DCU Linking Runs at MediaEval 2014: Search and Hyperlinking Task

Shu Chen
INSIGHT Research Center / 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
INSIGHT Research Center
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 Search and Hyperlinking task MediaEval 2014. The investigation focuses on how to efficiently identify target segments in a large BBC TV dataset. In our submission, Linear Discriminant Analysis is used to estimate fusion weights for multimodal features with the objective of improving the reliability of creating the potential links.

1. INTRODUCTION

For our participation in the Hyperlinking sub-task of Search and Hyperlinking of Television Content in MediaEval 2014, we developed a new strategy of target segment determination based on speaker identification. Linear Discriminant Analysis was used to estimate fusion weights. The paper is organized as follows: Section 2 describes a new strategy to determine target segments, Section 3 describes fusion weight determination based on Linear Discriminant Analysis, Section 4 gives our experimental results, and Section 5 concludes the paper.

2. SPEAKER-BASED TARGET SEGMENT DETERMINATION

This is the third year of the MediaEval Search and Hyperlinking task [8, 7]. The previous editions showed that state-of-the-art IR techniques can be applied to multimodal hyperlinking. However, identifying effective target segments is still an open issue. According to [7], a target segment should be moderate in length 10 to 120 seconds. In this paper, a simple and efficient target segment determination algorithm was developed by using the speaker identification from the LIMSI transcript [10].

Similar to searching model, the retrieval list of target segments are determined by the relevance to the query anchor. Existing researches [5, 3] use a sliding window to cover a variety of identifying target segments. We annotate that a target segment is a collection of multimodal features with time stamps, and those containing relevant information should be allocated a higher rank. As a result, we identify target segments as following. 1) Separate a video into a number of clips. The separating benchmark is defined according to the multimodal feature distribution. Each separated clip is defined as a seed segment. 2) Each seed segment increases the size of itself by merging adjacent segments. It stops when the length of newly merged segment reaches the target segment standard defined in [7]. 3) Repeat step 2 until all the seed segments have been expanded. 4) Identify each expanded segment as the target segment.

The kernel of the algorithm is how to determine the seed segments in a video and how to define the benchmark to merge adjacent segments. Existing researches purpose that the speech transcript often plays an important role in effective hyperlinking [5, 3]. Our investigation is based on LIMSI transcripts. Identifying a seed segment involves the transcript and speaker analysis. We assume that a group of spoken words, if presented by the same speaker, is content related. LIMSI transcript provides speaker information with each sentence. Define each transcript sentence as sen_i and its speaker ID as spk_i. Each video V_j is presented as a vector $V_j = [sen_1, sen_2, ...sen_i]$. If k continuous sentences $[sen_n, sen_{n+1}, ...sen_{n+k}]$ satisfy $spk_n = spk_{n+1} = ..spk_{n+k}$, they are merged to create a new seed segment sd_l, whose duration $|sd_l|$ is determined by the start time of sen_n and end time of sen_{n+k}. Our algorithm traverses all sentences to convert V_j into a new vector of seed segments $[sd_1, sd_2, ...sd_l]$. We assume that two transcripts presented by different speakers, if close enough, are possible to relate to the same topic. As a result, our algorithm to create target segments is defined as:

- Step 1: For each sd_i in V_j, compare the two nearby speech transcripts. Merge the one with lower time interval to create a new target segment mg_i. In this paper, the time interval threshold is set to be 10 seconds, following the minimum length of a target segment.
- Step 2: Continue step 1 until the length of merged segment mg_i is longer than 2 minutes.
- Step 3: Save the merged segment mg_i and process the sd_{i+1} in V_j.
- Step 4: Finally all merged segments $[mg_1, mg_2, ...mg_k]$ are regarded as potential target segments.

3. ESTIMATE FUSION WEIGHTS USING LINEAR DISCRIMINANT ANALYSIS

Linear Discriminant Analysis (LDA) algorithm [9]. was used to estimate linear fusion weights. The linear combination in LDA $w^T x$ can determine the fusing weights for

