An Integrated Knowledge-based and Machine Learning Approach for Chinese Question Classification

Min-Yuh Day[1,2], Cheng-Wei Lee[1], Shih-Hung Wu[3], Chorng-Shyong Ong[2], Wen-Lian Hsu[1]

[1] Institute of Information Science, Academia Sinica, Taipei

[2] Department of Information Management, National Taiwan University, Taipei

[3] Dept. of Computer Science and Information Engineering, Chaoyang Univ. of Technology, Taichung

myday@iis.sinica.edu.tw

Outline

- Introduction
  - Chinese Question Classification (CQC)
- Proposed Approach
  - Knowledge-based Approach: INFOMAP
  - Machine Learning Approach: SVM
  - Integration of SVM and INFOMAP
    - Hybrid Approach
- Experimental Results and Discussion
- Related Works
- Conclusions

Introduction

- Question Answering
  - TREC QA
  - QA@CLEF
  - NTCIR CLQA
- Chinese Question Classification
  - Goal: accurately classify a Chinese question into a question type and then map it to an expected answer type
  - Chinese Question: 奧運的發源地在哪裡？
    Where is the originating place of the Olympics?
  - Question Type: Q_LOCATION|地
- Question Types
  - Answer extraction and answer filtering
  - Improve the accuracy of the overall question answering system

Introduction

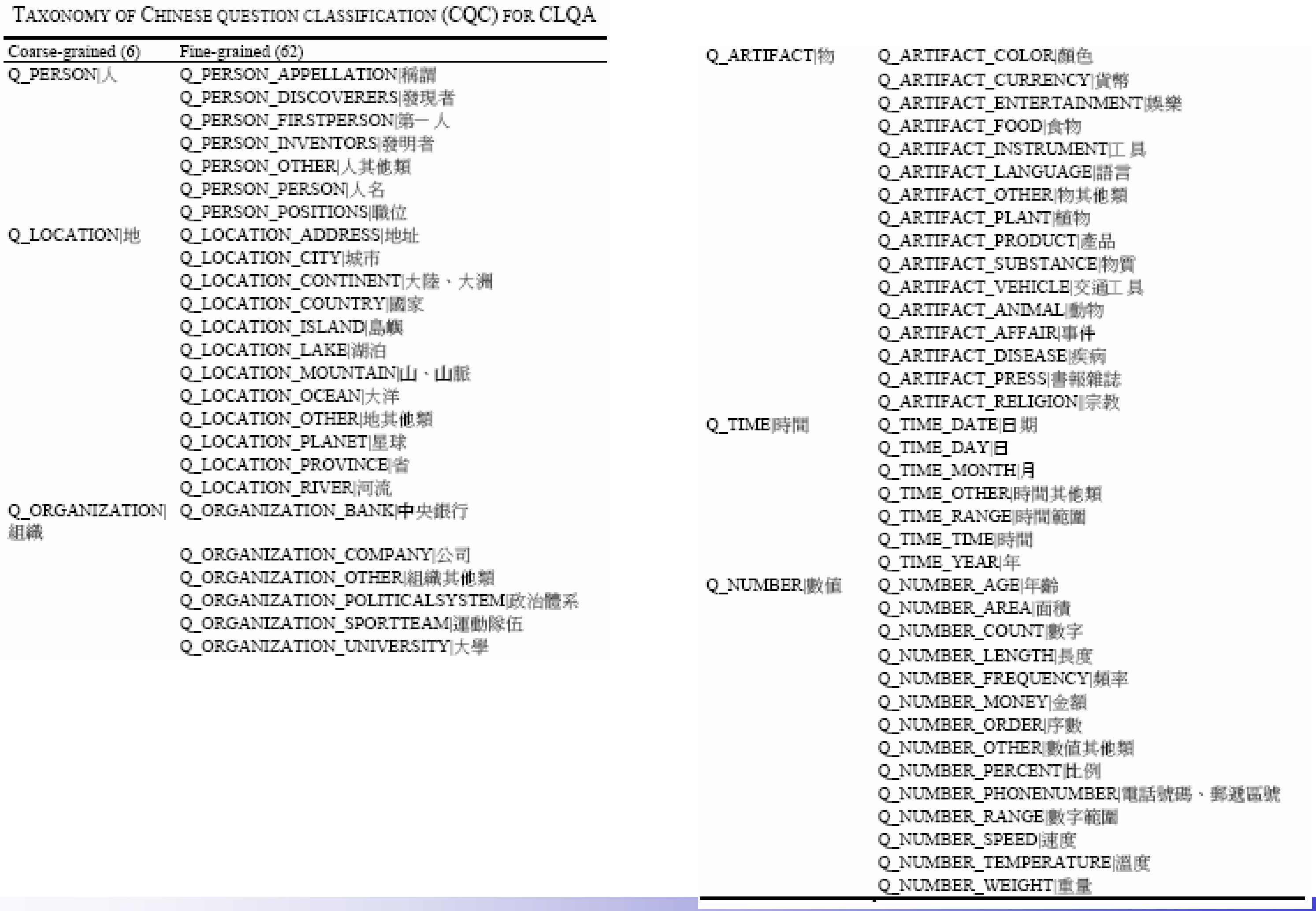
- Problem of Question Classification
  - 36.4% of the errors occur in the question classification module (Moldovan et al., 2003)
- Approaches to Question Classification (QC)
  - Rule-based approaches
  - Statistical approaches

Proposed Approach

- Chinese Question Taxonomy
- Question Type Filter for Expected Answer Type (EAT)
- Knowledge-based Approach: INFOMAP
- Machine Learning Approach: SVM
- Hybrid Approach: Integration of SVM and INFOMAP

Chinese Question Taxonomy

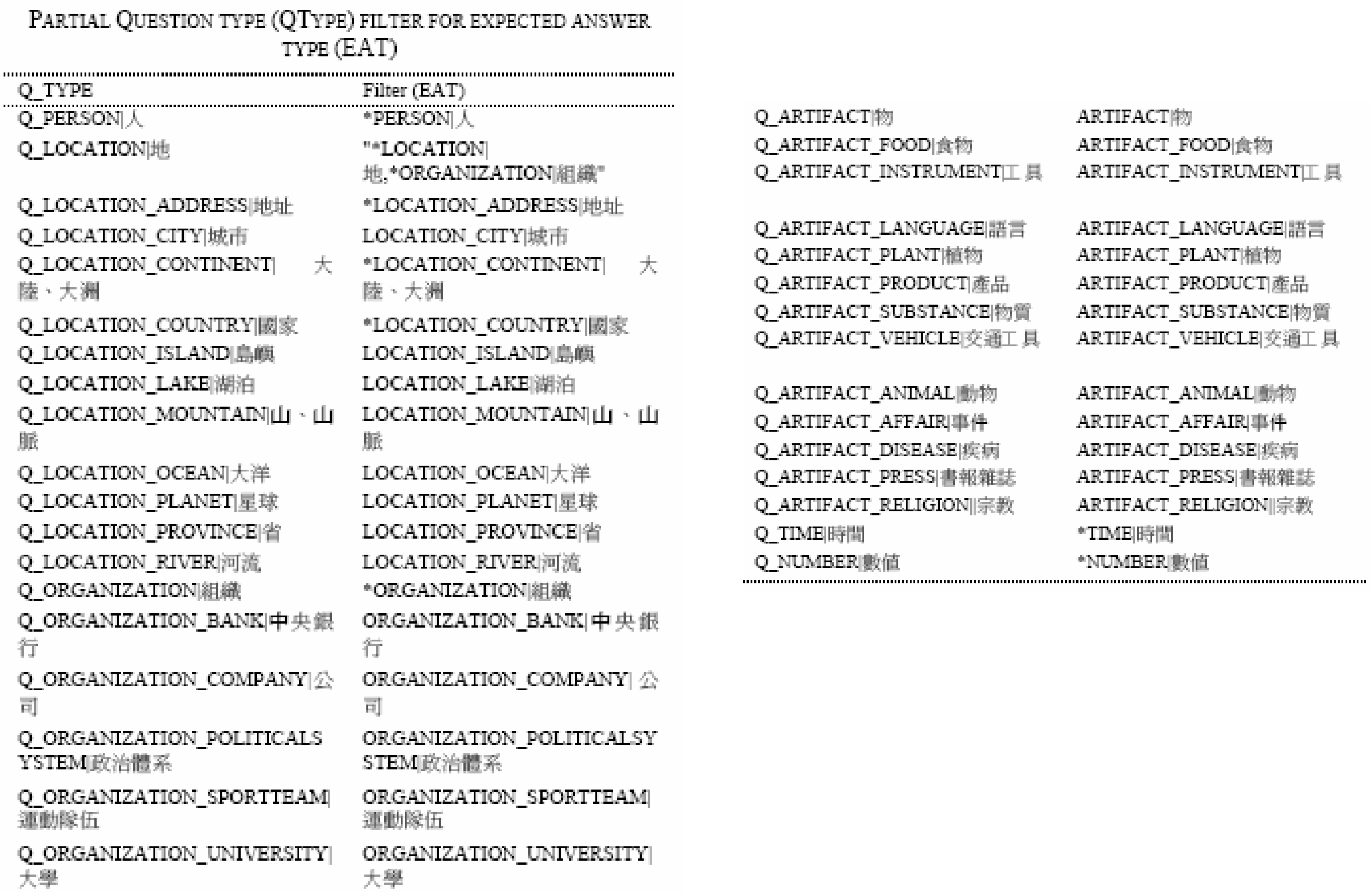
for NTCIR CLQA Factoid Question Answering



