Leveraging Multimodal LDA for Hyperlinking

Anca Roxana Simon, Ronan Sicre, Rémi Bois, Guillaume Gravier,
Pascale Sébillot

IRISA – France

Plan

1. Introduction

Hyperlinking: Linking video fragments

For machines and for humans

- "Advanced tasks" (e.g., video summarization)
- Media workers, companies (e.g., analytics)
- Generic user (e.g., recommendation)

For machines

- Near-duplicates (can be used for clustering or automatic summarization)
- Fragments that are part of the timeline (i.e. related events that happened just before or just after)

For humans

- *Diverse* targets to cover the potential interests of the user

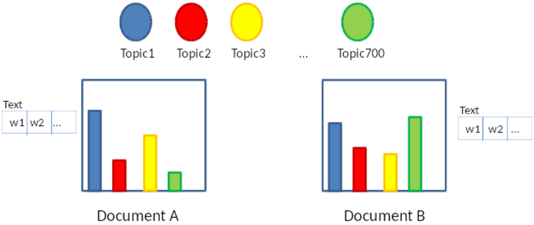
Plan

2 Joining Audio and Visual Informations

Latent Dirichlet Allocation

The idea

- Latent topics are extracted from a collection
- A document is represented by its topics probabilities

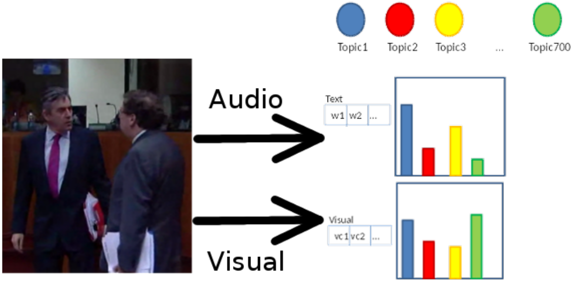


- Topics distributions can be compared
- Documents that do not share vocabulary can have a high similarity

Building conjointly two modalities

Using both audio and visual informations

- Idea: From comparable documents in two languages, build topics in both languages conjointly
- We use audio and visual informations as two different languages and build cross-modality topics



- For each visual topic, there exists a corresponding audio topic

Exemples of mappings between audio and visual

Most probable words from topic n°3 in our model:

Audio	love home feel day life baby made thing la
Visual	singer microphone sax concert master-of-ceremonies cornet flute trombone banjo

Most probable words from topic n°25 in our model:

Audio	years technology computer find key future power machine speed science
Visual	equipment machine tape-player computer appliance-recording memory-tape CD-player

From visual to audio

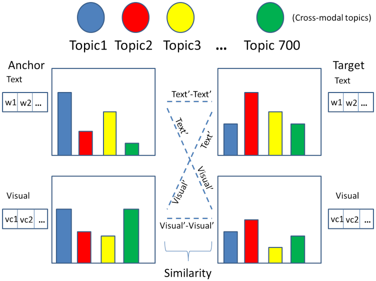
Objective

By learning this mapping, we can apply the usual topic similarities (i.e. audio → audio or visual → visual).
We can also apply cross-modality similarities (i.e. audio → visual or visual → audio).

New kinds of links

Cross-modality similarities correspond to:

- Seeing more about what is said
- Hearing more about what is shown



Plan

3 Data and Evaluation

Our system

What we used

- Automatic transcriptions from LIMSI
- Visual concepts from Leuven

Method and Reranking

Run1 Visual similarity (no topics) with visual reranking (top 50)

Run2 Audio to visual with visual reranking (top 50)

Run3 Visual to audio with ngram reranking (top 50)

Run4 Rank Aggregation

Reranking

- CNN trained on ImageNet ILSVRC (VGG 16) for visual reranking
- Unigram, bigram and trigram cosine similarity for ngram reranking

Plan

4 Results and Analysis

Near-Median scores but hard to compare

	Minimum		25%		50%	75%	Maximum
Prec 10	0.017		0.198		0.275	0.524	0.608
Run1				0.207			
Run2	0.017						
Run3				0.224			
Run4		0.156					

Table: Results for our four runs

Some of our relevant targets (RUN3)

Anchor 85

- Talks about the Ireland saying “No” to the Lisbon Treaty
- Europe is not happy, Mandelson (UK politician) is blamed by Nicolas Sarkozy but Gordon Brown supports Mandelson

Target 3

- Almost identical content (another news show 3 hours later)

Target 8

- Explanation of the successive difficulties of the EU in the ratification of treaties
- Focuses on times when referendum were used as opposed to parliamentary ratification

Some of our non-relevant targets

Anchor 85

- Talks about the Ireland saying “No” to the Lisbon Treaty
- Europe is not happy, Mandelson (UK politician) is blamed by Nicolas Sarkozy but Gordon Brown supports him

Target 6

- The UK Parliament debates on the answer that should be given to Ireland: push them to do another referendum or don’t pressure them
- Gordon Brown is in favor of pressuring them while the opposition calls for inaction

Suggestions for the evaluation

What we think

- Almost identical targets should be identified
- There should be several Turkers by anchor/target pair

What we know

- There would be a low inter-annotator agreement

Plan

5 Conclusion

Strengths and weaknesses

Strengths

- Brings more diversity
- A new way to exploit cross-modality
- More control over link creation

Weaknesses

- Works badly on some anchors (e.g., visual → audio showing an anchorman)

Push the community for more diversity