# Hierarchical Topic Models for Language-based Video Hyperlinking

Anca-Roxana Simon, Rémi Bois, Guillaume Gravier, Pascale Sébillot, Emmanuel Morin, Sien Moens

# Hierarchical Topic Models for Language-based Video Hyperlinking

Anca-Roxana Simon
Univ. Rennes 1
IRISA & Inria Rennes
anca.simon@irisa.fr

Rémi Bois
CNRS
IRISA & Inria Rennes
remi.bois@irisa.fr

Guillaume Gravier
CNRS
IRISA & Inria Rennes
guig@irisa.fr

Pascale Sébillot
INSA Rennes
IRISA & Inria Rennes
pascale.sebillot@irisa.fr

Emmanuel Morin
Univ. Nantes
LINA
emmanuel.morin@univ-nantes.fr

Sien Moens
KU Leuven
Computer Science Dpt.
sien.moens@cs.kuleuven.be

## ABSTRACT

We investigate video hyperlinking based on speech transcripts, leveraging a hierarchical topical structure to address two essential aspects of hyperlinking, namely, serendipity control and link justification. We propose and compare different approaches exploiting a hierarchy of topic models as an intermediate representation to compare the transcripts of video segments. These hierarchical representations offer a basis to characterize the hyperlinks, thanks to the knowledge of the topics who contributed to the creation of the links, and to control serendipity by choosing to give more weights to either general or specific topics. Experiments are performed on BBC videos from the Search and Hyperlinking task at MediaEval. Link precisions similar to those of direct text comparison are achieved however exhibiting different targets along with a potential control of serendipity.

## Categories and Subject Descriptors

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

## General Terms

Agorithms, Experimentation

## Keywords

Multimedia, NLP, hyperlinking, topic modeling

## 1. INTRODUCTION

Automatic generation of hyperlinks in videos is a subject with growing interest, as evidenced by the success of recent international benchmarks on the subject within the Mediaeval initiative and TRECVid, e.g., [4]. The goal of video hyperlinking within a collection is to create links between fragments of the collection, starting from an initial segment called *anchor* (by analogy with anchors in web hyperlinks). An anchor is a short video segment, selected by a human, from which one wants related content in the collection. In this context, the hyperlinking task boils down to selecting target segments for a given anchor. The relation between anchor and target is not defined beforehand and can range from similar information (almost same content) to new, surprising, information. Relevance is judged post-hoc, via crowd-sourcing in the framework of lab evaluations.

The hyperlinking task has been mostly handled as an information retrieval task after segmentation of the videos. In this two-step scheme, the first step consists in defining potential target segments. A number of strategies have been proposed to this end, e.g., fixed-length segmentation [5], shot [10] or topic [6] segmentation, or pseudo-sentences taken from the automatic transcripts [7]. Following the segmentation step, relevant target segments are selected for a given anchor. For this target selection step, most approaches rely on pairwise content-based proximity exploiting either text—subtitles or automatic transcripts—or visual content. Text and visual content may be enriched with additional information, e.g., named entities [3, 9], metadata [8], visual concepts [1] or prosodic information [5]. In most cases, a vector space model is used to represent the content of the anchor and the target along with standard similarity measures.

Human-based evaluations of hyperlinking systems done within MediaEval have revealed that the best systems were those that proposed targets very similar to the anchor. In [9], the authors observed that in some cases, having targets about the same people as in the anchor, though in different circumstances, was not found relevant. While this observation justifies the use of direct content-based comparison, the interest of the targets proposed in terms of informativeness and serendipity remains very limited. As a potential solution to this problem, we investigate transcript-based indirect content comparison mediated via a hierarchical topical structure. The key idea is to have a fine-grain control on the topics that are highlighted in the targets proposed for a given an anchor. The topical structure is composed of topics at different levels of granularity, from general to specific.

*SLAM'15,* October 30, 2015, Brisbane, Australia.
.

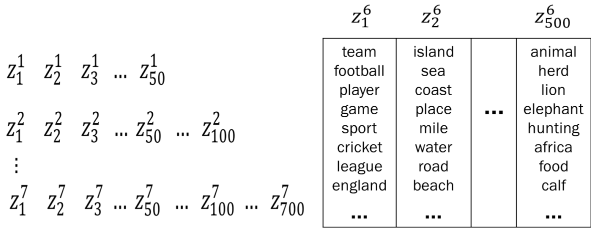
$z_1^6$ $z_2^6$ $z_{500}^6$
$z_1^1$ $z_2^1$ $z_3^1$ ... $z_{50}^1$
$z_1^2$ $z_2^2$ $z_3^2$ ... $z_{50}^2$ ... $z_{100}^2$
⋮
$z_1^7$ $z_2^7$ $z_3^7$ ... $z_{50}^7$ ... $z_{100}^7$ ... $z_{700}^7$
team football player game sport cricket league england ...
island sea coast place mile water road beach ...
...
animal herd lion elephant hunting africa food calf ...