TAXONOMY OF CHINESE QUESTION CLASSIFICATION (CQC) FOR CLQA

Coarse-grained (6)	Fine-grained (62)
Q_PERSON|人	Q_PERSON_APPELLATION|稱謂
	Q_PERSON_DISCOVERERS|發現者
	Q_PERSON_FIRSTPERSON|第一人
	Q_PERSON_INVENTORS|發明者
	Q_PERSON_OTHER|人其他類
	Q_PERSON_PERSON|人名
	Q_PERSON_POSITIONS|職位
Q_LOCATION|地	Q_LOCATION_ADDRESS|地址
	Q_LOCATION_CITY|城市
	Q_LOCATION_CONTINENT|大陸、大洲
	Q_LOCATION_COUNTRY|國家
	Q_LOCATION_ISLAND|島嶼
	Q_LOCATION_LAKE|湖泊
	Q_LOCATION_MOUNTAIN|山、山脈
	Q_LOCATION_OCEAN|大洋
	Q_LOCATION_OTHER|地其他類
	Q_LOCATION_PLANET|星球
	Q_LOCATION_PROVINCE|省
	Q_LOCATION_RIVER|河流
Q_ORGANIZATION|組織	Q_ORGANIZATION_BANK|中央銀行
	Q_ORGANIZATION_COMPANY|公司
	Q_ORGANIZATION_OTHER|組織其他類
	Q_ORGANIZATION_POLITICALSYSTEM|政治體系
	Q_ORGANIZATION_SPORTTEAM|運動隊伍
	Q_ORGANIZATION_UNIVERSITY|大學
Q_ARTIFACT|物	Q_ARTIFACT_COLOR|顏色
	Q_ARTIFACT_CURRENCY|貨幣
	Q_ARTIFACT_ENTERTAINMENT|娛樂
	Q_ARTIFACT_FOOD|食物
	Q_ARTIFACT_INSTRUMENT|工具
	Q_ARTIFACT_LANGUAGE|語言
	Q_ARTIFACT_OTHER|物其他類
	Q_ARTIFACT_PLANT|植物
	Q_ARTIFACT_PRODUCT|產品
	Q_ARTIFACT_SUBSTANCE|物質
	Q_ARTIFACT_VEHICLE|交通工具
	Q_ARTIFACT_ANIMAL|動物
	Q_ARTIFACT_AFFAIR|事件
	Q_ARTIFACT_DISEASE|疾病
	Q_ARTIFACT_PRESS|書報雜誌
	Q_ARTIFACT_RELIGION|宗教
Q_TIME|時間	Q_TIME_DATE|日期
	Q_TIME_DAY|日
	Q_TIME_MONTH|月
	Q_TIME_OTHER|時間其他類
	Q_TIME_RANGE|時間範圍
	Q_TIME_TIME|時間
	Q_TIME_YEAR|年
Q_NUMBER|數值	Q_NUMBER_AGE|年齡
	Q_NUMBER_AREA|面積
	Q_NUMBER_COUNT|數字
	Q_NUMBER_LENGTH|長度
	Q_NUMBER_FREQUENCY|頻率
	Q_NUMBER_MONEY|金額
	Q_NUMBER_ORDER|序數
	Q_NUMBER_OTHER|數值其他類
	Q_NUMBER_PERCENT|比例
	Q_NUMBER_PHONENUMBER|電話號碼、郵遞區號
	Q_NUMBER_RANGE|數字範圍
	Q_NUMBER_SPEED|速度
	Q_NUMBER_TEMPERATURE|溫度
	Q_NUMBER_WEIGHT|重量

Question Type (QType) Filter for Expected Answer Type (EAT)



PARTIAL QUESTION TYPE (QTYPE) FILTER FOR EXPECTED ANSWER TYPE (EAT)

<table>
<tr><th>Q_TYPE</th><th>Filter (EAT)</th></tr>
<tr><td>Q_PERSON|人</td><td>*PERSON|人</td></tr>
<tr><td>Q_LOCATION|地</td><td>"*LOCATION|地,*ORGANIZATION|組織"</td></tr>
<tr><td>Q_LOCATION_ADDRESS|地址</td><td>*LOCATION_ADDRESS|地址</td></tr>
<tr><td>Q_LOCATION_CITY|城市</td><td>LOCATION_CITY|城市</td></tr>
<tr><td>Q_LOCATION_CONTINENT| 大陸、大洲</td><td>*LOCATION_CONTINENT| 大陸、大洲</td></tr>
<tr><td>Q_LOCATION_COUNTRY|國家</td><td>*LOCATION_COUNTRY|國家</td></tr>
<tr><td>Q_LOCATION_ISLAND|島嶼</td><td>LOCATION_ISLAND|島嶼</td></tr>
<tr><td>Q_LOCATION_LAKE|湖泊</td><td>LOCATION_LAKE|湖泊</td></tr>
<tr><td>Q_LOCATION_MOUNTAIN|山、山脈</td><td>LOCATION_MOUNTAIN|山、山脈</td></tr>
<tr><td>Q_LOCATION_OCEAN|大洋</td><td>LOCATION_OCEAN|大洋</td></tr>
<tr><td>Q_LOCATION_PLANET|星球</td><td>LOCATION_PLANET|星球</td></tr>
<tr><td>Q_LOCATION_PROVINCE|省</td><td>LOCATION_PROVINCE|省</td></tr>
<tr><td>Q_LOCATION_RIVER|河流</td><td>LOCATION_RIVER|河流</td></tr>
<tr><td>Q_ORGANIZATION|組織</td><td>*ORGANIZATION|組織</td></tr>
<tr><td>Q_ORGANIZATION_BANK|中央銀行</td><td>ORGANIZATION_BANK|中央銀行</td></tr>
<tr><td>Q_ORGANIZATION_COMPANY|公司</td><td>ORGANIZATION_COMPANY| 公司</td></tr>
<tr><td>Q_ORGANIZATION_POLITICALSYSTEM|政治體系</td><td>ORGANIZATION_POLITICALSYSTEM|政治體系</td></tr>
<tr><td>Q_ORGANIZATION_SPORTTEAM|運動隊伍</td><td>ORGANIZATION_SPORTTEAM|運動隊伍</td></tr>
<tr><td>Q_ORGANIZATION_UNIVERSITY|大學</td><td>ORGANIZATION_UNIVERSITY|大學</td></tr>
<tr><td>Q_ARTIFACT|物</td><td>ARTIFACT|物</td></tr>
<tr><td>Q_ARTIFACT_FOOD|食物</td><td>ARTIFACT_FOOD|食物</td></tr>
<tr><td>Q_ARTIFACT_INSTRUMENT|工具</td><td>ARTIFACT_INSTRUMENT|工具</td></tr>
<tr><td>Q_ARTIFACT_LANGUAGE|語言</td><td>ARTIFACT_LANGUAGE|語言</td></tr>
<tr><td>Q_ARTIFACT_PLANT|植物</td><td>ARTIFACT_PLANT|植物</td></tr>
<tr><td>Q_ARTIFACT_PRODUCT|產品</td><td>ARTIFACT_PRODUCT|產品</td></tr>
<tr><td>Q_ARTIFACT_SUBSTANCE|物質</td><td>ARTIFACT_SUBSTANCE|物質</td></tr>
<tr><td>Q_ARTIFACT_VEHICLE|交通工具</td><td>ARTIFACT_VEHICLE|交通工具</td></tr>
<tr><td>Q_ARTIFACT_ANIMAL|動物</td><td>ARTIFACT_ANIMAL|動物</td></tr>
<tr><td>Q_ARTIFACT_AFFAIR|事件</td><td>ARTIFACT_AFFAIR|事件</td></tr>
<tr><td>Q_ARTIFACT_DISEASE|疾病</td><td>ARTIFACT_DISEASE|疾病</td></tr>
<tr><td>Q_ARTIFACT_PRESS|書報雜誌</td><td>ARTIFACT_PRESS|書報雜誌</td></tr>
<tr><td>Q_ARTIFACT_RELIGION||宗教</td><td>ARTIFACT_RELIGION||宗教</td></tr>
<tr><td>Q_TIME|時間</td><td>*TIME|時間</td></tr>
<tr><td>Q_NUMBER|數值</td><td>*NUMBER|數值</td></tr>
</table>

