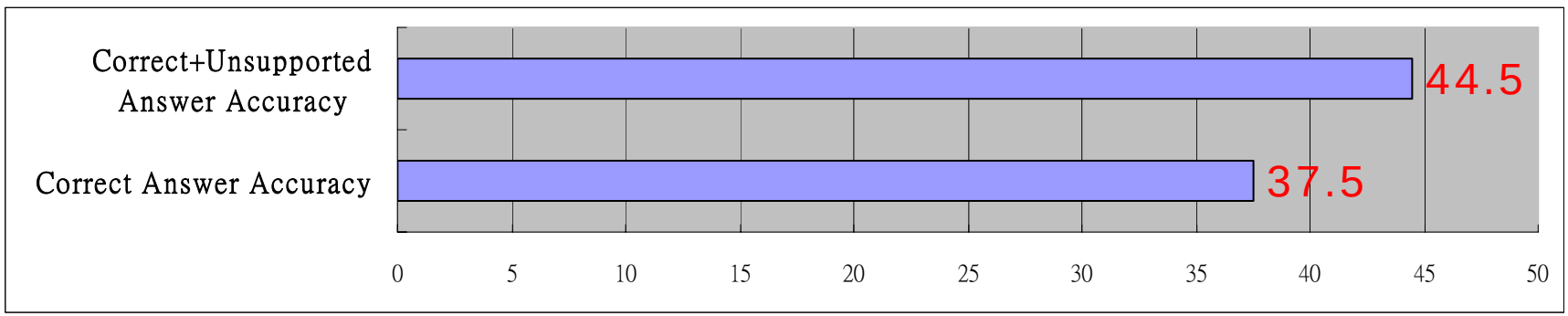
ASQA-FQ架構與流程



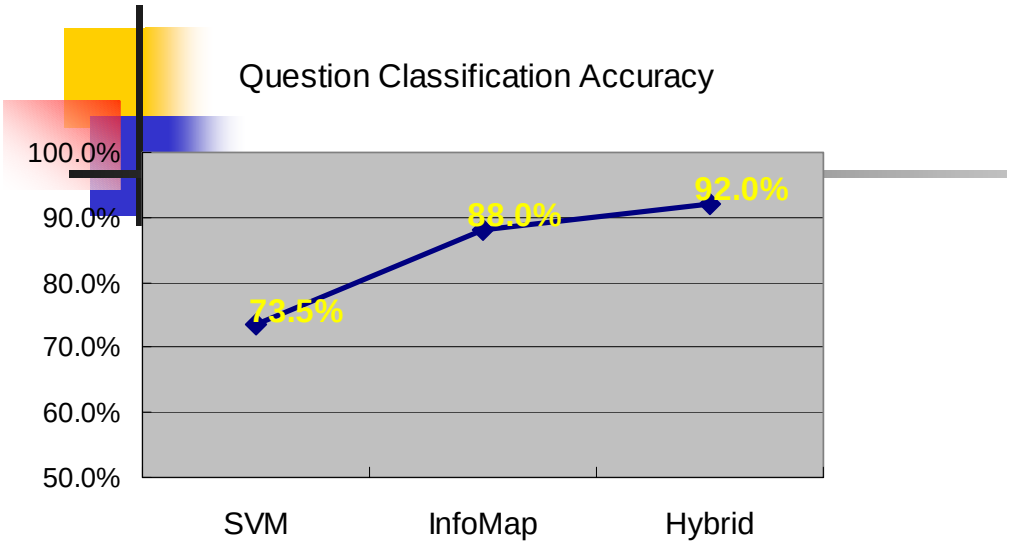
ASQA-FQ問題類型

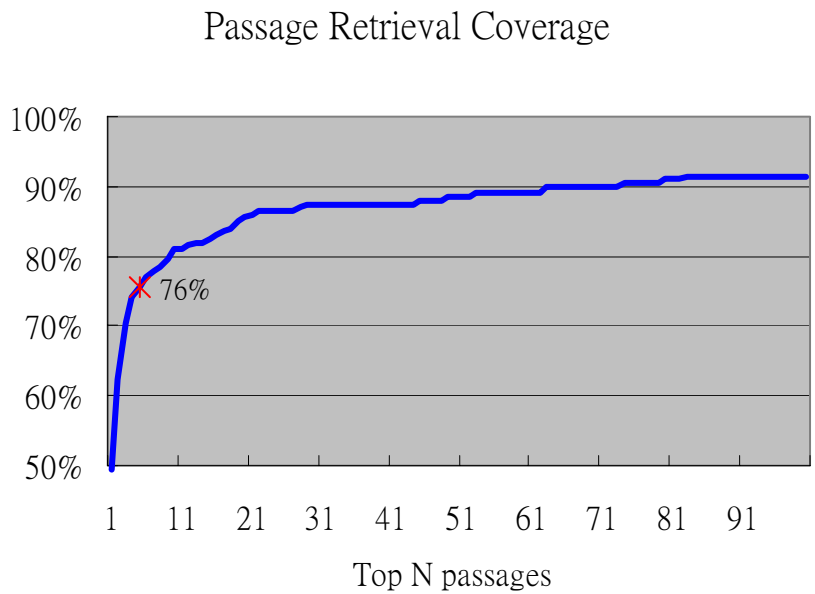

PERSON|人
APPELLATION|稱謂
DISCOVERERS|發現者
FIRSTPERSON|第一人
ARTIFACT|物
COLOR|顏色
CURRENCY|貨幣
ENTERTAINMENT|娛樂
ORG.|組織
BANK|中央銀行
COMPANY|公司
OTHER|組織其他類
NUMBER|數
AGE|年齡
AREA|面積
COUNT|數字
LENGTH|長度
FREQUENCY|頻率
MONEY|金額
TIME|時間
