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

Min-Yuh Day¹,², Cheng-Wei Lee¹, Shih-Hung Wu³, Chorng-Shyong Ong², Wen-Lian Hsu¹

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: 奥运的发源地在哪裡？
    Where is the originating place of the Olympics?
  - Question Type: Q_LOCATION

- **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

<table>
<thead>
<tr>
<th>Coarse-grained (6)</th>
<th>Fine-grained (32)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q_PERSON</strong> (人)</td>
<td>Q_ARTIFACT (物)</td>
</tr>
<tr>
<td>Q_PERSON_APPELLATION (稱謂)</td>
<td>Q_ARTIFACT_COLOR (颜色)</td>
</tr>
<tr>
<td>Q_PERSON_DISCOVERERS (發現者)</td>
<td>Q_ARTIFACT_CURRENCY (貨幣)</td>
</tr>
<tr>
<td>Q_PERSON_FIRSTPERSON (第一人)</td>
<td>Q_ARTIFACT_ENTERTAINMENT (娛樂)</td>
</tr>
<tr>
<td>Q_PERSON_INVENTORS (發明者)</td>
<td>Q_ARTIFACT_FOOD (食物)</td>
</tr>
<tr>
<td>Q_PERSON_OTHER (其他人)</td>
<td>Q_ARTIFACT_INSTRUMENT (工具)</td>
</tr>
<tr>
<td>Q_PERSON_NAME (姓名)</td>
<td>Q_ARTIFACT_LANGUAGE (語言)</td>
</tr>
<tr>
<td>Q_PERSON_POSITIONS (職位)</td>
<td>Q_ARTIFACT_OTHER (物其他類)</td>
</tr>
<tr>
<td>Q_PERSON_SOURCES (來源)</td>
<td>Q_ARTIFACT_PLANT (植物)</td>
</tr>
<tr>
<td>Q_PERSON_SOURCES (來源)</td>
<td>Q_ARTIFACT_PRODUCT (產品)</td>
</tr>
<tr>
<td>Q_PERSON_SOURCES (來源)</td>
<td>Q_ARTIFACT_SUBSTANCE (物質)</td>
</tr>
<tr>
<td>Q_PERSON_SOURCES (來源)</td>
<td>Q_ARTIFACT_VEHICLE (交通工具)</td>
</tr>
<tr>
<td>Q_PERSON_SOURCES (來源)</td>
<td>Q_ARTIFACT_ANIMAL (動物)</td>
</tr>
<tr>
<td>Q_PERSON_SOURCES (來源)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_PERSON_SOURCES (來源)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_LOCATION (地)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_LOCATION_ADDRESS (地址)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_LOCATION_CITY (城市)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_LOCATION_COUNTRY (國家)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_LOCATION_ISLAND (島嶼)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_LOCATION_LAKE (湖泊)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_LOCATION_MOUNTAIN (山脈)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_LOCATION_OCEAN (大海洋)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_LOCATION_OTHER (地其他類)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_LOCATION_PLANET (星球)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_LOCATION_PROVINCE (省)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_LOCATION_RIVER (河流)</td>
<td>Q_ARTIFACT_AFFAIR (事件)</td>
</tr>
<tr>
<td>Q_ORGANIZATION (組織)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_ORGANIZATION_BANK (央行)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_ORGANIZATION_COMPANY (公司)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_ORGANIZATION_SPORT (國)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_ORGANIZATION_SPORTTEAM (運動隊)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_ORGANIZATION_UNIVERSITY (大學)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL (數值)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_AGE (年齡)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_AREA (面積)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_COUNT (數字)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_LENGTH (長度)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_FREQUENCY (頻率)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_MONEY (金額)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_ORDER (字數)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_OTHER (數值其他類)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_PERCENT (百分比)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_PHONE (電話號碼)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_RANGE (數值範圍)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_SPEED (速度)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_TEMPERATURE (溫度)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
<tr>
<td>Q_NUMERICAL_WEIGHT (重量)</td>
<td>Q_NUMERICAL (數值)</td>
</tr>
</tbody>
</table>
### Partial Question Type (QType) Filter for Expected Answer Type (EAT)

