From Text to Image: Generating Visual Query for Image Retrieval

Wen-Cheng Lin, Yih-Chen Chang and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN

CLEF 2004

Outline

- Introduction
- Visual representation of textual query
- Query translation
- Experiment results and discussion
- Conclusion

Introduction

- Multimedia data
    - Language dependent
        - Text, speech
    - Language independent
        - Image, music
- Translingual transmedia information retrieval
    - Language translation
    - Media transformation

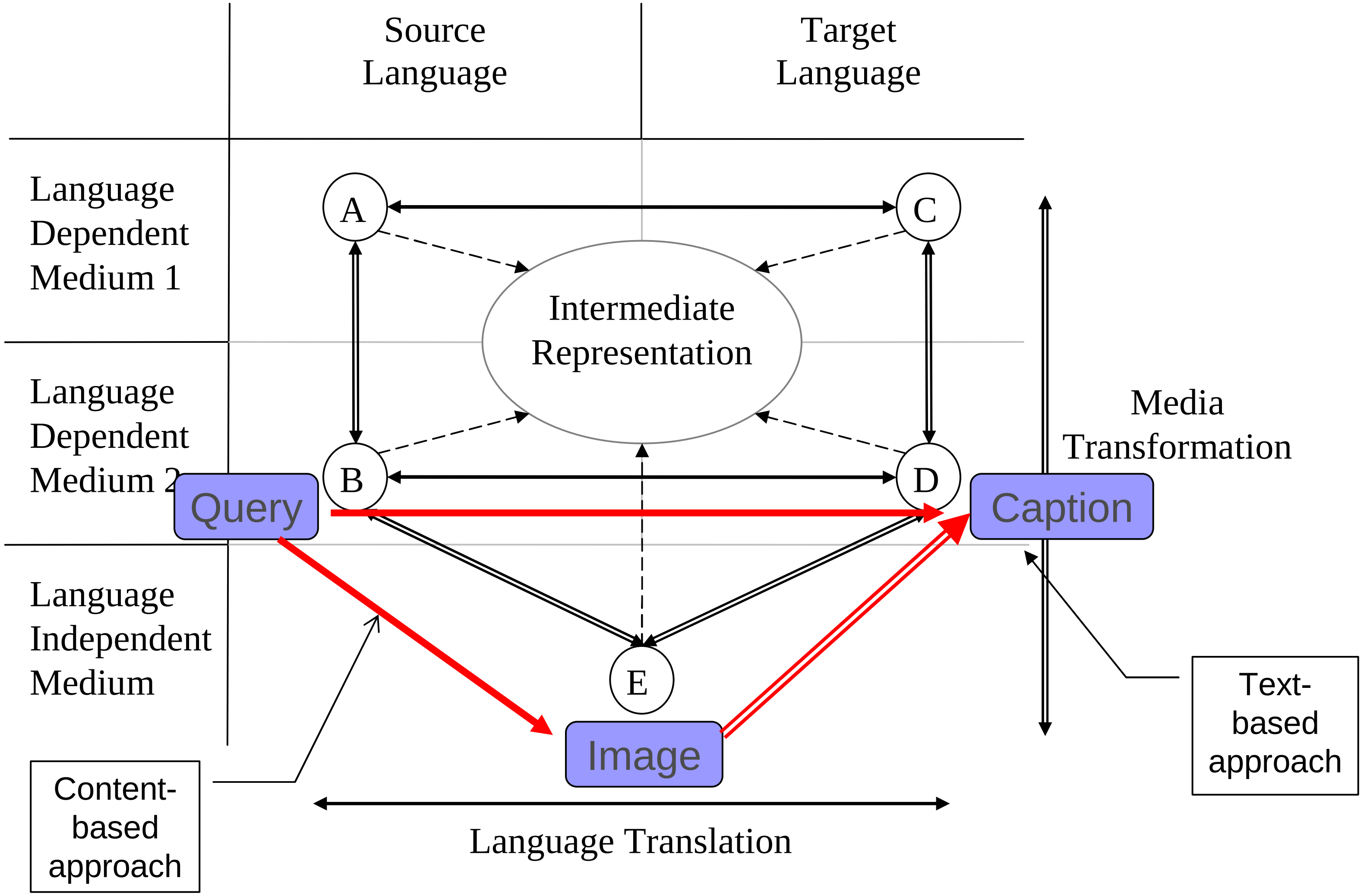

Figure 1. Media Transformation and Language Translation

Introduction

- We adopted text-based approaches in ImageCLEF 2003
  - Dictionary-based query translation
  - Unknown named entities were translated by similarity-based backward transliteration model
- This year, we explore the help of visual features to cross-language image retrieval

Introduction

- Transforms textual queries into visual representations
  - Model the relationships between text and images
  - Visual queries are constructed from textual queries using the relationships
- The retrieval results using textual and visual queries are combined to generate the final ranked list

Learning the relationships between images and text

- Learning the relationships between images and text from a set of images with textual descriptions
  - The textual description and image parts of an image is treated as an aligned sentence
  - The correlations between the textual and visual representations can be learned from the aligned sentences

Learning the relationships between images and text

- Use blobs to represent images
  - Images are segmented into regions using Blobworld
  - The regions of all images are clustered into 2,000 clusters by K-means clustering algorithm
  - Each cluster is assigned a unique label (blob token)
  - Each image is represented by the blobs that its regions belong to