DATE|日期
DAY|日
MONTH|月
YEAR|年
LOCATION|地
ADDRESS|地址
CITY|城市
CONTINENT|大陸、大洲
COUNTRY|國家
ISLAND|島嶼
LAKE|湖泊
MOUNTAIN|山、山脈
OCEAN|大洋
OTHER|地其他類
PLANET|星球
PROVINCE|省
RIVER|河流

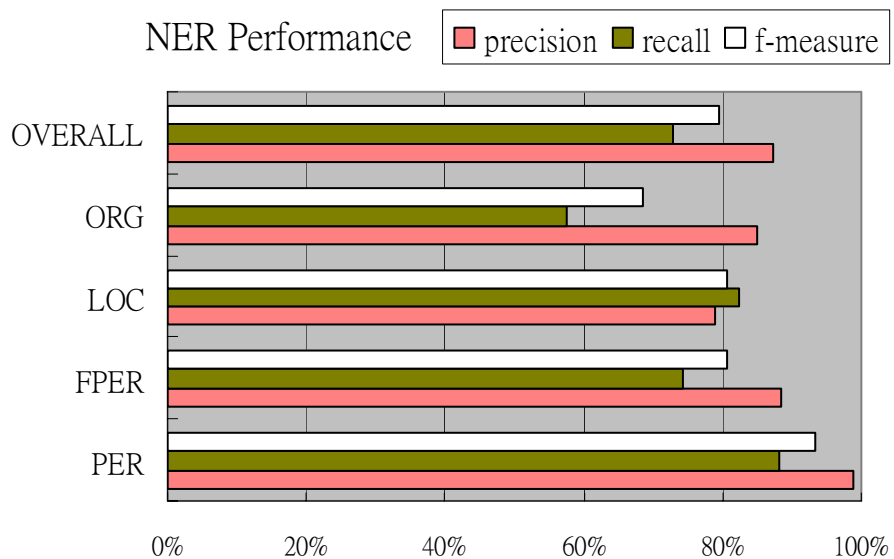
ASQA-FQ正確率

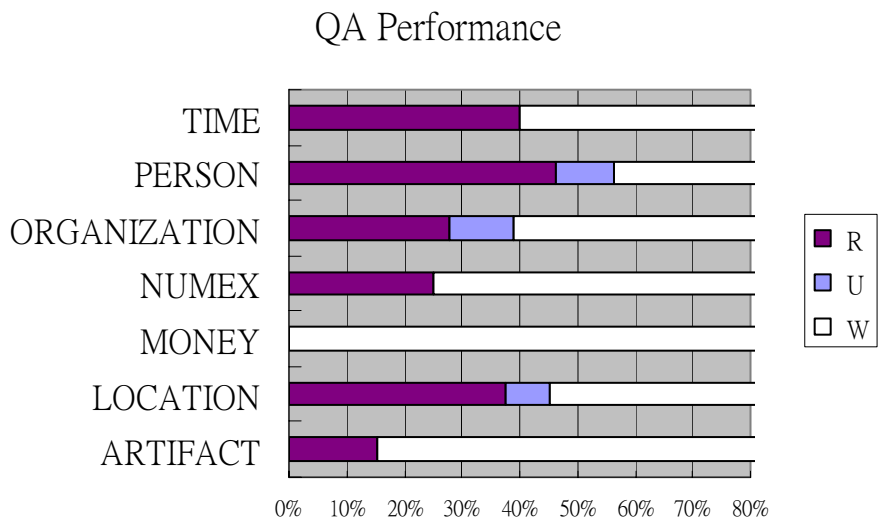

Correct+Unsupported Answer Accuracy
44.5
Correct Answer Accuracy
37.5
0
5
10
15
20
25
30
35
40
45
50