INFOMAP (Knowledge-based Approach)

- INFOMAP: Knowledge Representation Framework
  - Extracts important concepts from a natural language text
- Feature of INFOMAP
  - represent and match complicated template structures
    - hierarchical matching
    - regular expressions
    - semantic template matching
    - frame (non-linear relations) matching
    - graph matching
- We adopt INFOMAP as the knowledge-based approach for CQC
  - Using INFOMAP, we can identify the question category from a Chinese question

Knowledge Representation of Chinese Question

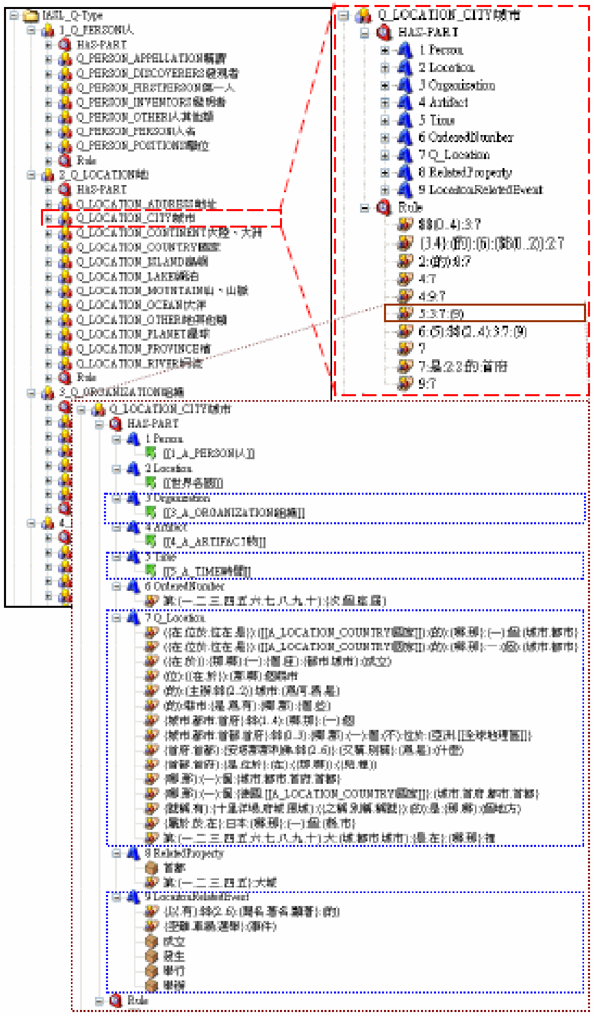
Chinese Question:

2004年奧運在哪一個城市舉行?

(In which city were the Olympics held in 2004?)

[5 Time]:[3 Organization]:[7 Q_Location]:([9 LocaitonRelatedEvent])

Knowledge representation for CQC in INFOMAP



```
IASL_Q-Type
  1_Q_PERSON人
    HAS-PART
    Q_PERSON_APPELLATION稱謂
    Q_PERSON_DISCOVERERS發現者
    Q_PERSON_FIRSTPERSON第一人
    Q_PERSON_INVENTORS發明者
    Q_PERSON_OTHER人其他類
    Q_PERSON_PERSON人名
    Q_PERSON_POSITION職位
    Rule
  2_Q_LOCATION地
    HAS-PART
    Q_LOCATION_ADDRESS地址
    Q_LOCATION_CITY城市
    Q_LOCATION_CONTINENT大陸、大洲
    Q_LOCATION_COUNTRY國家
    Q_LOCATION_ISLAND島嶼
    Q_LOCATION_LAKE湖泊
    Q_LOCATION_MOUNTAIN山、山脈
    Q_LOCATION_OCEAN大洋
    Q_LOCATION_OTHER地其他類
    Q_LOCATION_PLANET星球
    Q_LOCATION_PROVINCE省
    Q_LOCATION_RIVER河流
    Rule
  3_Q_ORGANIZATION組織
  4_
```

