DCU Linking Runs at MediaEval 2012: Search and Hyperlinking Task

Shu Chen
CLARITY: Centre for Sensor Web Technologies / CNGL
Dublin City University
Dublin 9, Ireland
shu.chen4@mail.dcu.ie

Gareth J. F. Jones
CNGL, School of Computing
Dublin City University
Dublin 9, Ireland
gjones@computing.dcu.ie

Noel E. O'Connor
CLARITY: Centre for Sensor Web Technologies
Dublin City University
Dublin 9, Ireland
Noel.OConnor@dcu.ie

ABSTRACT

We describe Dublin City University (DCU)'s participation in the Hyperlinking sub-task of the MediaEval 2012 Search and Hyperlinking Task. Our strategy involves combining textual metadata, automatic speech recognition (ASR) transcripts, and visual content analysis to create anchor summaries for each video segment available for linking. Two categories of fusion strategy, score-based and rank-based methods, were used to combine scores from different modalities to produce potential inter-item links.

Keywords

Hyperlinking, multimedia search, data fusion, information retrieval

1. INTRODUCTION

Search and Hyperlinking was offered as a Brave New Task at MediaEval 2012 as an incubator for emerging research challenges. This paper presents details of DCU's participation in the multimedia hyperlinking task. We describe our automatic multimedia hyperlinking construction strategy based on fusion of metadata, automatic speech recognition (ASR) transcripts and visual content. The reminder of this paper presents the overall system design, data fusion strategy and the results of the evaluation experiment with analysis of the results achieved.

The paper is organized as following: Section 2 describes previous research on multimedia content analysis and data fusion evaluation, Section 3 describes our automatic hyperlink generation strategy, including multimedia content extraction and data fusion between different modalities, Section 4 gives our experimental results, and Section 5 concludes the paper.

2. RELATED WORK

There has been extensive research in multimedia content analysis in recent years. Work of particular relevance to the MediaEval 2012 Hyperlinking task includes the Video Google system based on content-based visual search in which the content is described at the level of keyframes [6]. The visual features of each keyframe were represented using the scale-invariant feature transform (SIFT). All extracted visual features were then transformed into visual keywords by applying a k-means clustering algorithm based on the extracted SIFT descriptors. Finally, a *tf-idf* algorithm calculated a document score to produce the ranked output results. The AXES project is similarly a content-based video search engine [1] which combines text data with visual content to implement a multimedia retrieval system.

Data fusion is an effective method to combine results from multiple information retrieval systems. A significant challenge for data fusion is score normalization to achieve the most effective combination. A number of methods are evaluated in [8] and [7], which provide detailed guidelines underpinning the design of our data fusion strategy.

3. DESCRIPTION OF LINK TASK

The dataset for the Linking Task is described in [2]. A total of 30 queries were provided where each one is associated with an anchor segment in the video. The objective of the task was to link each anchor segment to related video material in the test collection. To identify a linked segment, we extracted textual information from metadata and automatic speech recognition (ASR) transcripts for the anchor segment as a summary. Moreover, SIFT visual features were extracted for keyframes for each video shot.

3.1 Text Processing

Shots extracted automatically from the videos were used as our units for linking. For each shot we created a summary using both the video metadata, including title and short description, and ASR transcripts within the corresponding time interval. ASR transcripts were provided by LIUM [5], where one-best hypotheses was used, and LIMSI [3]. Apache Lucene is used to index and search the text information [1].

3.2 Visual Content Analysis

Visual content extraction followed the strategy introduced in [6]. We used VLFeat (version 0.9.14) to implement the SIFT algorithm to extract the visual features of each keyframe. A total of 26,0931 keyframes in total were provided with the dataset. We selected 9,034 of them as training set, which were randomly picked from the keyframe set of each video. After extracting SIFT descriptors from the training set, a k-means algorithm clustered the descriptor vectors to create visual words. The number of cluster centres, which experimentally should not be over 1000, as a limitation of Matlab

[1] `http://lucene.apache.org/`

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

processing capacity, was set to 500. The weight vector of each keyframe was calculated based on visual words and its own SIFT descriptor. We used *ffmpeg* to extract the keyframe at the media time of the anchor segment interval as a query image. Finally, a *tf-idf* algorithm was used to calculate the score for ranking of potential link videos.

3.3 Data fusion

Data fusion was used to calculate final scores based on the results from text processing and visual content analysis. We used linear combination data fusion to combine them together. Our fusion investigation and experiment involved two fusion approaches: score-based and rank-based.

