# IRISA and KUL at MediaEval 2014: Search and Hyperlinking Task

Anca-Roxana Simon, Guillaume Gravier, Pascale Sébillot, Marie-Francine Moens

# IRISA and KUL at MediaEval 2014: Search and Hyperlinking Task


Anca-Roxana Şimon*, Guillaume Gravier**, Pascale Sébillot***
IRISA & INRIA Rennes
Univ. Rennes 1*, CNRS**, INSA***
35042 Rennes Cedex, France
firstname.lastname@irisa.fr

Marie-Francine Moens
KU Leuven
Celestijnenlaan 200A
B-3001 Heverlee, Belgium
sien.moens@cs.kuleuven.be



## ABSTRACT

This paper presents our approach and results in the hyperlinking sub-task at MediaEval 2014. A two step approach is implemented: relying on a topic segmentation technique, the first step consists in generating potential target segments; then, for each anchor, the best 20 target segments are selected according to two distinct strategies: the first one focuses on the identification of very similar targets using n-grams and named entities; the second one makes use of an intermediate structure built from topic models, which offers the possibility to control serendipity and to explain the links created.


## 1. INTRODUCTION

This paper presents the joint participation of IRISA and KUL at the MediaEval 2014 Search and Hyperlinking task, in which focus is set on the hyperlinking sub-task [2]. The goal is thus to create hyperlinks between predefined anchor segments and short video segments, called targets, which should offer complementary information not found at search time. Targets have to be automatically extracted from videos of a large collection.

The hyperlinking system we propose consists in a two step process: first, all potential target segments are extracted using a topic segmentation technique; then, the most relevant targets are selected for each anchor, with the help of content analysis and similarity measures.

In our 2013 participation [5], we focused on precise target selection and our most efficient system consisted in a direct comparison between anchor and target segments obtained through topic segmentation using bags of n-grams to represent content, thus resulting in links between anchor and very similar targets. We thus go on exploring this direction. However, we believe that besides precise target selection, a very important aspect of hyperlinking is to offer serendipity and to explain why two video segments are linked. To address these points, we propose an intermediate structure, obtained from topic models, that allows an indirect comparison between anchors and targets. Its first advantage is that segments which do not share a consistent part of the vocabulary but are semantically related can be linked. Moreover this structure provides a basis to investigate why certain links are created. Link justification is an interesting aspect of hyperlinking which, to our knowledge, has not been addressed before.



## 2. SYSTEM OVERVIEW

The aim of our approach is to find target segments of the same topic as the anchor, or of related topics. The hyperlink generation relies on content-based comparisons exploiting spoken data obtained from automatic transcripts and manual subtitles [4]. Data are first lemmatized and only nouns, non modal verbs and adjectives are kept. Subsections 2.1 and 2.2 respectively detail the two parts of our system: the generation of potential target segments and the selection of the top 20 targets for each anchor.

### 2.1 Generating potential target segments

Each video in the test collection is partitioned into topically coherent segments with the generic topic segmentation algorithm TextSeg [7], which is domain independent, needs no a priori information, is efficient on speech transcripts and on segments of varying lengths. Its main drawback concerns over-segmentation, which in our case is not problematic since the target segments must not last longer than 2 minutes.

### 2.2 Selection of hyperlinks targets

Each anchor segment is compared with each topically coherent segment previously obtained thanks to similarity measures. The comparison can be direct or indirect (i.e., using intermediate structures).

Four methods are proposed to select the hyperlinks targets. The baseline corresponds to the method for which we obtained our best results in 2013 (*Linear+ngrams*); a direct comparison between segments is done, contents being represented by bags of unigrams, bigrams and trigrams. The second method extends the previous one by using the Stanford Named Entity Recognizer (NER) [3], a 3 class (*person, organization and location*) entity tagger (*Linear+ngrams+NEs*). A Jaccard similarity coefficient is used to evaluate similarities in terms of shared named entities (NEs). The n-grams and NEs similarity scores are combined; weights of 0.3 and 0.7 respectively are chosen to favor precise alignments, thus not rewarding serendipity.

For the last two methods an intermediate structure is built using Latent Dirichlet Allocation (LDA) probabilistic topic models [1] learned on the manual transcripts of the development set (1335 hours of video). Each transcript is represented as a mixture of $K$ latent topics, where a latent topic is represented as a probability distribution over the vocabulary. Contrary to the bag of words representation, this one clusters semantically similar co-occurring terms. To construct the structure LDA is trained using Gibbs sampling, standard values for the hyperparameters $\alpha$=50/K and $\beta$=0.01 [6] with a varying number of latent topics $K$: 50, 100, 150, 200, 300, 500, 700. This range for the number of topics was chosen to learn general to more specific topics. Using this structure an indirect comparison between anchor and target segments can be performed.