ASQA-FQ分類與模組正確率



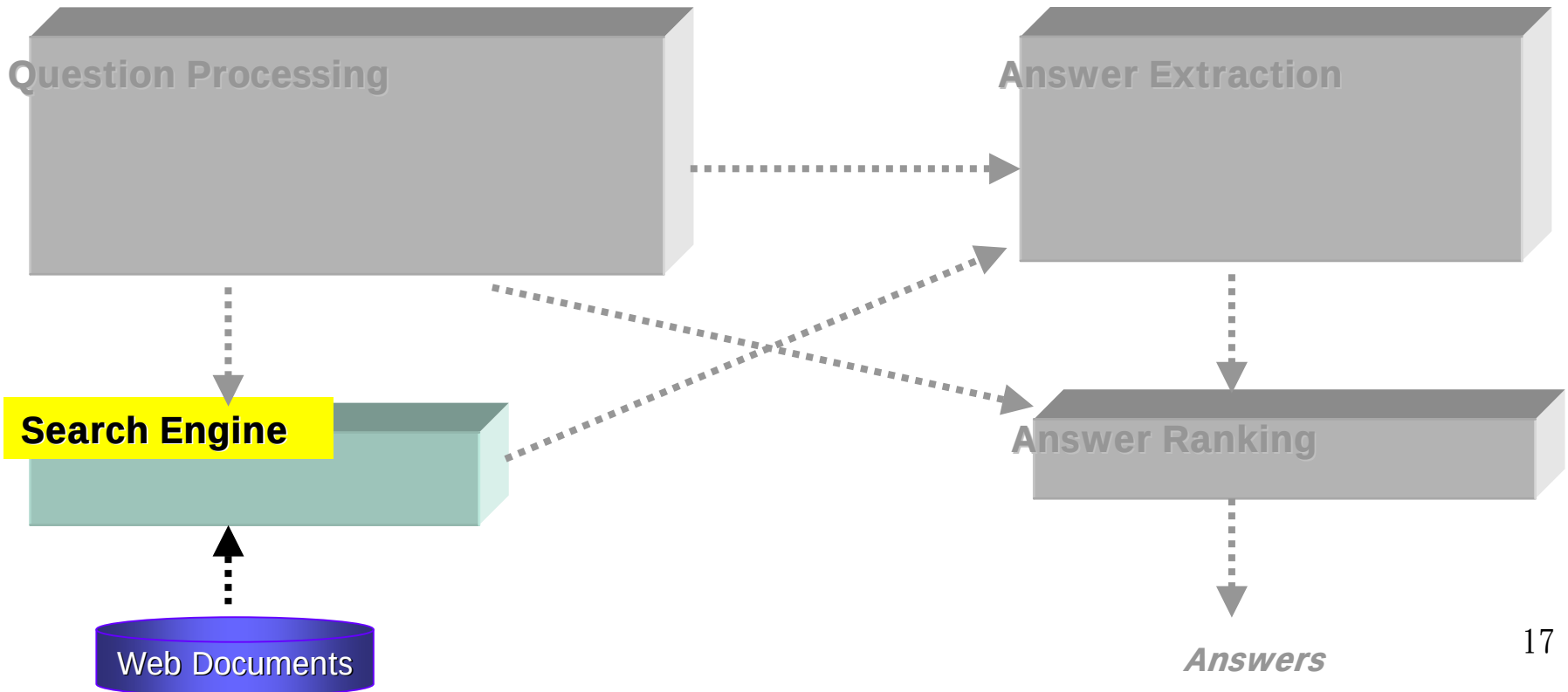






ASQA-WEB

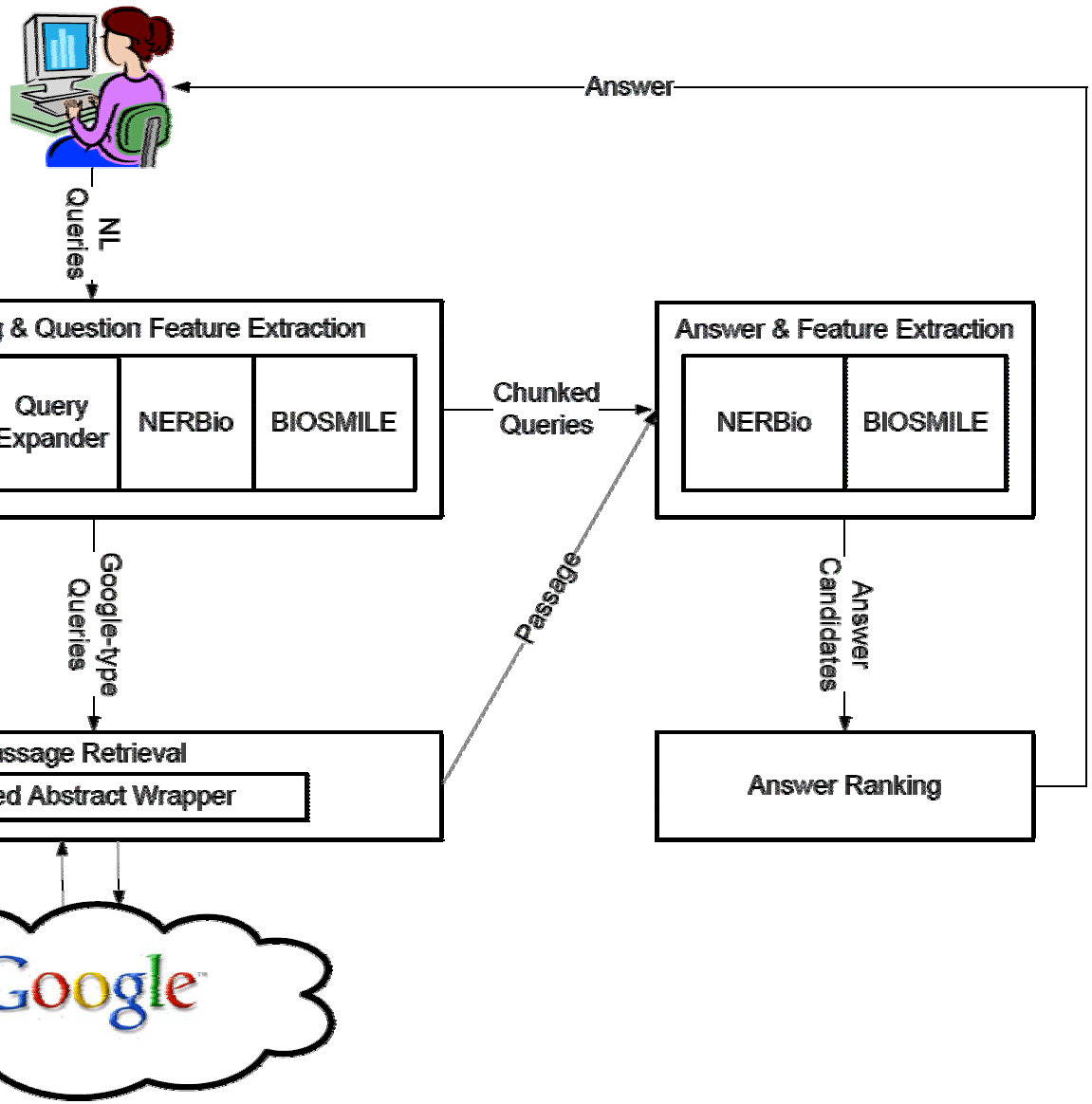
- 與ASQA-FQ處理的問題相同
- 將Passage Retrieval模組置換成搜尋引擎



BeQA
(Biomedical Question Answering system)

- 生物醫學文獻數量的激增，資訊檢索、整合與歸內對生物學家來說日益重要
- 問答系統提供自然語言溝通介面，方便生物學家或一般使用者自然而快速地在大量文章中搜尋出精準的答案
- BeQA採用了專為生物文獻量身打造的NERbio專有名詞辨識(NER)與BIOSMILE語意角色標記(SRL)元件
- 實驗結果顯示，BeQA能達到Top-1 51.9%答題正確率

BeQA流程與架構


Answer
NL Queries
Query Processing & Question Feature Extraction
Question Classifier
GENIA Tagger
Query Expander
NERBio
BIOSMILE
Chunked Queries
Answer & Feature Extraction
NERBio
BIOSMILE
Google-type Queries
Passage
Answer Candidates
Passage Retrieval
PubMed Abstract Wrapper
Answer Ranking
Google

International Competition

- 1st/9 in the NTCIR5 CLQA Chinese Question Answering Contest (44.5%)
- 6th/32 in the TREC Genomic IR Contest (24.5%)
- 5th/21 in the CoNLL-2005 Semantic Role Labeling Contest (74%)
- 1st/13 in the WS CityU closed track of the SIGHAN 2006 Word Segmentation Contest (97.2%)
- 2nd/10 in the WS CKIP closed track of the SIGHAN 2006 Word Segmentation Contest (95.7%)
- 2nd/8 in the NER CityU closed track of the SIGHAN 2006 Named Entity Recognition Contest (88%)

大綱

- 問答系統
  - 簡介
  - ASQA問答系統
    - FAQ問答系統
    - 仿真陳述(Factoid)問答系統
- 問答系統技術在數位典藏的重要性
  - 語言理解與智慧型介面
  - NER, SRL與未來的Metadata

