EURECOM
Sophia Antipolis

POLITECNICO DI TORINO

TRECVID 2017 Hyperlinking task

Eurecom-Polito team

Authors:

Benoit Huet
(EURECOM)
huet@eurecom.fr

Elena Baralis
Paolo Garza
Mohammad Reza Kavoosifar
(Politecnico di Torino)
{name.surname}@polito.it

Presented by:

Bernard Merialdo
(EURECOM)

System overview

- Our system is based on textual and visual feature analysis
  - ❖ The system is multimodal, however it starts with independent monomodal queries and combine the results of these queries to obtain the final result

- We used
  - o Automatic speech recognition (ASR) transcripts
    - o LIMSI
  - o Visual concepts
    - o extracted by using the Caffe framework with the BVLC GoogLeNet model
  - o Metadata
    - o Title, description and tags

- In order to have related text information, we also used:
  - o Named-entity recognition (NER)
    - o Stanford NER (From Stanford university), also known as CRFClassifier
  - o Concept mapping technique
    - ❖ Based on synonymous identified by means of Wordnet

System overview

- The core of all runs is composed of three stages:
  1. **Data segmentation**
     - We considered 120-seconds Fixed-segmentation
       - ❖ We didn't consider overlapping for this year
     - Stop words and punctuation removal tool
     - Word stemming is applied
  2. **Indexing and retrieval**
     - Apache Solr was used to index and retrieve data
  3. **Query formulation and segment retrieval**
     - Transforming the anchor (query) segment into a set of monomodal text-based query
       1. Including in the text of the query:
          - ✓ The words appearing in the LIMSI transcripts, or
          - ✓ The names of the identified visual concepts, or
          - ✓ The words appearing in the metadata
       2. Named-entity recognition and Concept mapping techniques are also applied to increase the importance of entities and the more relevant visual concepts
          - ❖ For increasing the importance, more weight is given to the entity when calculating the relevant score using TF-IDF
     - The prepared query is executed on Solr and returns the most relevant segments

System overview – query types

- LIMSI-based query + Named-entity recognition
  - For each anchor, a textual query is built by considering the words appearing in the LIMSI transcript of the anchor
  - The Name-entity recognition technique is used to identify the words associated with entities
    - A higher weight is assigned to those words in the query
  - The query is executed with respect to the LIMSI transcripts of the queried segments

- Visual concept based query + Concept mapping technique
  - For each anchor, a textual query is built by considering the “names” of the visual concept appearing in the anchor
    - Select only the visual concepts with a score/probability greater than 0.3
  - The Concept mapping technique selects the visual concepts related to the Metadata of the video
    - A higher weight is assigned to those concepts in the query
  - The query is executed with respect to the Visual concepts of the queried segments

System overview – query types

- Metadata based query for segment selection
  - For each anchor, a textual query is built by considering the metadata appearing in the video containing the anchor
    - Metadata are available only at the video level
  - The query is executed with respect to the LIMSI transcripts of the queried segments
    - Segments are returned

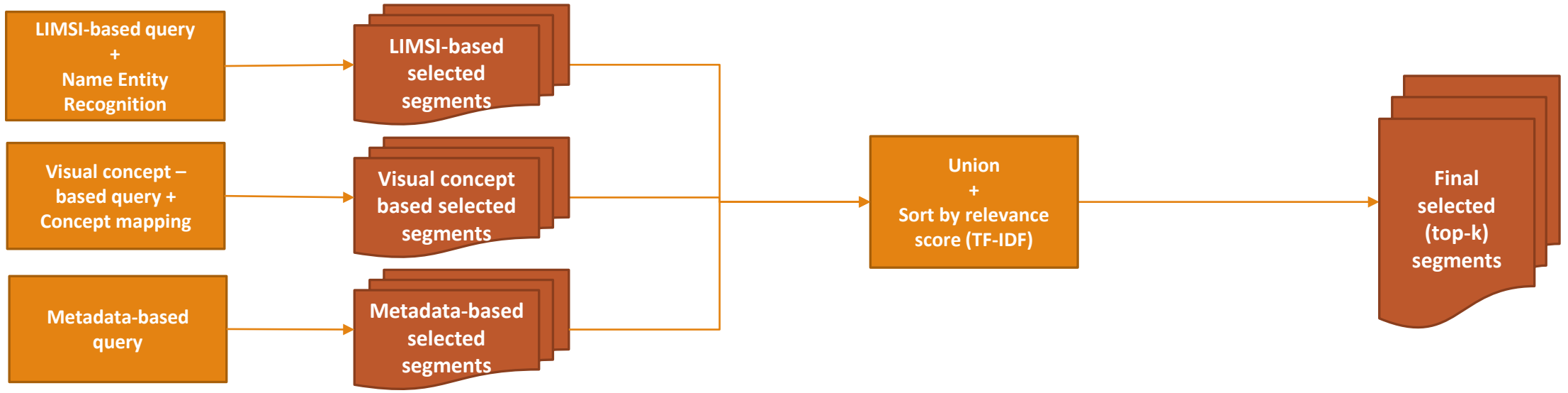
- Metadata based query for video selection
  - For each anchor, a textual query is built by considering the metadata appearing in the video containing the anchor
  - The query is executed with respect to the metadata information the queried videos
    - Videos are returned

