FXPAL Experiments for TRECVID 2016 Video Hyperlinking

Chidansh Bhatt and Matthew Cooper
FX Palo Alto Laboratory
Palo Alto, CA 94304 USA
{ bhatt, cooper }@fxpal.com

Abstract

This is a summary of our participation in the TRECVID 2016 video hyperlinking task (LNK). We submitted four runs in total. A baseline system combined on established vectorspace text indexing and cosine similarity. Our other runs explored the use of distributed word representations in combination with fine-grained inter-segment text similarity measures.

1 Introduction

The infrastructure for aggregating and distributing multimedia and video content has advanced rapidly in recent decades. However, available tools for efficiently navigating and accessing this content have not kept pace. The video hyperlinking (LNK) task at TRECVID 2016 [1] aims to foster progress in tools for effectively accessing video content. The task begins with an anchor video segment. The goal is to produce a ranked list of relevant segments to the anchor. A critical challenge here is the inherently ambiguous nature of inter-segment relevance which may reflect any aspect of the anchor content.

The LNK task focuses on the blip.tv data set [14] which consists of roughly 11,000 user generated videos from the blip.tv site. This data set also includes automatic speech recognition transcripts produced by the LIMSI team [8], as well as various user tags and metadata associated with each video.

We have designed variants of a basic pipeline for processing the videos to examine two questions. The first question is the effect of segmentation on video hyperlinking. Here we compare two established methods, TextTiling [5] which computes a content-based segmentation of the text transcript using a bag of words model, and TopicTiling [13] which integrates a generative topic representation.

The second question concerns quantifying inter-segment similarity. Hierarchical methods including topic modeling [4] are ubiquitous in multimedia processing to enhance both scalability and performance. However in empirical validations, some latent variable methods have compared poorly with highly optimized vector space indexing [9]. One possible explanation for this is the loss of exact word matches incurred by projecting word vectors to lower dimensional representations. In our runs, we explore the use of a recently proposed word mover's distance for text similarity [7]. This method attempts to model inter-segment similarity based on inter-word semantics, and allows us to use modern distributed word representations without resorting to segment-level aggregation which may conceal directly matching words.

2 Technical details

In this section we describe technical elements of our submitted runs. As has been frequently observed in previous evaluations, text-based analysis has driven higher performing systems. The bulk of our processing is also based on the LIMSI ASR transcripts to represent the video.

2.1 Baseline methods

Our baseline run (**L_D_I16_M_TFIDF-COSINE-KNN-T8V2-META-FUSION**) leverages established methods in text processing and information retrieval and builds on a system submitted to MediaEval 2013 [2] with very minor modification in the re-ranking scheme.

As shown in the Fig. 1, topic segmentation was performed over LIMSI ASR transcripts using the TextTiling implementation in the NLTK toolkit [3]. Inter-segment text similarity between anchors and all topic segments is computed using a vector

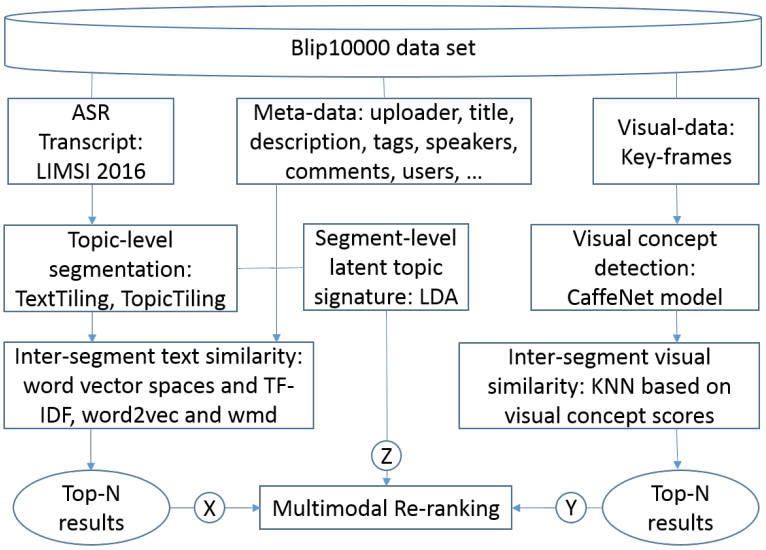

Figure 1: Overview of the proposed video hyperlinking (LNK) system.

space model and tf-idf weighting. We also considered the surrounding text from topic-segments that overlap with the anchor as well as the corresponding video-level meta-data (title, description, video uploader, tags, speakers, comments, users, etc.) to enrich segment similarity with context information. The most similar segments were found by cosine similarity.