數位典藏需面對的問題與解決方案

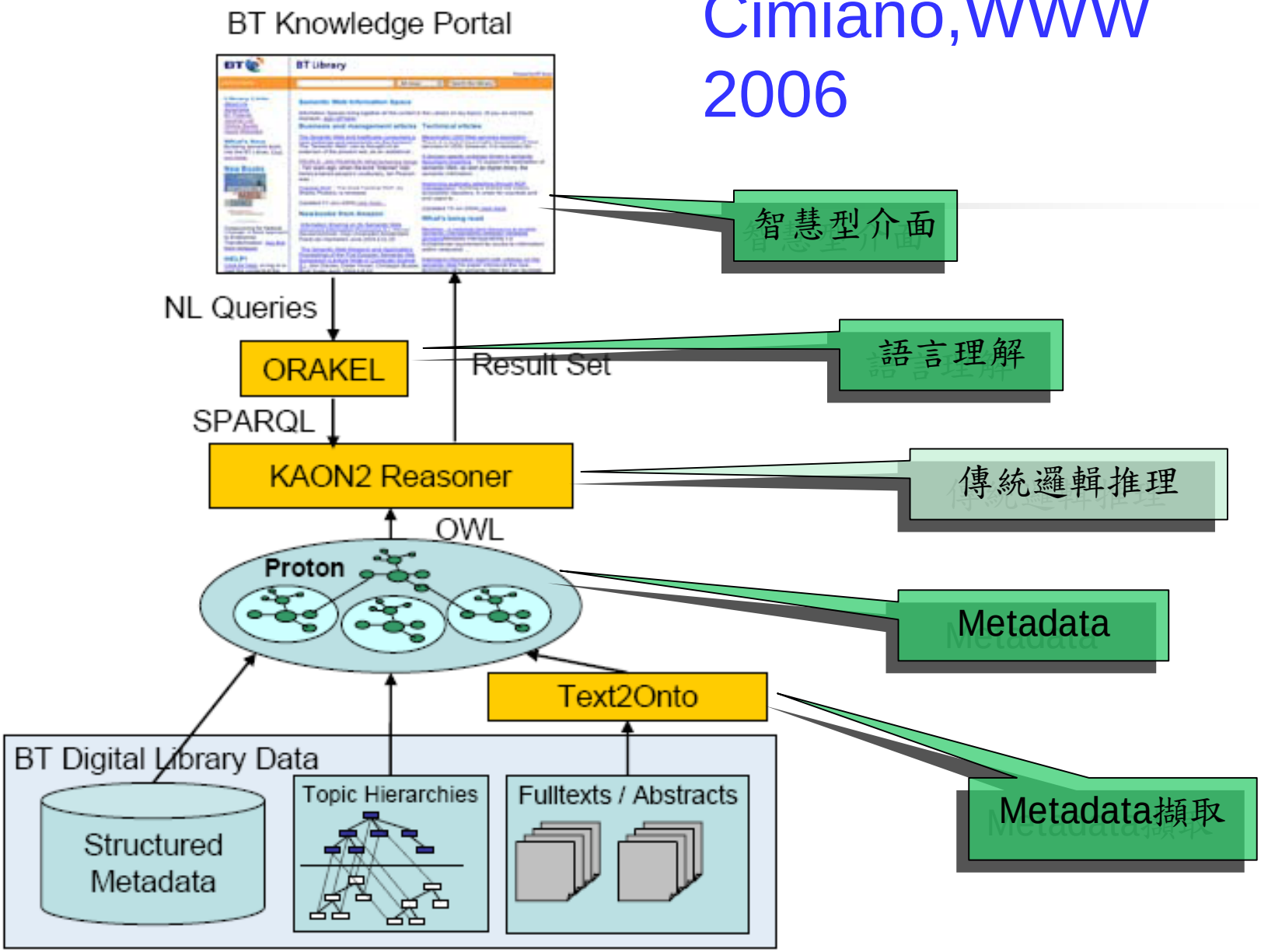
- 圖書館數位化
  - 瀏覽：實體圖書館 → 網路圖書館
  - 服務：圖書館員 → 冰冷電腦
  - 解決方案：以語言理解為基礎的智慧型介面

- 館藏數位化
  - 使用者如何在龐大的數位典藏中找到資料
  - 數位典藏品的長久保存與取得
  - 典藏品之互通、學習與共享
  - 解決方案：Metadata訂定與自動產生

大綱

- 問答系統
  - 簡介
  - ASQA問答系統
    - FAQ問答系統
    - 仿真陳述(Factoid)問答系統
- 問答系統技術對數位典藏的重要性
  - 語言理解與智慧型介面
  - NER, SRL與未來的Metadata

Cimiano,WWW 2006



知識表達
知識的Sensor

- 一個概念有許多的面向，對於這些個別面向瞭解的「總和」可以看成對這個概念理解的程度
  - 如何瞭解某某人？
- 我們以對一個概念的FAQ來看要表達這個概念的方式
- 我們建構了「知識地圖」，將抽象知識表達成其中的「座標」，再由電腦將FAQ準確地對應到這些座標上，來模擬「理解」的達成
- 其中最重要的概念就是context 和 resolution