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

Table 1: MAP evaluation on DCU Hyperlinking Runs

Evaluation	RUN 1	RUN 2
MAP	0.0791	0.0430
P@5	0.2800	0.1867
P@10	0.2800	0.1667
P@20	0.1850	0.1000
MAP_bin	0.0707	0.0415
MAP_tol	0.0397	0.0282

the multimodal features. We define a training group as $|X|$ and its corresponding sample groups as $|Y|$, where $y_i \in |Y|$ could be 0 or 1, meaning that the video segment to a specific query anchor can be either related or non-related. Therefore, the training data was separated into two classifications: $|X_1|$ where $x \in X_1$ satisfies $p(x_i|y_i = 0)$, and $|X_2|$, where $x \in X_2$ satisfies $p(x_i|y_i = 1)$. Assuming both dataset follows the normal distribution, we have the means μ_1, μ_2, the covariance Σ_1 and Σ_2. The ratio between them was defined as the class variance S_b, the within class variance S_w, and w as the vector of fusing weights. S_w and S_b can be defined as:

$$S_b = (w \cdot \mu_1 - w \cdot \mu_2)^2 \tag{1}$$

$$S_w = (w^T \cdot \Sigma_1 \cdot w + w^T \cdot \Sigma_2 \cdot w) \tag{2}$$

A linear combination was calculated by maximizing the criterion of between class variance and within class variance. Define the criterion c as following:

$$c = \frac{w \cdot (\mu_1 - \mu_2)^2}{w^T \cdot (\Sigma_1 + \Sigma_2)} \tag{3}$$

According to [2], a maximum separation can be achieved by maximizing c by making the weight vector to follow :

$$w \propto (\Sigma_1 + \Sigma_2)^{-1} \cdot (\mu_1 - \mu_2) \tag{4}$$

4. EXPERIMENT

The experiment data is introduced in MediaEval 2014 overview paper [6]. Our experiment is constructed based on only LIMSI transcripts. Two runs are investigated to evaluate the performance of speaker-based target segment determination and fusion weight estimation using LDA. All 2 runs creates the target segments using speaker-based determination algorithm. RUN 1 involves the late fusion model on metadata and LIMSI transcripts. RUN 2 involves only LIMSI transcripts. The fusion equation is defined as following:

$$\text{FusionScore} = w_m \cdot \text{MetadataScore} + w_t \cdot \text{LIMSIScore} \tag{5}$$

In RUN 1, the linear fusion weights w_m and w_m are determined by LDA algorithm. The hyperlinking retrieval on metadata and LIMSI transcripts are processed using *TD-IDF* model on text features. Lucene 4.9.0 is applied to implement indexing and searching. The details are provided in [4]. The retrieval model in RUN 2 uses Lucene implementation as well.

Table 1 shows the Mean Average Precision (MAP) value of each runs. The evaluation metrics is defined in [1]. The experiment reveals that LDA estimation on linear fusion work could improve the hyperlinking quality, as presented in Table 1 that RUN 1 has an overall increase on MAP, MAP_bin and MAP_tol.

5. CONCLUSIONS

This paper presented details of DCU's participation in the TV Data Hyperlinking task of MediaEval 2014. The evaluation shows that LDA algorithm can improve the hyperlinking performance by estimating multimodal fusion weights. Our future work will be placed on comparing the speaker-based target segment determination with other strategies described in previous MediaEval research [8, 7].

6. ACKNOWLEDGEMENT

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

7. REFERENCES

[1] R. Aly, M. Eskevich, R. Ordelman, and G. J. F. Jones. Adapting Binary Information Retrieval Evaluation Metrics for Segment-based Retrieval tasks. 2013.

[2] G. Balakrishnama. Linear Discriminant Analysis - A Brief Tutorial, 1998.

[3] S. Chen, M. Eskevich, G. J. F. Jones, and N. E. O'Connor. An Investigation into Feature Effectiveness for Multimedia Hyperlinking. In *MultiMedia Modeling*, pages 251–262. Dublin, Ireland, 2014.

[4] S. Chen, G. J. F. Jones, and N. E. O'Connor. DCU Linking Runs at Mediaeval 2013: Search and Hyperlinking Task. In *MediaEval 2013 Workshop*, Barcelona, Spain, 2013.

[5] 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, Dublin, Ireland, 2012.

[6] M. Eskevich, R. Aly, R. Ordelman, D. N. Racca, S. Chen, and G. J. F. Jones. The Search and Hyperlinking Task at Mediaeval 2014. In *In Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, 2014.

[7] M. Eskevich, G. J. F. Jones, S. Chen, R. Aly, and R. Ordelman. Search and Hyperlinking Task at MediaEval 2013. In *MediaEval 2013 Workshop*, Barcelona, Spain, 2013.

[8] 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*, Pisa, Italy, 2012.

[9] R. A. FISHER. The Use of Multimple Measurements in Taxonomic Problems. pages 179–188, 1936.

[10] 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.