For each segment, we generated a visual feature vector using the concepts with the highest scores from the keyframes of the segment. The visual concept detection scores per keyframe are extracted with pre-trained deep learning based image classification and annotation framework Caffe [6]. We employed the widely used AlexNet architecture trained on the ImageNet benchmark data set and concepts. Using k-nearest neighbor (KNN), we ranked all segments by decreasing visual similarity scores with the anchor. We generated the K nearest neighbors (segments) for each anchor using the Scikit-learn toolkit in Python, with a ball tree data-structure and Euclidean distance.

With a minor modification to multimodal reranking in [2], we integrate additional similarity measures based on topical information and uploader intent. Also, the weights are assigned dynamically for the fusion of all the similarity scores based on textual, visual, topical and user-intent features.

Finally, meta-data filtering is applied to generate the top 1000 submitted results for each run following guidelines received from the organizers of the task.

2.2 Transcript segmentation

We compare two approaches to segmenting the videos based on the LIMSI ASR transcripts. TextTiling [5] is a standard content-based text segmentation method and is used in our baseline run and one additional run. The second approach, TopicTiling [13] integrates topic modeling to enhance segmentation, and is used in our other two submitted runs. The TextTiling segmentation produced 174,747 segments for the corpus, while the TopicTiling produced 190,787 segments. There are total 1,981 ASR files with less than 150 words on which segmentation algorithms could not create any segments. The mean number of segments per video using TextTiling is 18 and using TopicTiling is 20.

To represent the anchors, we use the text in defined time interval of the transcript and augment it with adjacent text in neighboring (overlapping/containing) segments.

2.3 Distributed word representations

In the last several years, distributed word representations have been successfully developed and validated in a variety of text processing applications [10]. These dense vectors represent each word in a vocabulary such that words that frequently co-occur have similar (proximate) vector representations. The representations are learned to optimize the probability of word co-occurrence and thus encode semantics.

In three of our runs, we use standard dis-

tributed representations provided by Google [11] and learned from a large text corpus mined from Google News. These representations were used previously for video hyperlinking [12] with mixed results. One goal for our runs is to see if combining these representations with a fine grained inter-segment similarity measure that averts aggregation of the word vectors over the segment helps to improve performance.

2.4 Word mover's distance

Kusner *et al.* [7] introduced the word mover's distance (WMD) for comparing text documents for categorization and other tasks, largely in a nearest neighbors classification framework. The approach builds on the earth mover's distance which found widespread use in computer vision applications in years past. The distance measure compares two feature distributions and solves an optimization problem to most efficiently map one distribution to the other. This solution builds upon a "ground" distance that compares each possible pair of features. Consider two segments with word distributions $\mathbf{d}$ and $\mathbf{d}'$ over the n word vocabulary. d_i is the normalized frequency of word i in $\mathbf{d}$. Denote the cosine distance between individual word vectors w_1, w_2 by $c(w_1, w_2) = \langle w_1, w_2 \rangle / (\|w_1\| \cdot \|w_2\|)$. The WMD solves the linear program:

$$\begin{aligned} \min_{\mathbf{T} \geq 0} \quad & \textstyle\sum_{i,j=1}^{n} \mathbf{T}_{ij} c(w_i, w_j) \\ \text{subject to:} \quad & \textstyle\sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \;\; i = 1, \cdots, n \\ & \textstyle\sum_{i=1}^{n} \mathbf{T}_{ij} = d_j \;\; j = 1, \cdots, n \; . \end{aligned}$$

$\mathbf{T}_{ij}$ is the amount of word i in $\mathbf{d}$ that is mapped to word j in $\mathbf{d}'$ which incurs a proportional cost determined by the ground distance between the individual word vectors.

In the context of LNK, we represent two video segments by the collection of dense word vectors each segment contains, and their normalized frequencies. The inter-segment similarity is then the optimal transformation of one segment's word distribution in to the other's in terms of the inter-word cosine distances. Solving a linear programming problem to compare each possible pair of video segments is prohibitively expensive. We employ a greedy pruning scheme using approximation methods also described in [7]. We first use the cosine distance between standard word vectors to rank all segments in terms of similarity to the anchor and discard the least similar 50% of the segments. We then re-rank the remaining segments using the euclidean distance between the centroids of each segment's word vectors. Again, we discard half the remaining segments. Finally, we compute the WMD between the remaining segments and the anchor. These distances were in turn input to the multimodal re-ranking stage described in the next sub-section.

2.5 Multimodal Re-ranking

First, we trained the Latent Dirichlet Allocation (LDA) model with 100 topics on the segment corpus and anchors to generate their latent topic signature. Using union, intersection and difference of latent topic signature of the anchor and segment, we calculate their topical similarity and diversity reward score. Here, ratio of the length of latent topics intersection to the length of latent topics union represents amount of topical similarity between the anchor and the segment. Similarly, the ratio of the length of latent topics difference to the length of latent topics union represents amount of topical diversity. We combine the topical similarity and diversity score for the re-ranking. It will be useful to have at least 50% topic similarity and 50% topic diversity to consider significance of topical feature for re-ranking. Similarly, the anchor video uploader must be same as the segment video uploader or anchor video uploader should have commented on the segment video to consider significance of user-intent feature for re-ranking.