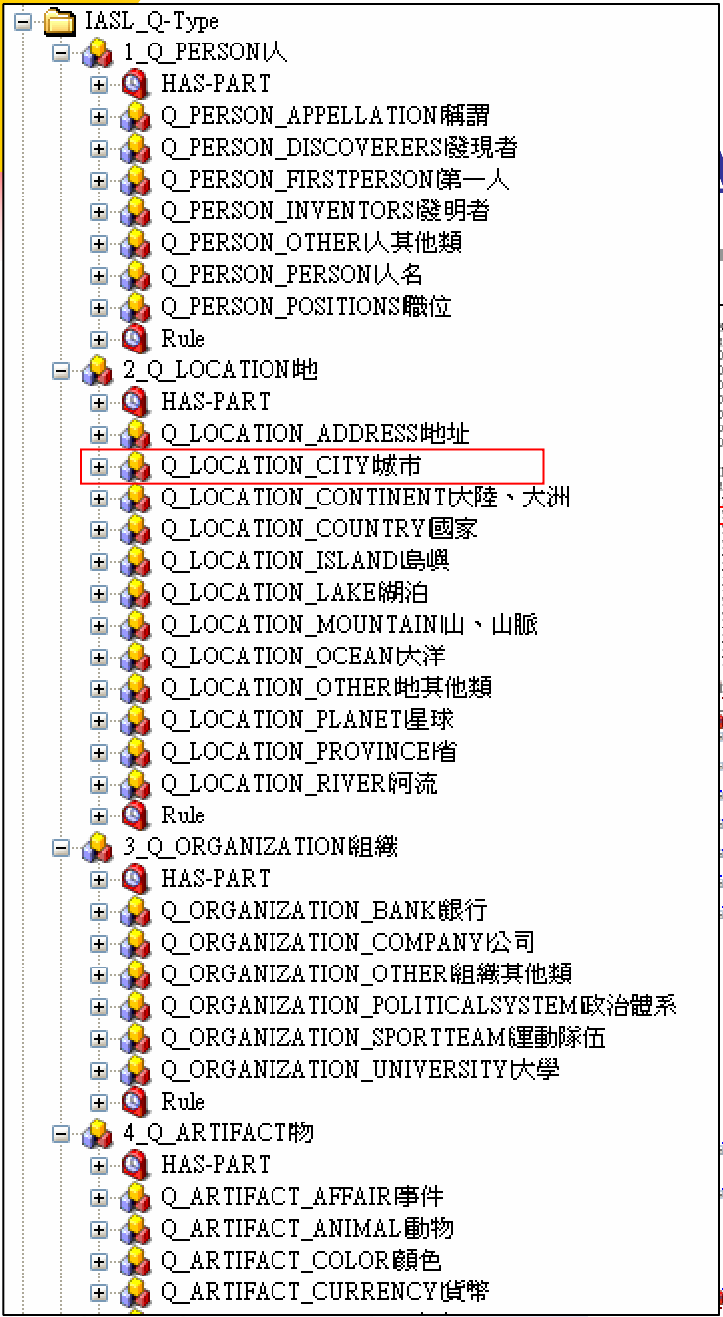
```
Q_LOCATION_CITY城市
  HAS-PART
    1 Person
    2 Location
    3 Organization
    4 Artifact
    5 Time
    6 OrderedNumber
    7 Q_Location
    8 RelatedProperty
    9 LocationRelatedEvent
  Rule
    $$(1..4):3:7
    {3.4}:{的}:{6}:{$$(1..2)}:2:7
    2:{的}:4:7
    4:7
    4:9:7
    5:3:7:{8}
    6:{5}:$$(2..4):3:7:{9}
    7
    7:是:2:2:的:首府
    9:7
```

```
Q_LOCATION_CITY城市
  HAS-PART
    1 Person
      [[1_A_PERSON人]]
    2 Location
      [[世界各國]]
    3 Organization
      [[3_A_ORGANIZATION組織]]
    4 Artifact
      [[4_A_ARTIFACT物]]
    5 Time
      [[5_A_TIME時間]]
    6 OrderedNumber
      第{一.二.三.四.五.六.七.八.九.十}:{次.個.屆.屆}
    7 Q_Location
      {在.位於.位在.是}:{[[A_LOCATION_COUNTRY國家]]}:{的}:{哪.那}:{一}:個:{城市.都市}
      {在.位於.位在.是}:{[[A_LOCATION_COUNTRY國家]]}:{的}:{哪.那}:一:{個}:{城市.都市}
      {在.於}:{哪.那}:{一}:{個.座}:{都市.城市}:{成立}
      {位}:{在.於}:{哪.那}:個縣市
      {的}:{主辦.$$(2..2)}:城市:{為何.為.是}
      {的}:都市:{是.為.有}:{哪.那}:{個.些}
      {城市.都市.首府}:$$(1..4):{哪.那}:{一}:個
      {城市.都市.首都.首府}:$$(0..3):{哪.那}:{一}:個:{不}:位於:{亞洲.[[全球地理區]]}
      {首府.首都}:{安哥拉.索馬利亞}:$$(2..6):{又稱.別稱}:{為.是}:{什麼}
      {首都.首府}:{是.位於}:{在}:{{哪.那}}:{{裡.裡}}
      {哪.那}:{一}:個:{城市.都市.首府.首都}
      {哪.那}:{一}:個:{德國.[[A_LOCATION_COUNTRY國家]]}:{城市.首府.都市.首都}
      {被稱.有}:{十里洋場.府城.風城}:{之稱.別稱.綽號}:{的}:是:{哪.那}:{個地方}
      {屬於.於.在}:日本:{哪.那}:{一}:個:{縣.市}
      第{一.二.三.四.五.六.七.八.九.十}大:{城.都市.城市}:{是.在}:{哪.那}:裡
    8 RelatedProperty
      首都
      第{一.二.三.四.五}:大城
    9 LocationRelatedEvent
      {以.有}:$$(2..6):{聞名.著名.舉行}:{的}
      {空難.車禍.選舉}:{事件}
      成立
      發生
      舉行
      舉辦
  Rule
```

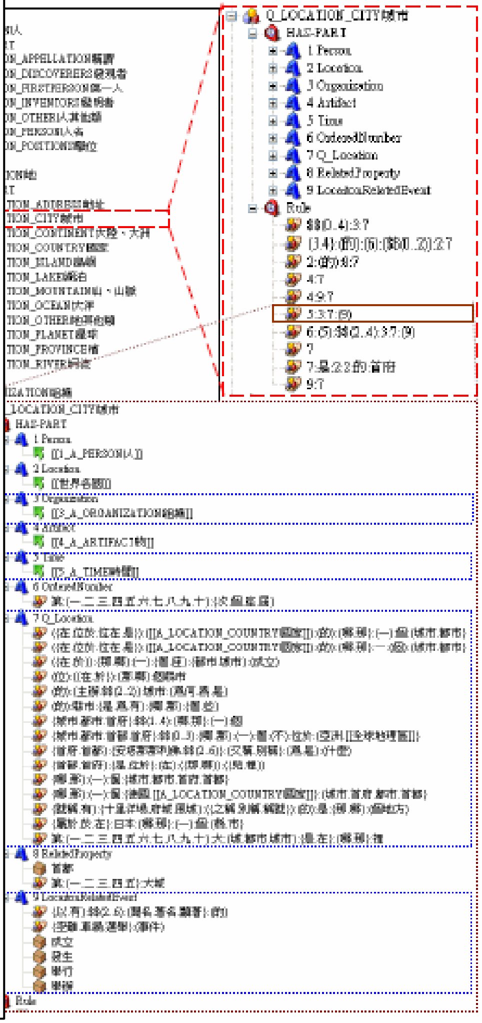
Fig.1. Knowledge representation for CQC in INFOMAP

2004年奧運在哪一個城市舉行？ (In which city were the Olympics held in 2004?)

representation for CQC
AP



IASL_Q-Type
- 1_Q_PERSON人
  - HAS-PART
  - Q_PERSON_APPELLATION稱謂
  - Q_PERSON_DISCOVERERS發現者
  - Q_PERSON_FIRSTPERSON第一人
  - Q_PERSON_INVENTORS發明者
  - Q_PERSON_OTHER人其他類
  - Q_PERSON_PERSON人名
  - Q_PERSON_POSITIONS職位
  - Rule
- 2_Q_LOCATION地
  - HAS-PART
  - Q_LOCATION_ADDRESS地址
  - Q_LOCATION_CITY城市
  - Q_LOCATION_CONTINENT大陸、大洲
  - Q_LOCATION_COUNTRY國家
  - Q_LOCATION_ISLAND島嶼
  - Q_LOCATION_LAKE湖泊
  - Q_LOCATION_MOUNTAIN山、山脈
  - Q_LOCATION_OCEAN大洋
  - Q_LOCATION_OTHER地其他類
  - Q_LOCATION_PLANET星球
  - Q_LOCATION_PROVINCE省
  - Q_LOCATION_RIVER河流
  - Rule
- 3_Q_ORGANIZATION組織
  - HAS-PART
  - Q_ORGANIZATION_BANK銀行
  - Q_ORGANIZATION_COMPANY公司
  - Q_ORGANIZATION_OTHER組織其他類
  - Q_ORGANIZATION_POLITICALSYSTEM政治體系
  - Q_ORGANIZATION_SPORTTEAM運動隊伍
  - Q_ORGANIZATION_UNIVERSITY大學
  - Rule
- 4_Q_ARTIFACT物
  - HAS-PART
  - Q_ARTIFACT_AFFAIR事件
  - Q_ARTIFACT_ANIMAL動物
  - Q_ARTIFACT_COLOR顏色
  - Q_ARTIFACT_CURRENCY貨幣