The data fusion algorithm first normalizes scores from different modalities for all multimedia documents. For score-based data fusion, we used the following formula for normalization:

$$score_{normalized} = \frac{s_i - s_{min}}{s_{max} - s_{min}} \quad (1)$$

s_{min} and s_{max} are used to denote the minimal and maximal score over all multimedia documents respectively, while s_i is the raw score of document at rank i. The normalization formula in rank-based method follows the binary logistic regression model introduced in [7]:

$$score_{normalized} = \frac{1}{1 + e^{-a-b\ln r_i}} \quad (2)$$

Here r_i is the rank of document i in different modalities. a and b are two parameters whose details are described in [4]. In our experiment, we provide a simple solution on setting the value of coefficients as $a = 1$, and $b = -1$.

The linear combination fusion formula was the same for both the score-based method and rank-based method, as follows:

$$s_f(d_i) = w_1 \cdot s_t(d_i) + w_2 \cdot s_v(d_i) \quad (3)$$

In equation 3, $s_t(d_i)$ means the normalized score of document d at rank i from text processing result, while $s_v(d_i)$ means the normalization result of visual content analysis. The coefficients w_1 and w_2 are both set as 0.5, meaning that text and visual content will equally influence fusion result.

4. EVALUATION RESULTS

Table 1 shows the details of our 6 submitted runs, and Table 2 shows mean average precision (MAP) value of each run. Our best result of 0.276 was obtained for run 2. We can observe that: (1) the LIUM transcripts perform better than the LIMSI transcripts, (2) rank-based fusion achieves better results. The reason is only top 10 videos are involved in evaluation, and a rank-based fusion strategy based on non-linear formula gives a much higher score to the videos ranked relatively higher.

5. CONCLUSIONS

This paper presented details of DCU's participation in the Linking task of MediaEval 2012. We used two different features to identify anchors to produce potential inter-item links: textual content from metadata and ASR transcripts, and content-based visual features. Furthermore, two fusion methods, rank-based and score-based were investigated. According to our evaluation results, LIUM transcripts give better results than LIMSI transcripts and the rank-based fusion strategy performs better than score-based fusion. In our future work, we will focus on how to utilize existing computer vision algorithms to improve link generation performance. We will seek to improve current visual content analysis program based on SIFT algorithm. Moreover, we believe the result of text indexing can be improved by involving named entity recognition process.

Run ID	Features Used	Acronym
run1	LIMSI, metadata	MSI_T
run2	LIUM, metadata	UM_T
run3	LIMSI, metadata, visual, score	MSI_T_V_S
run4	LIUM, metadata, visual, score	UM_T_V_S
run5	LIMSI, metadata visual, rank	MSI_T_V_R
run6	LIUM, metadata visual, rank	UM_T_V_R

Table 1: Run Details

RUN	MAP
MSI_T	0.197
UM_T	0.276
MSI_T_V_S	0.176
UM_T_V_S	0.223
MSI_T_V_R	0.187
UM_T_V_R	0.234

Table 2: Evaluation Results (MAP)

6. ACKNOWLEDGMENTS

This work is funded by the European Commission's Seventh Framework Programme (FP7) as part of the AXES project (ICT-269980).

7. REFERENCES

[1] S. Chen, K. McGuinness, R. Aly, N. O'Connor, and F. de Jong. The AXES-lite video search engine. In *Proceedings of WIAMIS 2012*, pages 1–4, May 2012.

[2] M. Eskevich, G. J. F. Jones, S. Chen, R. Aly, R. Ordelman, and M. Larson. Search and Hyperlinking Task at MediaEval 2012. In *MediaEval 2012 Workshop*, 2012.

[3] L. Lamel and J.-L. Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing (LNCS 5221)*, pages 4–15. Springer-Verlag, 2008.

[4] A. Le Calvé and J. Savoy. Database merging strategy based on logistic regression. *Information Processing Management*, 36(3):341–359, 2000.

[5] A. Rousseau, F. Bougares, P. Delĺeglise, H. Schwenk, and Y. Estĺĺv. LIUM's systems for the IWSLT 2011 Speech Translation Tasks. In *Proceedings of IWSLT 2011I*, 2011.

[6] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings of ICCV 2003*, pages 1470–1477, 2003.

[7] S. Wu. Linear combination of component results in information retrieval. *Data Knowledge Engineering*, 71(1):114–126, 2012.

[8] S. Wu, F. Crestani, and Y. Bi. Evaluating score normalization methods in data fusion. In *Proceedings of AIRS 2006*, pages 642–648. Springer-Verlag, 2006.