We re-ranked the top text-based results using fusion of similarity scores with weight $\mathbf{X}$ for textual, weight $\mathbf{Y}$ for visual, weight $\mathbf{Z}$ for topical and weight $\mathbf{W}$ for user-intent features. For our submissions, we chose $\mathbf{X} = 0.8$ and $\mathbf{Y} = 0.2$ whenever topical similarity score is below threshold (e.g., 0.25) and the segment video uploader do not match with anchor video uploader (or commentator). We chose $\mathbf{Z} = 0.1$ when topical similarity score is more than a threshold (e.g., 0.25) value. We assigned $\mathbf{W} = 0.1$ when there is a match between segment video uploader and anchor video uploader. Weight of textual and visual similarity will change to $\mathbf{X} = 0.6$ and $\mathbf{Y} = 0.3$ based on $\mathbf{Z} = 0.1$ or $\mathbf{W} = 0.1$.

Finally, we applied meta-data filtering to ignore the segments shorter than 10 seconds, divided larger segments into 2-minute segments and consider only non-overlapping segments to be valid submission according to the publicly available validation script provided by LNK task organizers.

Submitted run list						
#	Run Name	Text features	Segmentation	Similarity measure	Multi-modal re-ranking	Meta-data filtering
1	L_D_I16_M_W2V-WMD-TOPICTILING-KNN-T8V2-FUSION	word2vec	TopicTiling	WMD	yes	yes
2	L_D_I16_M_W2V-WMD-KNN-T8V2-META-FUSION	word2vec	TextTiling	WMD	yes	yes
3	L_D_I16_M_TFIDF-COSINE-KNN-T8V2-META-FUSION	TF-IDF	TextTiling	cosine	yes	yes
4	L_D_I16_M_W2V-WMD-TOPICTILING-NO-FUSION	word2vec	TopicTiling	WMD	no	yes

Table 1: Four systems submitted by team FXPAL for video hyperlinking (LNK) task evaluation.

3 Submitted Runs

The submitted runs appear in Table 1, including the segmentation method (TopicTiling or TextTiling) in combination with a text processing pipeline (TF-IDF with cosine similarity or word2vec with WMD). The second rightmost column indicates the use of multimodal re-ranking to validate each component for video hyperlinking.

Each of the runs is coded with a combination of letters indicating the information used by the system, following the TRECVID LNK task instructions, as follows: 'I16' for LIMSI transcripts 2016; 'L' for lexical cohesion segmentation; 'M' when metadata provided with the video; and 'D' when the visual features being used are derived using deep learning visual concept detection.

We look forward to revising this paper and providing additional analysis when the results and ground truth data become available.

References

[1] George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F. Smeaton, Georges Quéenot, Maria Eskevich, Robin Aly, and Roeland Ordelman. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *Proceedings of TRECVID 2016*. NIST, USA, 2016.

[2] Chidansh Bhatt, Nikolaos Pappas, Maryam Habibi, and Andrei Popescu-Belis. Idiap at mediaeval 2013: Search and hyperlinking task. In *MediaEval*. CEUR-WS.org, 2013.

[3] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[5] Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, March 1997.

[6] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[7] Matt J Kusner, Yu Sun, Nicholas I Kolkin, and Kilian Q Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pages 957–966, 2015.

[8] L. Lamel. Multilingual speech processing activities in quaero: Application to multimedia search in unstructured data. In *The Fifth International Conference Human Language Technologies - The Baltic Perspective*, 2012.

[9] Yue Lu, Qiaozhu Mei, and ChengXiang Zhai. Investigating task performance of probabilistic topic models: an empirical study of plsa and lda. *Information Retrieval*, 14(2):178–203, 2011.

[10] T Mikolov and J Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 2013.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. word2vec, 2014.

[12] Lei Pang and Chong-Wah Ngo. Vireo @ trecvid 2015: Video hyperlinking (lnk). In *Proceedings of TRECVID 2015*, 2015.

[13] Martin Riedl and Chris Biemann. How text segmentation algorithms gain from topic models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 553–557. Association for Computational Linguistics, 2012.

[14] Sebastian Schmiedeke, Peng Xu, Isabelle Ferrané, Maria Eskevich, Christoph Kofler, Martha A. Larson, Yannick Estève, Lori Lamel, Gareth J.F. Jones, and Thomas Sikora. Blip10000: a social video dataset containing spug content for tagging and retrieval. In *ACM Multimedia Systems Conference (MMSys 2013)*, 2013.