SUBMITTED RUNS

1. **Automatic Feature Selection (AFS)**
   - **Features:**
     - ❑ **Metadata, LIMSI, Visual concepts**
       - o **Also Named-entity recognition (NER) and Concept mapping techniques**
2. **Meta-data based approach**
   - **Features:**
     - ❑ **Metadata, LIMSI, Visual concepts**
       - o **Also Named-entity recognition (NER) and Concept mapping techniques**
3. **LIMSI-NER**
   - **Features:**
     - ❑ **LIMSI**
       - o **Also Named-entity recognition (NER)**
4. **Pipeline approach**
   - **Features:**
     - ❑ **LIMSI, Visual concepts**
       - o **Also Named-entity recognition (NER) and Concept mapping techniques**

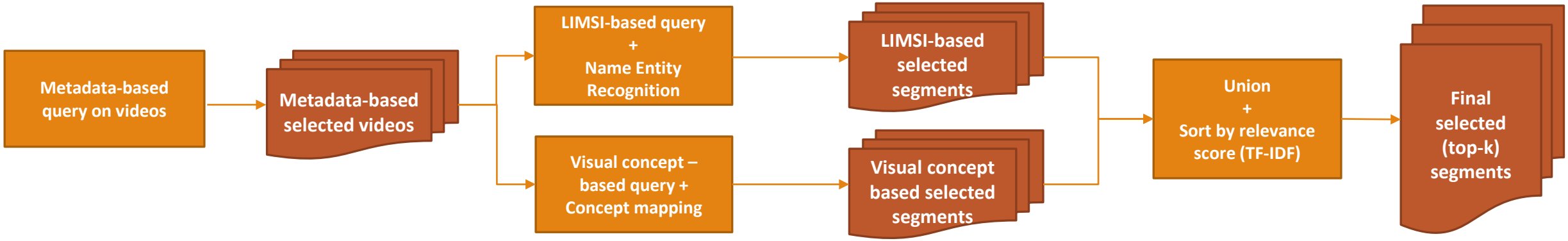
Run 1: Automatic Feature Selection (AFS)

- **Features**:
  - Metadata, LIMSI, Visual concepts
    - ✓ Also Named-entity recognition (NER) and Concept mapping techniques
- **For each anchor:**
  1. Select one set of relevant segments for each feature by considering one feature at a time (monomodal queries)
  2. Consider the union of the segments selected in Step 1, rank them by relevance score, and select the subset of segments with the highest relevance scores
     - We used the TF-IDF-based score returned by Solar to identify the relevance score of each of the selected segments