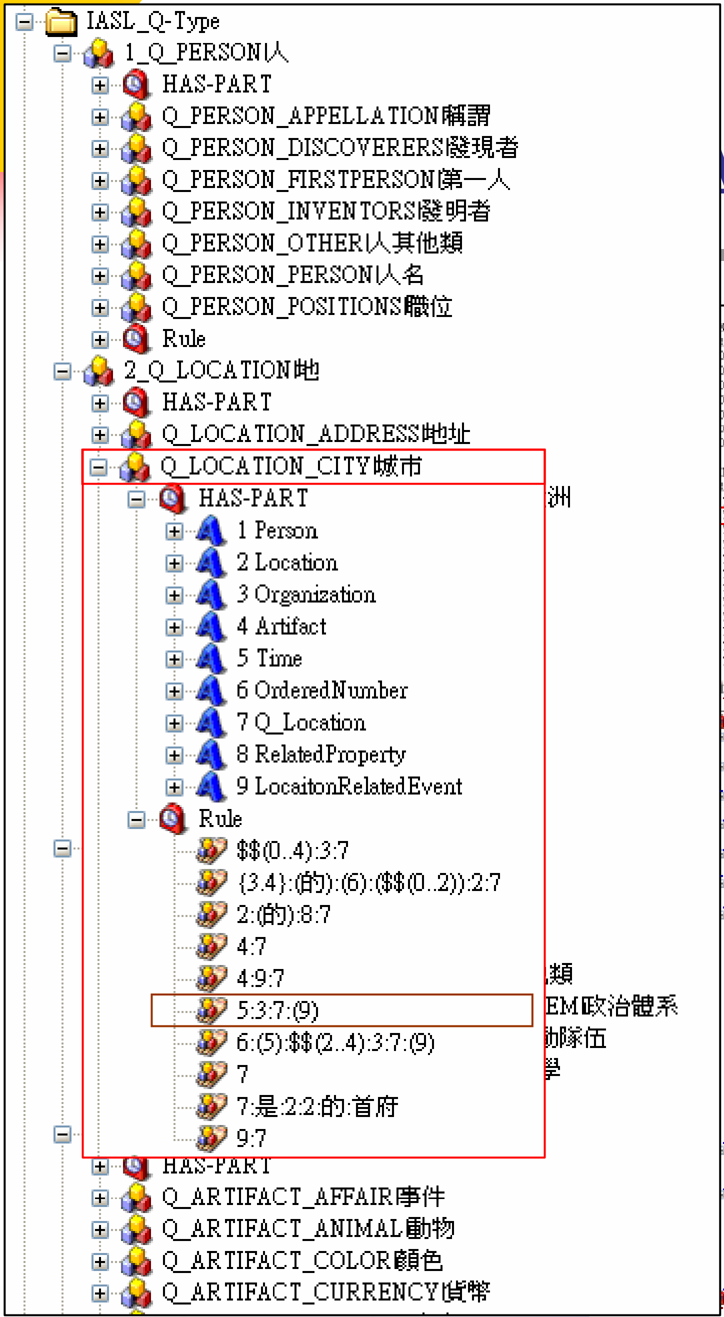


Q_LOCATION_CITY城市
- HAS-PART
  - 1 Person
  - 2 Location
  - 3 Organization
  - 4 Artifact
  - 5 Time
  - 6 OrderedNumber
  - 7 Q_Location
  - 8 RelatedProperty
  - 9 LocationRelatedEvent
- Rule
  - $$(0..4):3:7
  - (3.4):(的):(6):($$(0..2)):2:7
  - 2:(的):4:7
  - 4:7
  - 4:9:7
  - 5:3:7:(9)
  - 6:(5):$$(2..4):3:7:(9)
  - 7
  - 7:是:2:2:的:首府
  - 9:7

Fig.1. Knowledge representation for CQC in INFOMAP

2004年奧運在哪一個城市舉行? (In which city were the Olympics held in 2004?)

representation for CQC
AP



```
IASL_Q-Type
  1_Q_PERSON人
    HAS-PART
    Q_PERSON_APPELLATION稱謂
    Q_PERSON_DISCOVERERS發現者
    Q_PERSON_FIRSTPERSON第一人
    Q_PERSON_INVENTORS發明者
    Q_PERSON_OTHER人其他類
    Q_PERSON_PERSON人名
    Q_PERSON_POSITIONS職位
    Rule
  2_Q_LOCATION地
    HAS-PART
    Q_LOCATION_ADDRESS地址
    Q_LOCATION_CITY城市
      HAS-PART
        1 Person
        2 Location
        3 Organization
        4 Artifact
        5 Time
        6 OrderedNumber
        7 Q_Location
        8 RelatedProperty
        9 LocaitonRelatedEvent
      Rule
        $$(0..4):3:7
        {3.4}:(的):(6):($$(0..2)):2:7
        2:(的):8:7
        4:7
        4:9:7
        5:3:7:(9)
        6:(5):$$(2..4):3:7:(9)
        7
        7:是:2:2:的:首府
        9:7
    HAS-PART
    Q_ARTIFACT_AFFAIR事件
    Q_ARTIFACT_ANIMAL動物
    Q_ARTIFACT_COLOR顏色
    Q_ARTIFACT_CURRENCY貨幣
```

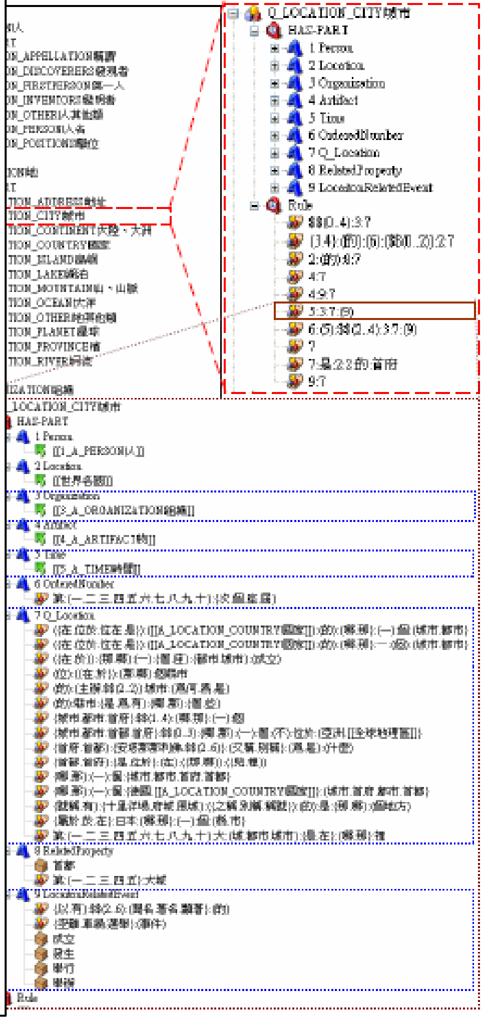
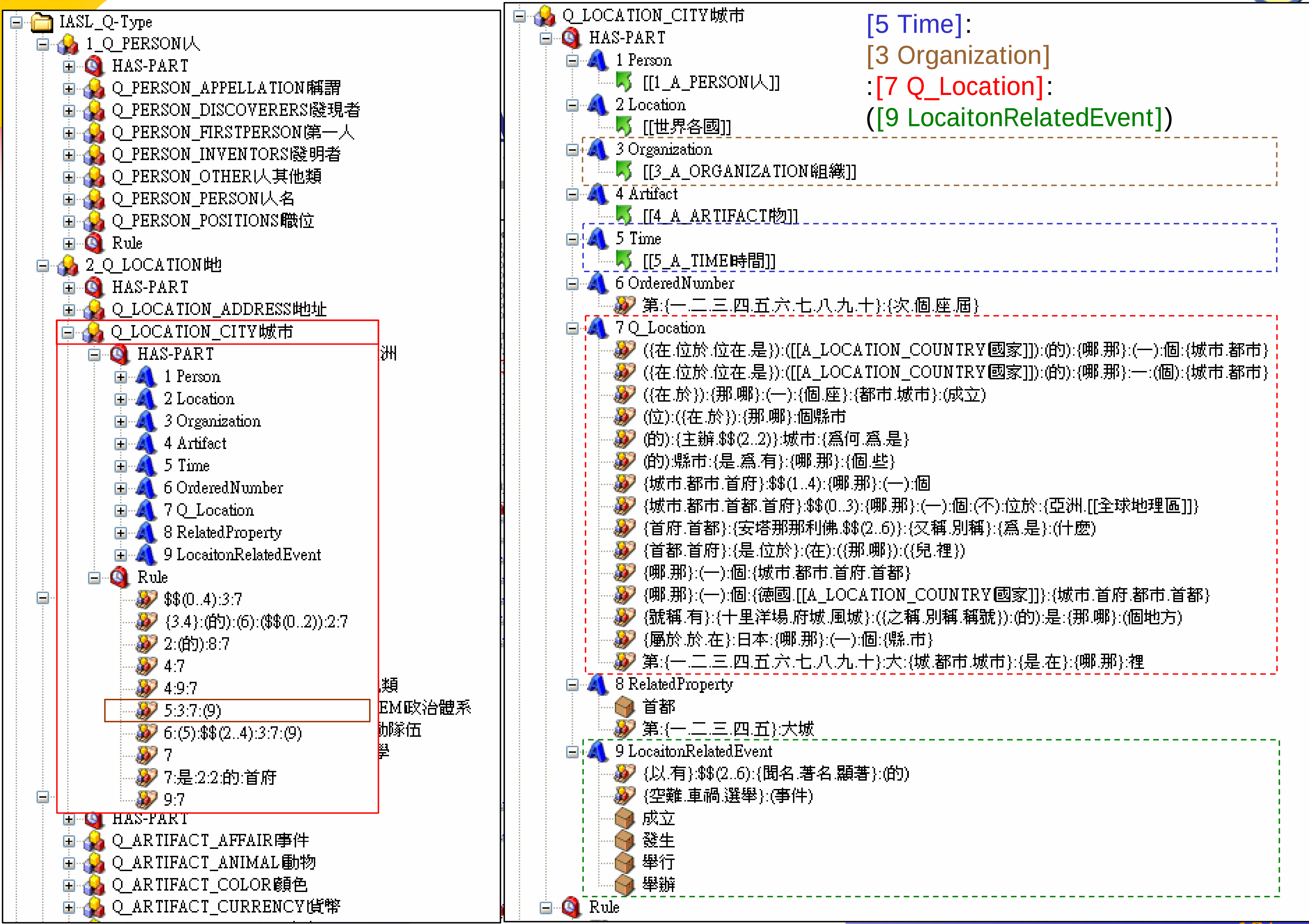


