Idiap at MediaEval 2013: Search and Hyperlinking Task

Chidansh Bhatt
Idiap Research Institute
Martigny, Switzerland
cbhatt@idiap.ch

Nikolaos Pappas
Idiap and EPFL
Martigny, Switzerland
npappas@idiap.ch

Maryam Habibi
Idiap and EPFL
Martigny, Switzerland
mhabibi@idiap.ch

Andrei Popescu-Belis
Idiap Research Institute
Martigny, Switzerland
apbelis@idiap.ch

ABSTRACT

The Idiap system for search and hyperlinking uses topic-based segmentation, content-based recommendation algorithms, and multimodal re-ranking. For both sub-tasks, our system performs better with automatic speech recognition output than with manual subtitles. For linking, the results benefit from the fusion of text and visual concepts detected in the anchors.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems

Keywords

Topic segmentation; video search; video hyperlinking.

1. INTRODUCTION

This paper outlines the Idiap system for the MediaEval 2013 Search and Hyperlinking task [3]. The search sub-task required finding a determined segment of a show (from 1260 hours of broadcast TV material provided by BBC) based on a query that had been built with this "known item" in mind. The hyperlinking sub-task required finding items from the collection that are related to "anchors" from known items. We propose a unified approach to both sub-tasks, based on techniques inspired from content-based recommender systems [6], which provide the most similar segments to a given text query or to another segment, based on words. For hyperlinking, we also use the visual concepts detected in the anchor in order to rerank answers based on visual similarity.

2. SYSTEM OVERVIEW

The Idiap system makes use of three main components, shown at the center of Fig. 1. We generate the data units, namely topic-based segments, from the subtitles or the ASR transcripts (either from LIMSI[4] or from LIUM[7]) using TextTiling in NLTK [1]. For search, we compute word-based similarity (from transcript and metadata) between queries and all segments in the collection, using a vector space model based and TF-IDF weighting. Similarly, for hyperlinking, we first rank all segments based on similarity with the anchor. In addition, we use the visual concept detection provided by the organizers (key frames from Technicolor[5], concepts detected by Visor[2]) to generate a score matrix and then the list of nearest neighbors. Scores from text and visual similarity are fused to re-rank final linking results.

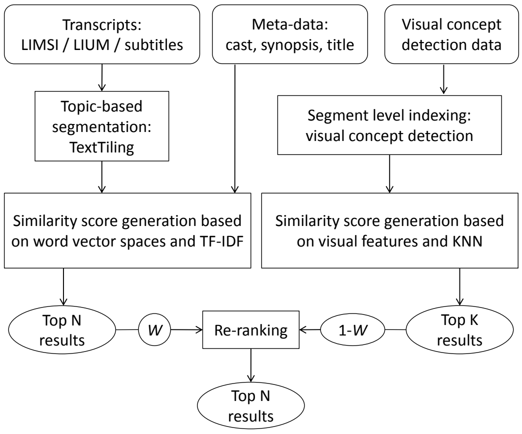
Transcripts: LIMSI / LIUM / subtitles
Meta-data: cast, synopsis, title
Visual concept detection data
Topic-based segmentation: TextTiling
Segment level indexing: visual concept detection
Similarity score generation based on word vector spaces and TF-IDF
Similarity score generation based on visual features and KNN
Top N results
W
Re-ranking
1-W
Top K results
Top N results

Figure 1: Overview of the Idiap system.

Topic segmentation was performed over subtitles or transcripts using TextTiling as implemented in the NLTK toolkit [1]. Topic shifts are based on the analysis of lexical co-occurrence patterns, computed from 20-word pseudo-sentences. Then, similarity scores are assigned at sentence gaps using block comparison. The peak differences between the scores are marked as boundaries, which we fit to the closest speech segment break. The total number of segments for subtitles / LIMSI / LIUM is respectively 114,448 / 111,666 / 84,783, with average segment sizes of 53 / 53 / 68 seconds and a STD of 287 / 68 / 64 s. The longer size of the LIUM segments and the large variability of subtitles segments should be noted. We found some mismatches between the durations in metadata files and the timing found in the subtitle and LIMSI transcript files (488 resp. 956 videos) and discarded the corresponding segments.

Segment search was performed by indexing the text segments in a word vector space with TF-IDF weights, representing each textual query (and words from the "visual cues") into the same space, and retrieving the most similar segments to the query using cosine similarity. We first tokenized the text and removed stop words. We tested several parameters on the small development set with the LIMSI transcript: the order of n-grams (1, 2, or 3) and the size of the vocabulary (10k, 20k, 30k, 40k, 50k words). The best

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

scores (ranks of known items in the results) were reached for 50k words with unigrams, bigrams and trigrams. With these features, we found on the development set that the LIMSI transcript performed best, followed by LIUM, LIUM with metadata, and subtitles. We submitted 4 runs for the search sub-task: 3 were based on each transcript/subtitle words, and the fourth used the LIUM transcript but appended to each segment the words from the metadata (cast, synopsis, series, and episode name).