概念理解

- 電腦概念理解的實務性定義：
  - 假設每個概念都associate一個題庫
  - 電腦針對題庫的random測試能達到標準即算理解
- 這個定義的主要貢獻：利用智慧型的知識管理，可以將『理解』轉化成知識庫的『檢索』
- 概念理解的背後需要一個powerful知識系統
- 可能的應用範圍

中文檢索
自動答詢系統
中英混合字串檢索(容錯)
英文改錯（文法、語意）
文件摘要
文件自動分類

交談系統（dialogue）
email自動過濾、回覆
語音輸入後處理
英漢翻譯
數學、英語教學
知識管理

知識表達系統要能夠同時處理的問題

- Event frames
- 中文人名、機構名、地名、時間
- 未知詞
- 多重語意樹
- global 同義詞，local 同義詞的機制
- 中文詞組（複合詞）、短語、句型
- 英文字尾、時態變化
- 英文詞組、短語、句型
- scenario 理解（以影像辨識的概念來進行語言分析）
- 動態知識表達（理解一個事件之後，並『執行』其內容）
- Script, Story

知識的「座標」(I)

- 以GIS為例，一般抽象的知識要使用怎樣的「座標」才比較容易搜尋？
  - 自然語言的模糊性很高，不同的人可能有不同的解釋方式
    - 「台大」可否當成座標？
      - 台大校本部，台大圖書館，台大醫院
    - 「中央研究院」可能就比較合適
- 目前搜尋引擎普遍使用的是「關鍵詞」
  - 在許多情況，關鍵詞不夠細緻，以致於找到的文件相關性不夠高
- 理想的知識座標是能讓我們從大處著眼，小處著手

知識的「座標」(II)

- 如何在大範圍的知識座標中帶出較小的知識座標？
  - 這牽涉到概念的描述方式
    - 如何描述「台灣大學」，「中央研究院」？
  - 描述概念必須用到自然語言，但是如何處理自然語言意義模糊的問題？
  - 必須「聚焦」，先限定大的context，再由其中限定較小的context。如此，一層層地階段性描述，才容易清楚。

Information Map (資訊地圖)

- 在我們設計的概念理解模式下

  理解 ⟷ InfoMap的正確搜尋

- 要電腦幫人類正確地搜尋資料，必須提供一個良好的guidance。
- 我們提出的新工具—資訊地圖
  - 描述資訊的廣度、深度以及相互關係。在性質上是抽象的，但功能上則類似於地理地圖。
- 資訊地圖將一個概念的相關概念做進一步的分類。也就做一個概念與相關概念的結構描述
- 概念描述必須Context sensitive. 找到了中研院之後，才有可能找到（中研院的）資訊所。

股票的Info MAP

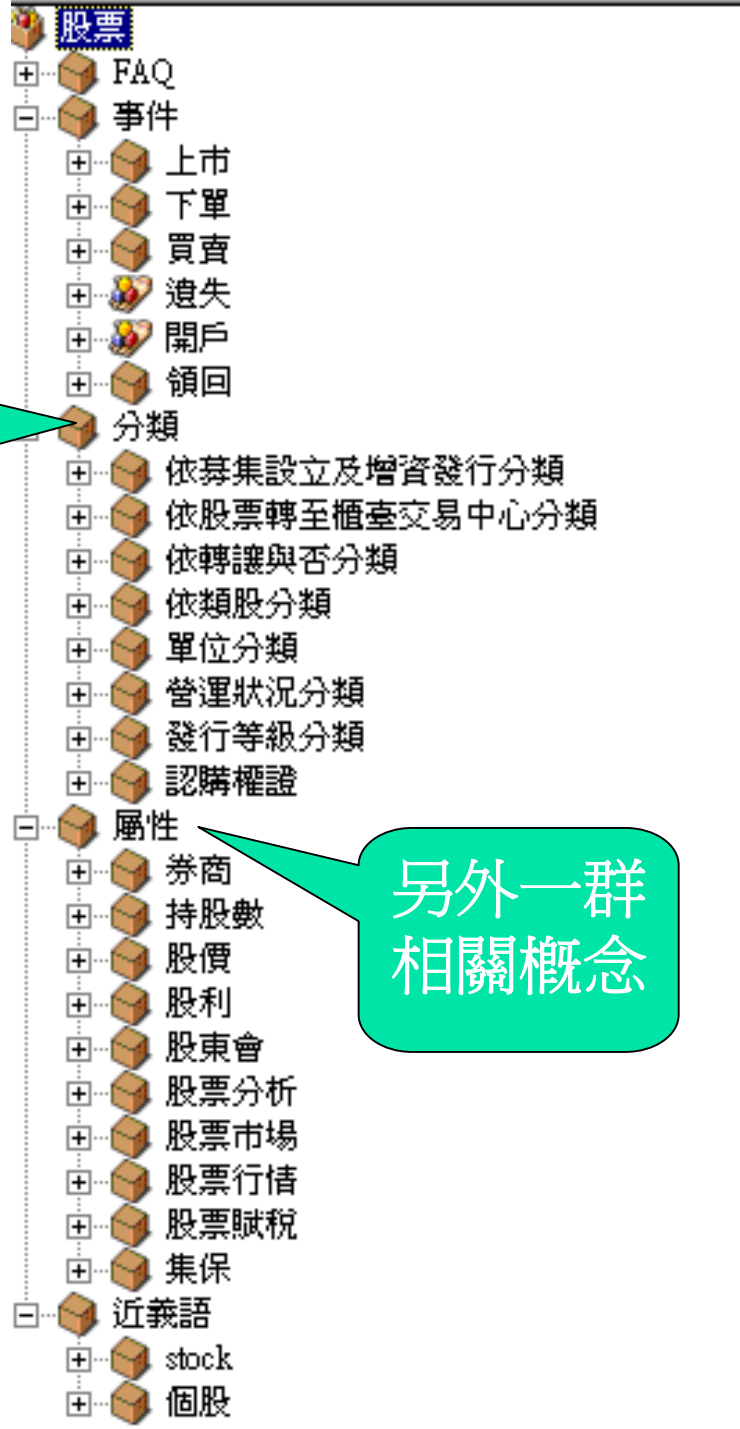
- 事件
  - 上市、下單…
- 分類
  - 依類股分類、單位分類…
- 屬性
  - 券商、股價…