Fig.1. Knowledge representation for CQC in INFOMAP

2004年奧運在哪一個城市舉行? (In which city were the Olympics held in 2004?)



```
IASL_Q-Type
  1_Q_PERSON人
    HAS-PART
    Q_PERSON_APPELLATION稱謂
    Q_PERSON_DISCOVERERS發現者
    Q_PERSON_FIRSTPERSON第一人
    Q_PERSON_INVENTORS發明者
    Q_PERSON_OTHER人其他類
    Q_PERSON_PERSON人名
    Q_PERSON_POSITIONS職位
    Rule
  2_Q_LOCATION地
    HAS-PART
    Q_LOCATION_ADDRESS地址
    Q_LOCATION_CITY城市
      HAS-PART
        1 Person
        2 Location
        3 Organization
        4 Artifact
        5 Time
        6 OrderedNumber
        7 Q_Location
        8 RelatedProperty
        9 LocaitonRelatedEvent
      Rule
        $$(0..4):3:7
        {3.4}:(的):(6):($$(0..2)):2:7
        2:(的):8:7
        4:7
        4:9:7
        5:3:7:(9)
        6:(5):$$(2..4):3:7:(9)
        7
        7:是:2:2:的:首府
        9:7
    HAS-PART
    Q_ARTIFACT_AFFAIR事件
    Q_ARTIFACT_ANIMAL動物
    Q_ARTIFACT_COLOR顏色
    Q_ARTIFACT_CURRENCY貨幣
```

```
Q_LOCATION_CITY城市
  HAS-PART
    1 Person
      [[1_A_PERSON人]]
    2 Location
      [[世界各國]]
    3 Organization
      [[3_A_ORGANIZATION組織]]
    4 Artifact
      [[4_A_ARTIFACT物]]
    5 Time
      [[5_A_TIME時間]]
    6 OrderedNumber
      第:{一.二.三.四.五.六.七.八.九.十}:{次.個.座.屆}
    7 Q_Location
      ({在.位於.位在.是}):([[A_LOCATION_COUNTRY國家]]):(的):{哪.那}:(一):個:{城市.都市}
      ({在.位於.位在.是}):([[A_LOCATION_COUNTRY國家]]):(的):{哪.那}:一:(個):{城市.都市}
      ({在.於}):{那.哪}:(一):{個.座}:{都市.城市}:(成立)
      (位):({在.於}):{那.哪}:個縣市
      (的):{主辦.$$(2..2)}:城市:{爲何.爲.是}
      (的):縣市:{是.爲.有}:{哪.那}:{個.些}
      {城市.都市.首府}:$$(1..4):{哪.那}:(一):個
      {城市.都市.首都.首府}:$$(0..3):{哪.那}:(一):個:(不):位於:{亞洲.[[全球地理區]]}
      {首府.首都}:{安塔那那利佛.$$(2..6)}:{又稱.別稱}:{爲.是}:(什麼)
      {首都.首府}:{是.位於}:(在):({那.哪}):({兒.裡})
      {哪.那}:(一):個:{城市.都市.首府.首都}
      {哪.那}:(一):個:{德國.[[A_LOCATION_COUNTRY國家]]}:{城市.首府.都市.首都}
      {號稱.有}:{十里洋場.府城.風城}:({之稱.別稱.稱號}):(的):是:{那.哪}:(個地方)
      {屬於.於.在}:日本:{哪.那}:(一):個:{縣.市}
      第:{一.二.三.四.五.六.七.八.九.十}:大:{城.都市.城市}:{是.在}:{哪.那}:裡
    8 RelatedProperty
      首都
      第:{一.二.三.四.五}:大城
    9 LocaitonRelatedEvent
      {以.有}:$$(2..6):{聞名.著名.顯著}:(的)
      {空難.車禍.選舉}:(事件)
      成立
      發生
      舉行
      舉辦
  Rule
```

[5 Time]:
[3 Organization]
:[7 Q_Location]:
([9 LocaitonRelatedEvent])

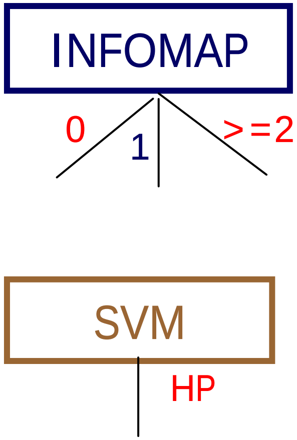
Fig.1. Knowledge representation for CQC in INFOMAP

SVM (Machine Learning Approach)

- Two types of feature used for CQC
    - Syntactic features
        - Bag-of-Words
            - character-based bigram (CB)
            - word-based bigram (WB)
        - Part-of-Speech (POS)
            - AUTOTAG
                - POS tagger developed by CKIP, Academia Sinica
    - Semantic Features
        - HowNet Senses
            - HowNet Main Definition (HNMD)
            - HowNet Definition (HND)

Integration of SVM and INFOMAP (Hybrid Approach)

- The integrated module selects the question type with the **highest confidence score** from the INFOMAP or the SVM model
    - If the question matches the templates or rules represented in INFOMAP and obtains the question type, we use the question type obtain from INFOMAP first.
    - If no question type can be obtained from INFOMAP, we use the result from the SVM model.
    - If multiple question types are obtained from INFOMAP, we choose the one obtained from SVM first.
    - If one question type with a high positive score is obtained from SVM and one question type obtained from INFOMAP, which is not the same as the one from SVM, we choose the one from SVM with a high positive score.