<table>
<thead>
<tr>
<th>Q_TYPE</th>
<th>Filter (EAT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q_Person</td>
<td><em>PERSON</em></td>
</tr>
<tr>
<td>Q_Location</td>
<td><em>LOCATION</em></td>
</tr>
<tr>
<td>Q_Location_Address</td>
<td><em>LOCATION_ADDRESS</em></td>
</tr>
<tr>
<td>Q_Location_City</td>
<td><em>LOCATION_CITY</em></td>
</tr>
<tr>
<td>Q_Location_Continent</td>
<td><em>LOCATION_CONTINENT</em></td>
</tr>
<tr>
<td>Q_Location_Country</td>
<td><em>LOCATION_COUNTRY</em></td>
</tr>
<tr>
<td>Q_Location_Lake</td>
<td><em>LOCATION_LAKE</em></td>
</tr>
<tr>
<td>Q_Location_Mountain</td>
<td><em>LOCATION_MOUNTAIN</em></td>
</tr>
<tr>
<td>Q_Location_Ocean</td>
<td><em>LOCATION_OCEAN</em></td>
</tr>
<tr>
<td>Q_Location_Planet</td>
<td><em>LOCATION_PLANET</em></td>
</tr>
<tr>
<td>Q_Location_Province</td>
<td><em>LOCATION_PROVINCE</em></td>
</tr>
<tr>
<td>Q_Location_River</td>
<td><em>LOCATION_RIVER</em></td>
</tr>
<tr>
<td>Q_Organization</td>
<td><em>ORGANIZATION</em></td>
</tr>
<tr>
<td>Q_Organization_Bank</td>
<td><em>ORGANIZATION_BANK</em></td>
</tr>
<tr>
<td>Q_Organization_Company</td>
<td><em>ORGANIZATION_COMPANY</em></td>
</tr>
<tr>
<td>Q_Organization_PoliticalSystem</td>
<td><em>ORGANIZATION_POLITICALSYSTEM</em></td>
</tr>
<tr>
<td>Q_Organization_SportTeam</td>
<td><em>ORGANIZATION_SPORTTEAM</em></td>
</tr>
<tr>
<td>Q_Organization_University</td>
<td><em>ORGANIZATION_UNIVERSITY</em></td>
</tr>
<tr>
<td>Q_Number</td>
<td><em>NUMBER</em></td>
</tr>
</tbody>
</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
2004年奧運在哪一個城市舉行？ (In which city were the Olympics held in 2004?)

Fig 1. Knowledge representation for CQC in INFOMAP
Min-Yuh Day (SINICA; NTU)

Knowledge representation for CQC in INFOMAP

Fig 1: Knowledge representation for CQC in INFOMAP
In which city were the Olympics held in 2004?

Fig 1. Knowledge representation for CQC in INFORMAP
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 obtained 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 built 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

Chinese Question Classification (CQC)

Accuracy

- Machine Learning Approach (SVM)
- Knowledge-based Approach (INFOMAP)
- Hybrid Approach (SVM + INFOMAP)

Accuracy:
- Machine Learning Approach: 73.5%
- Knowledge-based Approach: 88.0%
- Hybrid Approach: 92.0%
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

  - 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 NTCI R5 Cross-Language Question Answering (CLQA 2005) task
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？

The Answer Is: 雅典

Keyword
- 奧運
- 舉行
- *市
- 2004年

News Source

Candidate Passage Content

雅典 2004年奧運將在雅典舉行
希臘首都2004年奧運將在希臘首都雅典舉行
雅典 2004年奧運將在希臘首都雅典舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
希臘首都2004年奧運將在希臘首都雅典舉行
希臘首都2004年奧運將在希臘首都雅典舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
希臘首都2004年奧運將在希臘首都雅典舉行
希臘首都2004年奧運將在希臘首都雅典舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
希臘首都2004年奧運將在希臘首都雅典舉行
希臘首都2004年奧運將在希臘首都雅典舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
希臘首都2004年奧運將在希臘首都雅典舉行
希臘首都2004年奧運將在希臘首都雅典舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
希臘首都2004年奧運將在希臘首都雅典舉行
希臘首都2004年奧運將在希臘首都雅典舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
希臘首都2004年奧運將在希臘首都雅典舉行
希臘首都2004年奧運將在希臘首都雅典舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
希臘首都2004年奧運將在希臘首都雅典舉行
希臘首都2004年奧運將在希臘首都雅典舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
希臘首都2004年奧運將在希臘首都雅典舉行
希臘首都2004年奧運將在希臘首都雅典舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
希臘首都2004年奧運將在希臘首都雅典舉行
希臘首都2004年奧運將在希臘首都雅典舉行
雅典 2002年世界女壘賽預定7月在加拿大多倫多舉行
雅典 2002年世界女壘賽預定7月在加拿多く
Q & A

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

Min-Yuh Day\textsuperscript{1,2}, Cheng-Wei Lee\textsuperscript{1}, Shih-Hung Wu\textsuperscript{3}, Chorng-Shyong Ong\textsuperscript{2}, Wen-Lian Hsu\textsuperscript{1}

(戴敏育\textsuperscript{1,2}, 李政緯\textsuperscript{1}, 吳世弘\textsuperscript{3}, 翁崇雄\textsuperscript{2}, 許聞廉\textsuperscript{1})

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

myday@iis.sinica.edu.tw

IEEE NLPKE 2005