台積電的股票如何買賣？

概念分類 如何 事件

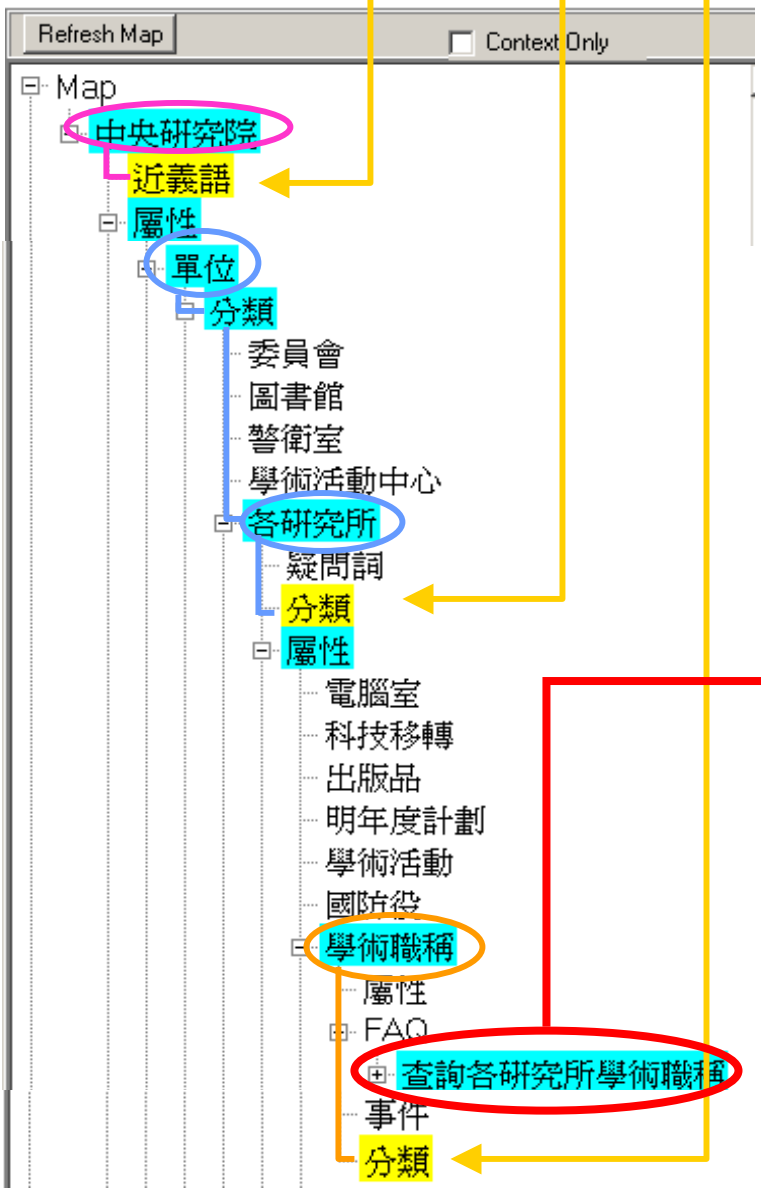
台積電的股價是多少？

概念分類 的 屬性

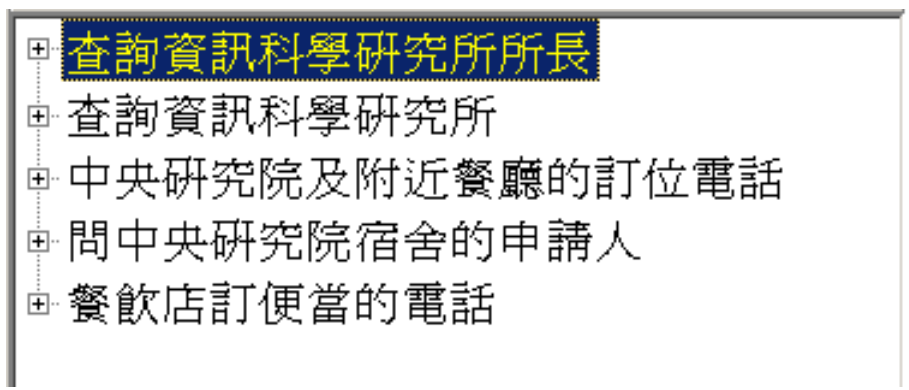


ASQA-FAQ 系統背後的機制

請問中研院資訊所所長的電話



Result of Top 1~5 FAQ



大綱

- 問答系統
  - 簡介
  - ASQA問答系統
    - FAQ問答系統
    - 仿真陳述(Factoid)問答系統
- 問答系統技術對數位典藏的重要性
  - 語言理解與智慧型介面
  - NER, SRL與未來的Metadata

今日的Metadata

- 數位典藏計畫Metadata工作小組FAQ：

- Metadata記載有關資料的元素或屬性(名稱、大小、資料類型等)、有關紀錄或資料結構(長度、欄位、行列等)、或有關資料的資料(位置、關聯性、擁有者等)。Metadata可以包含有關背景、品質和狀況、或資料特徵等的描述性資料。
- 功用
  - 對數位典藏品的提供者、擁有者與管理者來說， Metadata可以協助儲存、控制、管理、散布和交換數位資源。
  - 對數位典藏品的使用者來說， Metadata可以協助搜尋、辨識、選擇、詮釋、獲取和使用數位資源。

Metadata與內容資訊擷取

- 現行Metadata幾乎都是「表面」的資料（如作者、類別、品質等）
- 資訊擷取是問答系統成功的關鍵，也是未來「內容」Metadata能否成功的關鍵
- 「內容」Metadata可能包括如標記某文章所記載「事件」、文章人物間的「關係」等等。
- 「內容」Metadata使得館藏資料搜尋方式更彈性、更容易分析

未來的Metadata

- 雖然現階段Metadata都仍在將傳統館藏Metadata數位化階段，但更豐富的Metadata為未來系統與研究走向

ACE標準 -NE

Type	Subtypes
FAC (Facility)	Airport, Building-Grounds, Path, Plant, Subarea-Facility
GPE (Geo-Political Entity[3])	Continent, County-or-District, GPE-Cluster, Nation, Population-Center, Special, State-or-Province
LOC (Location)	Address, Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body