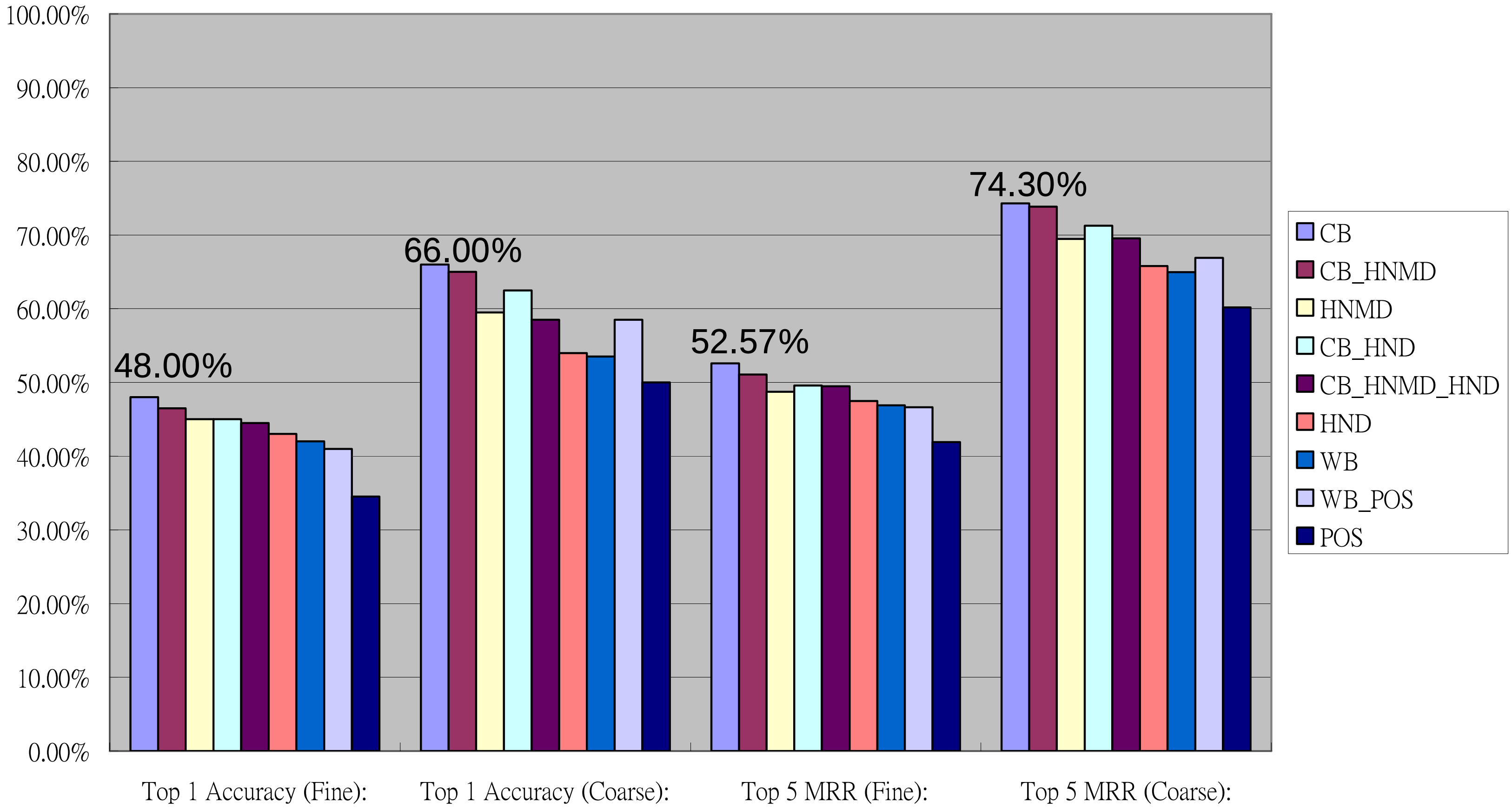


Experimental Results and Discussion

- Datasets
  - Training: 1350 questions
    - 500 questions from CLQA's **development** dataset
      - 300 questions for Japanese news
      - 200 questions for Traditional Chinese news
    - 850 questions manually build for our proposed question taxonomy
      - 518 questions in SVM
      - 332 questions in INFOMAP
  - Testing: 200 questions
    - 200 Questions from CLQA's **formal run** dataset
  - We use different features to train the SVM model based on a total of 1350 questions and their labeled question type

Experimental Results of CLQA’s development dataset

CQC training CLQAS300 model for testing CLQAS200N



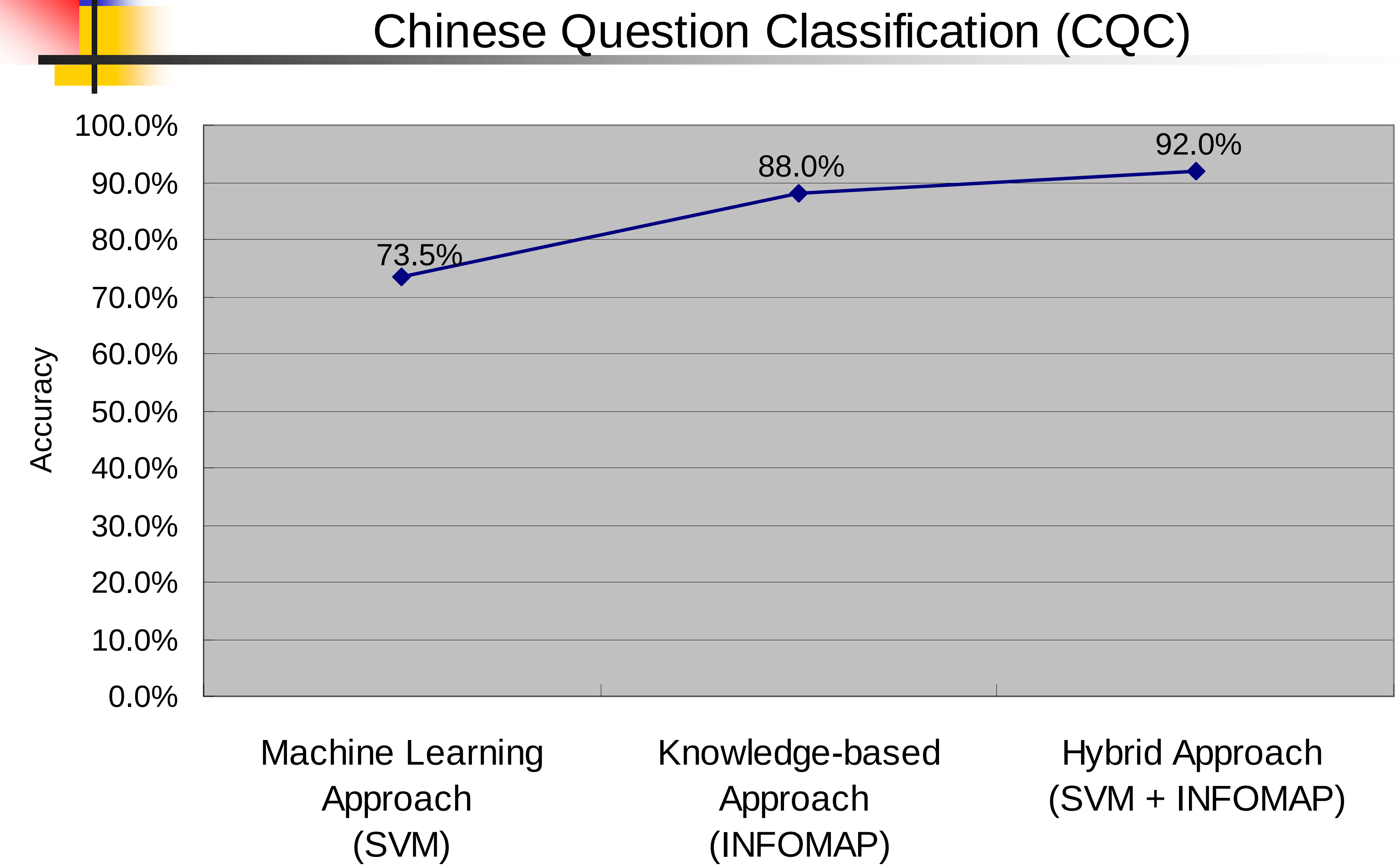
SVM Training data: CLQAS300 (300 questions for Japanese news)

SVM Testing data: CLQAS200 (200 questions for Chinese news)

Experimental Results of CLQA's Formal Run dataset

- Training dataset: 1350 questions
  - 300 (Development dataset for Japanese News) + 200 (Development dataset for Chinese News) + 518 (SVM) + 332 (INFOMAP)
- Features: CB+HNMD
- Testing dataset: 200 questions
  - CLQA's formal run

Experimental Results of CLQA’s Formal Run dataset



Discussion

- **Integrated approach** performs better than the individual knowledge-based or machine learning approach
- **knowledge-based approach** performs well with **easy questions** using the templates and rules
  - Easy questions are defined as follows:
    - **Clear words** that show the question type and indicate the words that are not question types
      - Ex: “誰(Who)”, “哪一位(Which person)”, “首位(the first person)”
    - **Explicit words** that identify the question type. If words are easy to identify, it means they overlap with a question type
      - Ex: “隊伍(team)” and “運動隊伍(sports team)”
    - **Interrogative words** that connect with question type words in question
      - Ex: “那個人(Which Person)”