Learning the relationships between images and text

- Correlation measurement

$$MI(x,y) = p(x,y) \times \log \frac{p(x,y)}{p(x)\,p(y)}$$

  - □ *p(x):* the occurrence probability of word x in text descriptions
  - □ *p(y)*: the occurrence probability of blob y in image blobs
  - □ *p(x,y)*: the probability that x and y occur in the same image

Generating visual query

- In a textual query, nouns, verbs, and adjectives are used to generate blobs
- For a word w_i, the top 30 blobs whose MI values with w_i exceed a threshold, i.e., 0.01, are selected
- The set of selected blobs is the generated visual query

Query translation

- Chinese query → English query
  - Segmentation, POS tagging, named entity identification
  - For each Chinese query term, find its translations by looking up a Chinese-English bilingual dictionary
  - First-two-highest-frequency
    - The first two translations with the highest frequency of occurrence in the English image captions are selected

Query translation

- Unknown named entities
  (Lin, W.C., Yang, C., and Chen, H.H. (2003). “Foreign Name Backward Transliteration in Chinese-English Cross-Language Image Retrieval”)
  - Apply the transformation rules to identify the name part and keyword part of a name
  - Keyword part → first-two-highest-frequency
  - Name part → similarity-based backward transliteration

Query translation

- Build English name candidate list
  - The personal names and the location names in the English image captions are extracted (3,599 names)
- For each Chinese name, 300 candidates are selected from the 3,599 English names using an IR-based candidate filter
- The similarities of the Chinese name and the 300 candidates are computed at the phoneme level
- The top 6 candidates with the highest similarities are considered as the translations of the Chinese name

Combining textual and visual information

- Two indices
  - Textual index ← English captions
  - Visual index ← image blobs
    - Treat blobs as a language in which each blob token is a word
    - Indexed by text retrieval system
- For each image, the similarity scores of textual and visual retrieval are normalized and combined using linear combination

Experiment

- IR system
  - Okapi system
  - BM25
- Learning correlations
  - English captions were translated into Chinese by SYSTRAN system
- 4 Chinese-English runs + 1 English monolingual run

Official results

Run	Merging Weight			Average Precision
	Textual Query	Example Image	Generated Visual Query	
NTU-adhoc-CE-T-W	1.0	-	-	0.3977
NTU-adhoc-CE-T-WI	0.9	-	0.1	0.3969
NTU-adhoc-CE-T-WE	0.7	0.3	-	0.4171
NTU-adhoc-CE-T-WEI	0.7	0.2	0.1	0.4124
NTU-adhoc-EE-T-W				0.5463

- Error of index
  - Long captions were truncated, thus some words were not indexed

Unofficial results

Run	Merging Weight			Average Precision
	Textual Query	Example Image	Generated Visual Query	
NTU-CE-T-W-new	1.0	-	-	0.4395
NTU-CE-T-WI-new	0.9	-	0.1	0.4409
NTU-CE-T-WE-new	0.7	0.3	-	0.4589
NTU-CE-T-WEI-new	0.7	0.2	0.1	0.4545
NTU-EE-T-W-new				0.6304

Discussion

- Textual query only
  - 69.72% of monolingual retrieval (CLEF2004)
  - 55.56% of monolingual retrieval (CLEF2003)
    - several named entities are not translated into Chinese in Chinese query set of ImageCLEF 2004
- Example image only
  - Average precision: 0.0523
  - The top one entry is the example image itself and is relevant to the topic except Topic 17 (the example image of Topic 17 is not in the pisec-total relevant set of Topic 17)

Discussion

- Generated visual query only
  - Average precision: 0.0103 (24 topics)
- The help of generated visual query is limit
  - The performance of image segmentation is not good enough
    - the majority of images are in black and white
  - The performance of clustering affects the performance of blobs-based approach

Discussion

- The quality of training data
  - English captions are translated into Chinese by MT system
  - Monolingual experiment (English)
    - Use English captions for training
    - Generate visual query from English query
    - English textual query + generated visual query
      - Average precision: 0.6561

Discussion

- Which word in a query should be used to generate visual query
  - Not all words are relative to the content of images or discriminative
  - Manually select query terms to generate visual query
    - Average precision: 0.0146 (18 topics)
    - textual query run + manually selecting run
      - Average precision: 0.4427

Discussion

- In some topics, the retrieved images are not relevant to the topics, while they are relevant to the query terms that are used to generate visual query
  - Topic 13: 1939年聖安德魯斯高爾夫球公開賽 (The Open Championship golf tournament, St. Andrews 1939)
  - “高爾夫球” (golf) and “公開賽” (Open Championship)

Top 10: the Open Championship golf tournament, but are not the one held in 1939

Conclusion

- We propose an approach that transforms textual queries into visual representations
- The retrieval results using textual and visual queries are combined
  - □ the performance is increased in English monolingual experiment
  - □ generated visual query has little impact in cross-lingual experiments

Conclusion

- Using generated visual query could retrieve images relevant to the query terms that the visual query is generated from
- How to select appropriate terms to generate visual query and how to integrate textual and visual information effectively will be further investigated

Thank you!