Our third method (*TopicM*) consists in computing, for every $K$, the probabilities of the anchor and target segments given the topics. For each $K$ and for each anchor-target pair, two vectors are obtained in which $component_i$ corresponds to the probability of $topic_i$, given the words contained in the segment. Then a similarity measure is computed between these vectors, leading to a score for each anchor-target pair. To select the top targets, for each anchor a linear combination of the scores resulted for each $K$ is done with more importance to the most specific topics.

For the last method (*HierTopicM*), the previous structure with topic models is extended to form a tree-like hierarchy between the topics. This hierarchy relies on a similarity measure between the topics obtained with $K_i$ and $K_j$, where $K_j < K_i$. Thus, each topic learned with $K = 700$ will be connected to the most similar topic learned with $K = 500$ and so on for the other $K$ values. Having this representation a path for each anchor can be selected, starting with the bottom of the tree ($K = 700$) and choosing the topic with the highest probability given the words in the anchor and going on with the selection of the parents for that topic until the first level in the tree ($K = 50$) is reached. The targets are then selected by a linear combination of the probability values of the topics in an anchor path, given the words in the target. Thus, in the end only a part of the topics in the structure will be considered, allowing more precise control of serendipity and justification of links.

Using the structure of topics (hierarchical or not) to select the hyperlinks, serendipity can be controlled by giving more weights to values attained with more general or more specific topics. Moreover links can be justified, by looking at the top words of the topics that contributed most in the selection of those links. This structure can be used a posteriori for other methods to understand the links creation.

Considering the evaluation rules, the target segments proposed for each anchor should have a duration between 10 sec and 2 min. Therefore, a post-processing of the final targets is done to verify these constraints. If a segment is longer than 2 min, it is resegmented with a sliding window of 2 min and the best scoring window within the segment will represent the final target segment. If a segment is shorter than 10 sec it is combined with the best scoring neighbor, resulted from the topic segmentation, until the minimum length is reached.

# 3. RESULTS

Official evaluation results are reported in Table 1. Only the top 10 results were evaluated for each method, via crowdsourcing on Amazon Mechanical Turk (AMT). Other results were evaluated automatically based on the evaluation of the selected results from all participants. The best results are obtained with the topic segmentation with ngrams on the manual subtitles (*Linear+ngrams*). This means that turkers prefer that the content of the target segments is closely related to that of the anchor. As anticipated, on the automatic transcripts the precision decreases. Unexpectedly, adding NEs to this method and therefore favoring segments about the same people, places, organizations, diminished the precision. From a manual assessment of several target segments proposed using NEs, it seems that having targets speaking about same people (e.g., Madonna, Beckham) in different circumstances (i.e, shows on different subjects: diets, charity) is not relevant. Possibly, giving less weight to the NEs shared between anchors and targets, the precision could be improved.

| | P_5 | P_10 | P_20 |
|---|---|---|---|
| Linear+ngrams (LIMSI) | 0.1 | 0.096 | 0.06 |
| Linear+ngrams (MANUAL) | 0.19 | 0.16 | 0.1 |
| Linear+ngrams+NEs (LIMSI) | 0.046 | 0.036 | 0.02 |
| TopicM (LIMSI) | 0.093 | 0.093 | 0.048 |
| HierTopicM (LIMSI) | 0.03 | 0.02 | 0.01 |

**Table 1: Precision values obtained for all proposed methods on the 2014 test set.**

Regarding the topic model-based methods, the one that uses all the topics learned (*TopicM*) yields better results, comparable to those obtained with the *Linear+ngrams* method. From a manual assessment of the results it seems that the topics, even those learned with $K = 700$ are too general. The targets that were considered relevant are those for which the anchor addresses a more general topic (e.g., wildlife). The problem of generality appears also for the *HierTopicM* method. However having only a part of the topics considered the results are worse than with *TopicM*. Using such intermediate structures could be improved by learning more specific topics. Still, having the topwords of the topics that best explain the anchors and the targets can help interpret and justify the links. Additionally, with an intermediate structure, anchors are linked to targets even without sharing much vocabulary.

# 4. REFERENCES


[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. F. Jones. The Search and Hyperlinking task at MediaEval 2014. In *Working notes of the MediaEval 2014 Workshop*, Barcelona, Spain, 2014.

[3] J. R. Finkel, T. Grenager, and C. D. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Association for Computational Linguistics*, Ann Arbor, USA, 2005.

[4] J.-L. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.

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

[6] M. Steyvers and T. Griffiths. Probabilistic Topic Models. *Handbook of Latent Semantic Analysis*, 427(7):424–440, 2007.

[7] M. Utiyama and H. Isahara. A statistical model for domain-independent text segmentation. In *Association for Computational Linguistics*, Toulouse, France, 2001.