**Figure 1: Representation of the independent topic models for $K = 50 \rightarrow 700$.**

A first advantage of this structure over the direct bag-of-words representation is the ability to link related anchor-target pairs that do not share a consistent part of vocabulary. Additionally, the hierarchical topic structure of the model, linking fine-grain topics with coarse-grain themes, allows for an increased control over serendipity and has the potential to explain the nature of the link (i.e., why is this linked proposed?).

## 2. LINKING WITH TOPIC MODELS

The topical structure that we propose relies on a hierarchy of topics from coarse-grain topics at the top to fine-grain ones at the bottom. The different topics are obtained independently at each level of the hierarchy based on the latent Dirichlet allocation (LDA) probabilistic model [2]. Three variants are considered. In the first ones, topics at each level of the hierarchy are considered as independent. In the last two, links are established between topics at consecutive levels in the hierarchy, thus forming a tree structure of topics.

### 2.1 Building LDA topic models

At each level of the hierarchy, a number of LDA topic models are estimated based on the transcripts of the collection of videos. In this model, each transcript is represented as a mixture of $K$ latent topics, where each latent topic is characterized by a probability distribution over the set of words in the transcript (the vocabulary). LDA models were estimated using Gibbs sampling with standard values for the hyperparameters $\alpha = 50/K$. To define the various levels of the topic hierarchy, we trained model for different numbers of latent topics, namely $K \in \{50, 100, 150, 200, 300, 500, 700, 1000, 1500, 1700\}$. This range of values for the number of topics was chosen to obtain topics that go from being general to highly specific and to have a large number of granularity levels for a better control of link creation. At each level $l$, a word distribution $z_i^l$ is obtained for each topic $i \in [1, K_l]$, where $K_l$ is the number of latent variables at level $l$ ($K_1 = 50, \ldots, K_{10} = 1700$). The result of the process of building LDA topic models at 10 levels is illustrated in Figure 1, where the most likely words for some topics obtained with $K_6 = 500$ are given on the right side. Clearly, the first topic is about sport while the second one is about the sea.

### 2.2 Independent topic levels

The simplest model that can be built from the set of topic models $z_i^l$ is to consider each level independently. Segment probabilities given by topic models at different levels are then used, either independently or in combination, to compare an anchor with a target segment.

Given a segment $x$, the word distribution $z_i^l$ for the $i$-th topic at level $l$ enables the computation of the probability that $x$ was obtained from $z_i^l$ according to

$$p(x|z_i^l) = \sqrt[n_x]{\prod_{j=1}^{n_x} p(w_j|z_i^l)} \ , \tag{1}$$

where $n_x$ is the size of the vocabulary in $x$ and $w_j$ is the $j$-th word in $x$. The word probabilities are given by

$$p(w_j|z_i^l) = \frac{n(z_i^l, w_j) + \beta}{\sum_{k=1}^{n} n(z_i^l, w_k) + \beta|V|} \ . \tag{2}$$

These probabilities are estimated on the entire collection, with $n(z_i^l, w_j)$ being the number of times topic $z_i^l$ was assigned to word $w_j$ occurring at a certain position in the training documents. The denominator thus corresponds to the total number of words assigned to topic $z_i^l$. $V$ represents the number of distinct words in the entire vocabulary and $\beta$ is the Dirichlet prior. Based on (1), we represent $x$ at level $l$ by the vector gathering topic-wise probabilities of $x$, i.e.,

$$x_l = (p(x|z_1^l), p(x|z_2^l), ..., p(x|z_{K_l}^l)) \ . \tag{3}$$

For efficiency reasons, we use a sparse version of $x_l$, zeroing all but the 10 top-scoring topics. The discarded probability mass is redistributed evenly on the top-10 topics.

Comparing two segments $x$ and $y$ is done via the respective representations $x_l$ and $y_l$ according to

$$S_1(x, y) = -\sum_l \alpha_l \log(x_l y_l) \ . \tag{4}$$

The weights $\alpha_l$ allows to control the relative weights of the topic levels, for instance, to select one single level or to emphasize fine-grain levels over general topics. We compare three weighting variants: equal importance to all topics ($\mathrm{IT}_{\mathrm{Comb}=}$), increasing importance ($\mathrm{IT}_{\mathrm{Comb}<}$) as going from general topics to specific ones and conversely ($\mathrm{IT}_{\mathrm{Comb}>}$).