For ***hyperlinking segments from anchors***, indexing is performed as above, though using only unigrams and a vocabulary of 20,000 words. For scenario A (anchor information only), we extended the anchor text with text from segments containing/overlapping the anchor boundaries. For the scenario C, we considered the text within the start time and end time of the provided know-item, along with text from segments containing/overlapping the know-item boundaries. We enriched the subtitle/ASR text using the textual metadata (title, series, episode) and webdata (cast, synopsis). The segments and anchors were indexed into a vector space with TF-IDF weights, and the top N most similar segments were found by cosine similarity.

Then, we reranked results based on visual feature similarity, using the visual concept detection scores per keyframe (provided by the organizers). Keyframes were first aligned to topic-based segments using shot information [5], with an average of 5 keyframes per segment. Similarly, this was performed for the anchors (8 frames) and anchors + contexts (55 frames). For each segment, we generated a visual feature vector using the concepts with the highest scores from the keyframes of the segment. Using KNN, we ranked all segments by decreasing similarity to an anchor. Then, we reranked text-based results using visual information, respectively with weight W (for text) and $1 - W$ (for visual). We chose $W = 0.8$ in the case of subtitles (assuming a higher accuracy) and $W = 0.6$ for transcripts. Finally, we ignored segments shorter than 10 s and chunked larger segments into 2-minute segments. We submitted 3 runs: two with the subtitle words (scenarios A and C) and one with the LIMSI transcript (C).

3. RESULTS

The official ***search results*** (Table 1) show the same ranking as on the development set. Using LIMSI transcript outperforms the LIUM one, which is not helped by metadata (this might be due to low-frequency features in the metadata). Surprisingly, subtitles yield the lowest scores.

The overall low scores (esp. on mGAP and MASP) could be due to the short average size of our segments, which were not calibrated to match the average size of known items.

Analyzing results per query, in 12 out of 50 test queries our best run gets the known item in the top 10 answers. These queries are not "easy", as they vary across runs (with exceptions like item_18). On the contrary, for 14 queries the known-item is not found among the top 1000 results.

Submission	MRR	mGAP	MASP
Subtitles	0.064	0.044	0.044
LIUM + Meta	0.085	0.054	0.053
LIUM	0.090	0.058	0.057
LIMSI	**0.110**	**0.060**	**0.060**

Table 1: Official Idiap results for the search task.

The ***linking runs*** (Table 2) were scored after the deadline, separately from the other submissions, due to a time conversion problem undetected on submission. Here also, using the LIMSI transcript (first line) outperforms subtitles. This might be due to the higher weight of visual concepts when using transcripts (0.4) vs. subtitles (0.2).

When using subtitles (2nd and 3rd rows), a higher MAP value was found when context was *not* used, indicating that this might actually add noise, esp. with our strategy of extending context boundaries to the closest segments. Therefore, we hypothesize that using LIMSI transcripts for the A task would lead to an even higher MAP.

The precision of our system increases from top 5 to top 10 and 20. Our best system reaches close-to-average MAP on anchors 31, 32 and 39 (respectively 0.49, 0.39 and 0.38), while the MRR of the corresponding search queries (item_23 for 31 and 32, item_25 for 39) is close to zero. This is an indication that the visual features may be helpful.

Submission	P_5	P_10	P_20	MAP
L_V_M_O_T6V4_C	**0.673**	**0.687**	**0.615**	**0.576**
S_V_M_O_T8V2_A	0.460	0.540	0.593	0.542
S_V_M_O_T8V2_C	0.493	0.520	0.523	0.492

Table 2: Idiap results for hyperlinking: precision at top 5, 10 and 20, and mean average precision.

4. ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation (AROLES project n. 51NF40-144627) and by the European Union (inEvent project FP7-ICT n. 287872). We would like to thank Maria Eskevich and Robin Aly for their valuable help with the task.

5. REFERENCES

[1] S. Bird. NLTK: the Natural Language Toolkit. In *COLING/ACL Interactive Presentations*, Sydney, 2006.

[2] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, pages 76.1–76.12, 2011.

[3] M. Eskevich, G. J. Jones, S. Chen, and R. Aly. The Search and Hyperlinking task at MediaEval 2013. In *MediaEval 2013*, Barcelona, 2013.

[4] J.-L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37(1-2):89–108, 2002.

[5] A. Massoudi, F. Lefebvre, C.-H. Demarty, L. Oisel, and B. Chupeau. A video fingerprint based on visual digest and local fingerprints. In *ICIP*, Atlanta, GA, 2006.

[6] N. Pappas and A. Popescu-Belis. Combining content with user preferences for ted lecture recommendation. In *11th Int. Workshop on Content Based Multimedia Indexing (CBMI)*, Veszprém, 2013.

[7] H. Schwenk, P. Lambert, L. Barrault, C. Servan, H. Afli, S. Abdul-Rauf, and K. Shah. LIUM's SMT machine translation systems for WMT 2011. In *6th Workshop on Statistical Machine Translation*, pages 464–469, Edinburgh, 2011.