ORG (Organization)	Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non-Governmental, Religious, Sports
PER (Person)	Group, Indeterminate, Individual
VEH (Vehicle)	Air, Land, Subarea-Vehicle, Underspecified, Water
WEA (Weapon)	Biological, Blunt, Chemical, Exploding, Nuclear, Projectile, Sharp, Shooting, Underspecified

ACE標準 -Relation

Type	Subtype
ART (artifact)	User-Owner-Inventor-Manufacturer
GEN-AFF (Gen-affiliation)	Citizen-Resident-Religion-Ethnicity, Org-Location
METONYMY*	*none*
ORG-AFF (Org-affiliation)	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
PART-WHOLE (part-whole)	Artifact, Geographical, Subsidiary
PER-SOC* (person-social)	Business, Family, Lasting-Personal
PHYS* (physical)	Located, Near

ACE標準 -Event

Types	Subtype
Life	Be-Born, Marry, Divorce, Injure, Die
Movement	Transport
Transaction	Transfer-Ownership, Transfer-Money
Business	Start-Org, Merge-Org, Declare-Bankruptcy, End-Org
Conflict	Attack, Demonstrate
Contact	Meet, Phone-Write
Personnel	Start-Position, End-Position, Nominate, Elect
Justice	Arrest-Jail, Release-Parole, Trial-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon

結語

- 問答系統有各種不同類別
- 問答系統整合多個研究領域
- 問答系統為建構智慧型介面之基礎
- 專有名詞辨識與語意角色標記研究，可提供未來metadata標準與自動標記技術

Q&A