### 2.3 Tree-structured topic levels

Exploiting explicit links between topics at different levels of the hierarchy—e.g., meronymy, hyperonymy—appears as appealing for a better control of the diversity of the targets and of the relation between anchor and target. We thus propose two strategies to turn the independent 10 levels of LDA models into a tree structure.

A straightforward way to build a tree structure exploits the similarity between topics at two consecutive levels, where the similarity between topic $i$ at level $l$ and topic $j$ at $l+1$ is given by $-\log\left(z_i^l z_j^{l+1}\right)$. The tree is obtained by connecting a topic to the most similar topic at the previous level. Formally, $z_j^{l+1}$ is linked to $z_k^l$ such that $k = \mathrm{argmin}_i \log\left(z_i^l z_j^{l+1}\right)$. We call such links 'hard' links, meaning that every node as a unique parent (except at $l = 1$) but not necessarily a sibling or a child.

The 'hard link' tree-structured (HLT) hierarchy of topics is used to define a new representation of an anchor $x$ depicting the path in the tree that ends at $l = 10$ with the best matching fine-grain topic. For an anchor segment $x$, we first identify the best matching topic at the lowest level, i.e., $k = \mathrm{argmax}_j\, p(x|z_j^{10})$. By construction of the tree structure, this node has a unique parent and we follow the path from $z_k^{10}$ to the first level in the tree. This path corresponds to a sequence of topics $\mathbf{t}^x = \{t_1^x, \ldots, t_{10}^x\}$, where $t_{10}^x = z_k^{10}$, and $t_l^x = z^l_{\text{parent}(t^x_{l+1})}$ for $l = 9$ to 1. Given $\mathbf{t}^x$, the similarity between a target segment $y$ and the anchor $x$ is defined as

$$S_2(x, y) = \sum_{l=1}^{10} \alpha_l p(y|t_l^x) \ . \tag{5}$$

The 'hard link' tree structure is rather simple and, by construction, some nodes might be unreachable from the lower level. Such nodes are useless because they cannot appear in the best path used in (5). We thus propose another tree construction algorithm where we enforce a more complex (and balanced) structure where each node have at least two children. The resulting tree-structure guarantees that no topic will be left aside, and allows the use of richer relations between nodes. Integer linear programming (ILP) is employed to obtain an optimal structure[1], maximizing the weight of the links created. More formally, for link creation between levels $l$ and $l+1$, the ILP optimization consists of maximizing

$$\sum_{i \in [1, K_l], j \in [1, K_{l+1}]} \text{sim}(i, j) \ \text{link}(i, j) \tag{6}$$

subject to

$$\sum_{i \in [1, K_l]} \text{link}(i, j) = 1 \quad \forall j \in [1, K_{l+1}] \tag{7}$$

and

$$\sum_{j \in [1, K_{l+1}]} \text{link}(i, j) \geq 2 \quad \forall i \in [1, K_l] \tag{8}$$

where $\text{link}(i, j) = 1$ if a link is created between topic $i$ at level $l$ and topic $j$ at level $l+1$, 0 otherwise, and where $\text{sim}(i, j)$ is the cosine similarity between the two topics. : Because every topic is represented as a distribution over the words in the vocabulary, the similarity between two topics corresponds to a simple cosine between their sets of words, where each word is weighted by the probability in the respective topic. Eq. (7) ensures that every node has only one parent while (8) ensures that each parent has at least two children. At hyperlinking time, the ILP tree-structure is used as the HLT one to generate a path from the best matching node at the lowest level to the coarsest level.

# 3. EXPERIMENTAL EVALUATION

The different selection strategies with hierarchical topic models are evaluated on data from the Search and Hyperlinking task of the 2013 and 2014 MediaEval benchmarks. Both evaluations are based on the same video data, with different anchors over the two years.