Related Works

- Li and Roth (2002)
  - 6 coarse classes and 50 fine classes for TREC factoid question answering
  - Sparse Network of Windows (SNoW)
  - Over 90% accuracy
- Zhang and Lee (2003)
  - Support Vector Machines (SVMs)
  - Surface text features (bag-of-words and bag-of-ngrams)
    - coarse-grained: 86% accuracy
    - fine-grained: approximately 80% accuracy.
  - Adding syntactic information
    - coarse-grained: accuracy of 90%
- Suzuki et al. (2003)
  - Hierarchical SVM
  - Four feature sets

    (1) words only
    (2) words and named entities
    (3) words and semantic information
    (4) words and NEs and semantic information
  - Coarse-grained: 95% (depth 1)
  - Fine-grained: 75% (depth 4)

Comparison with related works

- Question classification in Chinese
- The accuracy of CQC
  - SVM: 73.5%
  - INFOMAP: 88%
  - Hybrid Approach (SVM+INFOMAP): 92%

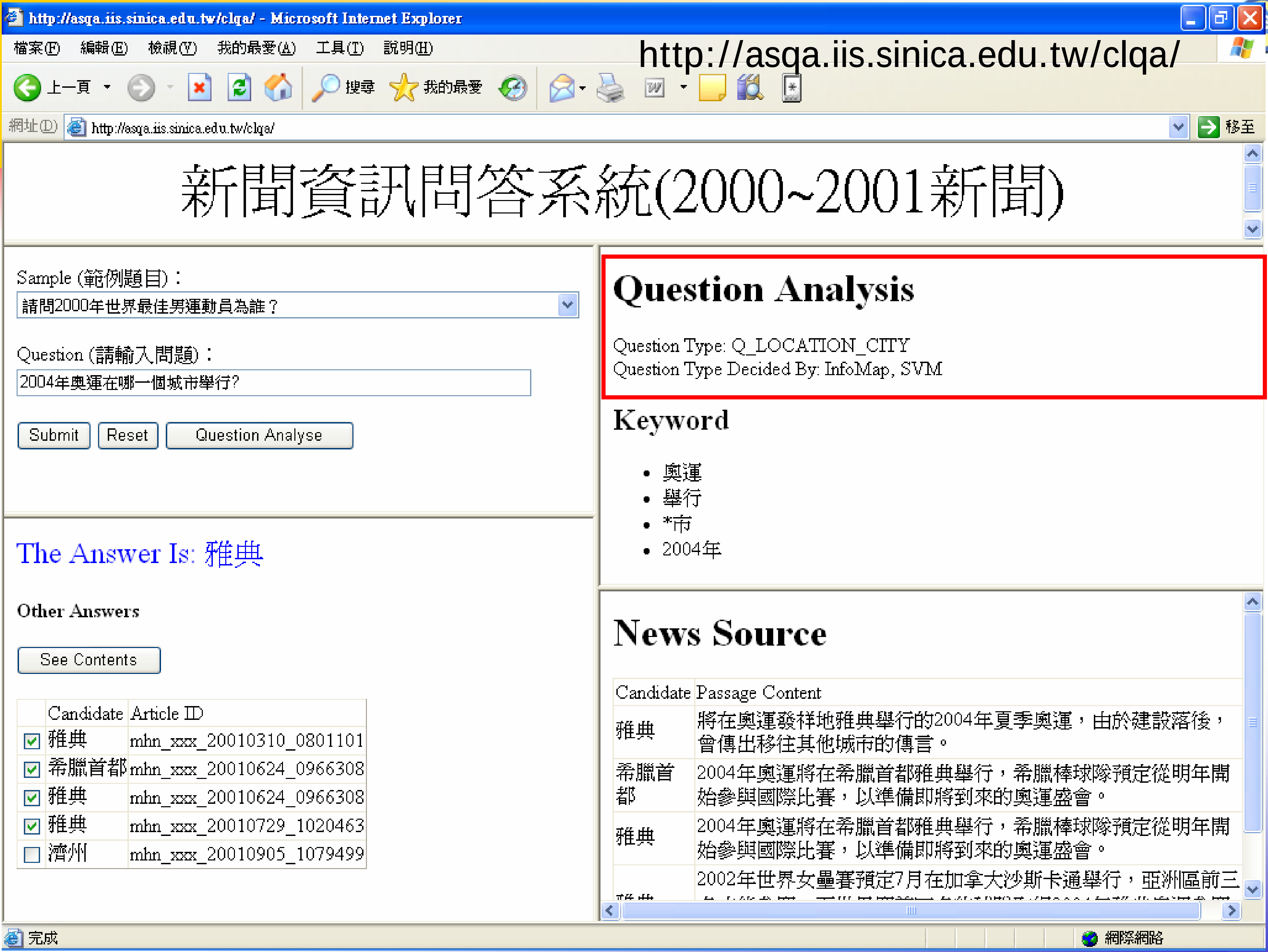
Conclusions

- We have proposed a Hybrid approach to Chinese question classification (CQC) for NTCIR CLQA factoid question-answering
    - Hierarchical coarse-grained and fine-grained question taxonomies
        - 6 coarse-grained categories and 62 fine-grained categories for Chinese questions
    - Mapping method for question type filtering to obtain expected answer types (EAT)
- The integrated knowledge-based and machine learning approach achieves significantly better accuracy rate than individual approaches

Applications: ASQA (Academia Sinica Question Answering system)

- **ASQA** (IASL-IIS-SINICA-TAIWAN)
  - First place in the Chinese-Chinese (C-C) subtask of the NTCIR5 Cross-Language Question Answering (CLQA 2005) task

http://asqa.iis.sinica.edu.tw/clqa/



http://asqa.iis.sinica.edu.tw/clqa/ - 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
網址(D)
http://asqa.iis.sinica.edu.tw/clqa/
移至
完成
網際網路

新聞資訊問答系統(2000~2001新聞)

Sample (範例題目)：
請問2000年世界最佳男運動員為誰？

Question (請輸入問題)：
2004年奧運在哪一個城市舉行?

Submit
Reset
Question Analyse

Question Analysis

Question Type: Q_LOCATION_CITY
Question Type Decided By: InfoMap, SVM

Keyword

- 奧運
- 舉行
- *市
- 2004年

The Answer Is: 雅典

Other Answers

	Candidate	Article ID
☑	雅典	mhn_xxx_20010310_0801101
☑	希臘首都	mhn_xxx_20010624_0966308
☑	雅典	mhn_xxx_20010624_0966308
☑	雅典	mhn_xxx_20010729_1020463
☐	濟州	mhn_xxx_20010905_1079499

News Source

Candidate	Passage Content
雅典	將在奧運發祥地雅典舉行的2004年夏季奧運，由於建設落後，曾傳出移往其他城市的傳言。
希臘首都	2004年奧運將在希臘首都雅典舉行，希臘棒球隊預定從明年開始參與國際比賽，以準備即將到來的奧運盛會。
雅典	2004年奧運將在希臘首都雅典舉行，希臘棒球隊預定從明年開始參與國際比賽，以準備即將到來的奧運盛會。
	2002年世界女壘賽預定7月在加拿大沙斯卡通舉行，亞洲區前三

Q & A

An Integrated Knowledge-based and Machine Learning Approach for Chinese Question Classification

Min-Yuh Day[1,2], Cheng-Wei Lee[1], Shih-Hung Wu[3],
Chorng-Shyong Ong[2], Wen-Lian Hsu[1]

(戴敏育[1,2], 李政緯[1], 吳世弘[3], 翁崇雄[2], 許聞廉[1])

[1] Institute of Information Science, Academia Sinica, Taipei
[2] Department of Information Management, National Taiwan University, Taipei
[3] Dept. of Computer Science and Information Engineering, Chaoyang Univ. of Technology, Taichung

myday@iis.sinica.edu.tw