## 3.1 Data

The video data set consists in a collection of BBC videos amounting to approximately 4,000 hours, with an average length of 45 minutes for a single video. All videos were transcribed by several ASR systems. We report results with the LIMSI transcripts, preliminary experiments having shown little difference in the conclusions with different ASR systems. Anchors were defined by approx. 30 users, aged between 18 and 30, who selected segments that they found interesting to follow on or relevant of the collection. For evaluation purposes, 30 anchors were selected for each year's evaluation, with an average duration of 32.2 s in 2013 and significantly reduced to 22.9 s in 2014. from the difference in duration, use of context not

Topic models are solely used for the selection step and do not intervene with the segmentation step. We thus limit evaluation to the capacity of the various models to rank a set of targets given an anchor, casting evaluation into a classical information retrieval framework. For the sake of having annotations, we took the targets submitted by the various participants along with their relevance judgment as the set of targets to rerank. In other words, for a given anchor, the set of target segments proposed by the various competing systems in the framework of the benchmark and assessed on Mechanical Turk forms a list of potential targets that we rerank for selection of relevant targets. Over the 30 anchors evaluated in 2013, 9,973 target segments were assessed where 29.9 % were considered as relevant. In 2014, 12,340 target segments are considered of which 15.3 % were deemed relevant. The differences between 2013 and 2014 can be explained by the larger number of participants in 2014 and by the fact that changes in evaluation rules made the task harder. Note also that the segments that we consider for reranking come from a variety of systems, using textual content, visual content, or a combination of both possibly with additional sources (metadata, prosody, etc.). While the segments are biased in the sense that they were proposed for some reasons, the variety of systems ensures diversity.

## 3.2 Results

In the experiments reported in this section, all transcripts were tagged and lemmatized[2]. Only the lemmas corresponding to nouns were kept after removal of stop-words. To limit the influence of transcription errors in building the hierarchical topic structures used, the LDA models were trained for each level on the manual subtitles of the videos from the 2013 collection.

Results are gathered in Table 1 for various settings. The baseline setting (DirectH) corresponds to the direct comparison of the anchor and target transcripts with the cosine similarity. Results for the structure with independent levels were tested with are given for four variants: $\text{IT}_{150}$ corresponds to comparison with $S_1(x, y)$ considering 150 topics (i.e., level 2, for which the best results were obtained); the next three results are for the linear combination in $S_1(x, y)$ with resp. equal weights for all level ($\text{IT}_{\text{Comb=}}$), increasing importance from coarse to fine grain levels ($\text{IT}_{\text{Comb<}}$) and vice-versa ($\text{IT}_{\text{Comb>}}$). Results for the HLT structure are given with $K = 50, 10, 300, 700, 1500$ ($\text{HLT}_1$) and limiting to $K = 50, 150, 300, 700$ (($\text{HLT}_2$)). This last setting enables comparison with the ILP tree structure, which was obtained with only 4 levels due to the computational complexity of the ILP optimization.

Results obtained with independent topic levels are comparable to those obtained with the direct comparison of the transcripts and the differences are not statistically significant (paired t-test, $p < 0.05$). This holds both for the 2013 dataset and for the more difficult 2014 dataset. We observed that varying the number of latent topics did not yield any significant differences either. Interestingly though, an informal assessment of the links revealed that the anchors for

[1] We used `https://www.gnu.org/software/glpk` as solver

[2] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger`

|  | 2013 |  |  | 2014 |  |  |
|---|---|---|---|---|---|---|
| method | @5 | @10 | @20 | @5 | @10 | @20 |
| DirectH | 0.71 | 0.66 | 0.62 | 0.41 | 0.41 | 0.38 |
| $\text{IT}_{150}$ | 0.67 | 0.64 | 0.58 | 0.45 | 0.4 | 0.35 |
| $\text{IT}_{\text{Comb}=}$ | 0.7 | 0.67 | 0.63 | 0.34 | 0.33 | 0.31 |
| $\text{IT}_{\text{Comb}<}$ | 0.68 | 0.66 | 0.62 | 0.31 | 0.33 | 0.32 |
| $\text{IT}_{\text{Comb}>}$ | 0.71 | 0.68 | 0.63 | 0.35 | 0.35 | 0.33 |
| $\text{HLT}_1$ | 0.54 | 0.49 | 0.43 | 0.43 | 0.38 | 0.35 |
| $\text{HLT}_2$ | 0.44 | 0.44 | 0.39 | 0.43 | 0.43 | 0.37 |
| ILP | 0.4 | 0.39 | 0.37 | 0.43 | 0.44 | 0.41 |

**Table 1: Mean average precision at 5, 10 and 20 for all methods on the 2013 and 2014 datasets.**

|  |  | % difference |  |
|---|---|---|---|
| System 1 | System 2 | 2013 | 2014 |
| $\text{IT}_{700}$ | DirectH | 93 | 86 |
| $\text{IT}_{700}$ | $\text{IT}_{\text{Comb}>}$ | 82 | 90 |
| $\text{IT}_{700}$ | $\text{HLT}_2$ | 98 | 93 |
| $\text{HLT}_2$ | $\text{HSLT}_S$ | 29 | 43 |
| $\text{IT}_{\text{Comb}=}$ | $\text{HSLT}_2$ | 94 | 95 |

**Table 2: Percentage of anchor/target pairs proposed and that differ between two runs.**

which relevant targets were found were mostly the ones addressing more general topics. The combination strategies bring marginal improvement on the 2013 data but not on the 2014 data. Globally, the results clearly establish that topic models can be efficiently used to create relevant links. As we will illustrate in the next paragraph, the links created with topic models are different from those created with direct content comparison, yet they are relevant. Comparing tree-based topic structures with direct comparison and independent topics, we see a drop in performance for 2013 that does not occur on the 2014 data. A plausible explanation lies in the shorter and more realistic anchors defined in 2014, combined with the absence of context. These features are detrimental for direct content comparison and benefits topic-based matching. In such a situation, tree-based topic matching performs well and offers more insight on the control of the links thanks to the sibling relations.

An in-depth analysis of the links created by the different methods was performed. For instance, Tab. 2 reports the proportion of targets that differ between two systems. While all systems exhibit comparable MAP, the pairwise comparison shows that a large proportion of the links proposed differs between two systems. This proves that the different strategies proposed here are complementary and hints that all those techniques can be leveraged to propose a wider variety of links than those offered by direct content comparison. We also studied the distribution of the cosine similarity between an anchor and the relevant targets proposed by the various methods. As the topic structure gets more complex, from independent topics to tree-structures, the median cosine similarity between anchor and targets gets lower, particularly on the 2013 data. This fact again highlights the potential interest of topic-based hyperlinking to provide links between segments that share little vocabulary and potentially exhibit serendipity.

## 4. DISCUSSION

Video hyperlinking based on language has mostly exploited so far the direct comparison of content using either bag-of-words representations or entity matching. The main drawback of these methods is the lack of diversity in the links generated, as well as a limited control of the links proposed. Experimental evaluation of hyperlinking with hierarchical topic models shows that various topic structures are able to provide equally relevant links that significantly differ from those obtained with direct comparison. These preliminary results call for further investigation of the flexibility offered to hyperlinking by hierarchical topic structures. In particular, we believe that such structures have the potential for serendipitous hyperlinking, as suggested by the results reported in this paper. Having explicit topics along with relations between topics also offers ground for link justification, i.e., being able to tell users why the link was proposed and what new aspect of the topic will be covered by following the link.

## 5. ACKNOWLEDGEMENTS

This work received a French state support via the Comin-Labs excellence laboratory financed by the National Research Agency under reference ANR-10-LABX-07-01.

## 6. REFERENCES

[1] C. Bhatt, N. Pappas, M. Habibi, and A. Popescu-Belis. Idiap at MediaEval 2013: Search and hyperlinking task. In *Working Notes MediaEval Workshop*, 2013.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[3] T. De Nies, W. De Neve, E. Mannens, and R. Van de Walle. Ghent University-iMinds at MediaEval 2013: an unsupervised named entity-based similarity measure for search and hyperlinking. In *Working Notes MediaEval Workshop*, 2013.

[4] M. Eskevich, G. J. F. Jones, R. Aly, and et al. Multimedia information seeking through search and hyperlinking. In *ACM Intl. Conf. on Multimedia Retrieval*, 2013.

[5] P. Galuscáková, M. Krulis, J. Lokoc, and P. Pecina. CUNI at MediaEval 2014 search and hyperlinking task: Visual and prosodic features in hyperlinking. In *Working Notes MediaEval Workshop*, 2014.

[6] C. Guinaudeau, A.-R. Şimon, G. Gravier, and P. Sébillot. HITS and IRISA at MediaEval 2013: Search and hyperlinking task. In *Working Notes MediaEval Workshop*, 2013.

[7] C. Guinaudeau, G. Gravier, and P. Sébillot. IRISA at MediaEval 2012: Search and hyperlinking task. In *Working Notes MediaEval Workshop*, 2012.

[8] J. Preston, J. Hare, S. Samangooei, J. Davies, N. Jain, D. Dupplaw, and P. H. Lewis. A unified, modular and multimodal approach to search and hyperlinking video. In *Working Notes MediaEval Workshop*, 2013.

[9] A.-R. Simon, G. Gravier, P. Sébillot, and M.-F. Moens. IRISA and KUL at MediaEval 2014: Search and hyperlinking task. In *Working Notes MediaEval Workshop*, 2014.

[10] C. Ventura, M. Tella-Amo, and X. G. Nieto. UPC at MediaEval 2013 hyperlinking task. In *Working Notes MediaEval Workshop*, 2013.