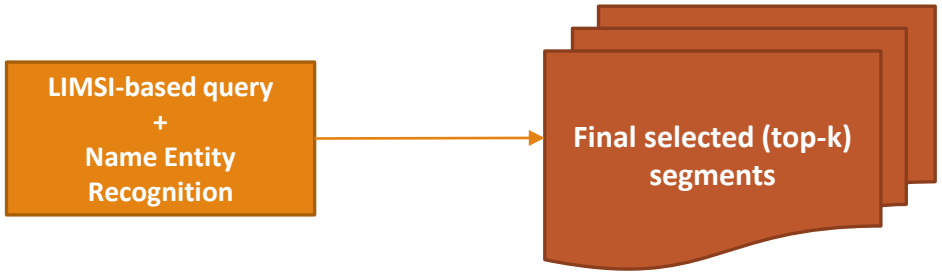
Run 2: Meta-data based approach

- **Features**:
  - Metadata, LIMSI, Visual concepts
    - ✓ Also Named-entity recognition (NER) and Concept mapping techniques
  - Differently from Run 1, Meta-data are used to perform an initial filter on the videos that could contain interesting segments.
- **For each anchor:**
  1. Select relevant videos by using metadata for querying the video collection
  2. Select the most relevant segments from the selected videos by using LIMSI and visual concepts
     - Combine the results of two monomodal queries
     - We used the TF-IDF-based score returned by Solar to identify the relevance score of each of the selected segments



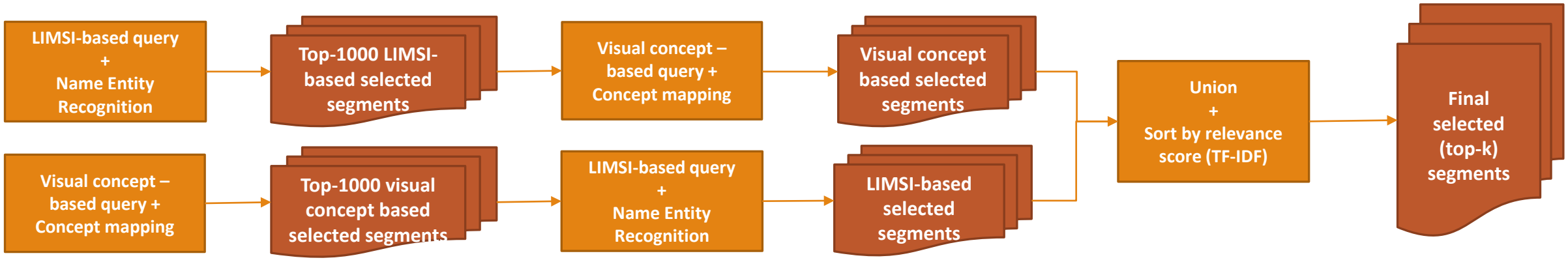
Run 3: LIMSI-NER

- **Features**:
  - LIMSI
    - ✓ Also Named-entity recognition (NER) technique
- **For each anchor:**
  1. Select relevant segments by using LIMS for querying the video collection
     - We used the TF-IDF-based score returned by Solar to identify the relevance score of each of the selected segments
- **Monomodal algorithm**
  - The aim of this algorithm to analyze the differences between monomodal and multimodal approaches
  - The LIMSI transcript feature, on the development anchors, performs better than the other features



Run 4: Pipeline approach

- **Features:**
  - LIMSI, Visual concepts
    - ✓ Also Named-entity recognition (NER) and Concept mapping techniques
- **For each anchor:**
  - **Step 1-1:** Select the top-1000 relevant segments by using LIMSI for querying the video collection
  - **Step 1-2:** Select the most relevant segments from the segments selected in Step 1-1 by using visual concepts
  - **Step 2:** Repeat Step 1 by switching the roles of LIMSI and visual concepts
  - **Step 3:** Consider the union of the segments selected in Step 1, rank them by relevance score, and select the subset of segments with the highest relevance scores



RESULTS

- Run 1 (Automatic Feature Selection) yields the best results in term of all the considered metrics
- Run 2 (the Meta-data based approach) achieved the lowest result
  - The Meta-data-based video pre-filtering step selects very few related videos for some anchors
- The achieved results show that the proper combination of several features performs better than single features

RUN	Name	P @ 5	P @ 10	MAP	MAiSP
1	**Automatic Feature selection (AFS)**	**0.8400**	**0.8080**	**0.1638**	**0.2527**
2	Metadata based approach	0.7040	0.5560	0.0815	0.1320
3	LIMSI-NER	0.7250	0.6667	0.0930	0.1547
4	Pipeline approach	0.8080	0.7480	0.1135	0.1851

RESULTS - Automatic Feature selection (AFS)

- The table below demonstrates:
  1. the total percentage of occurrence for each modality for the top 10 segments
  2. the percentage of average performance per modality over all the queries.

❖ It is completely clear that most of the relevant segments are returned by visual concepts, However, Metadata has a high proportion for the relevant segments

❖ However, based on the average performance per modality:

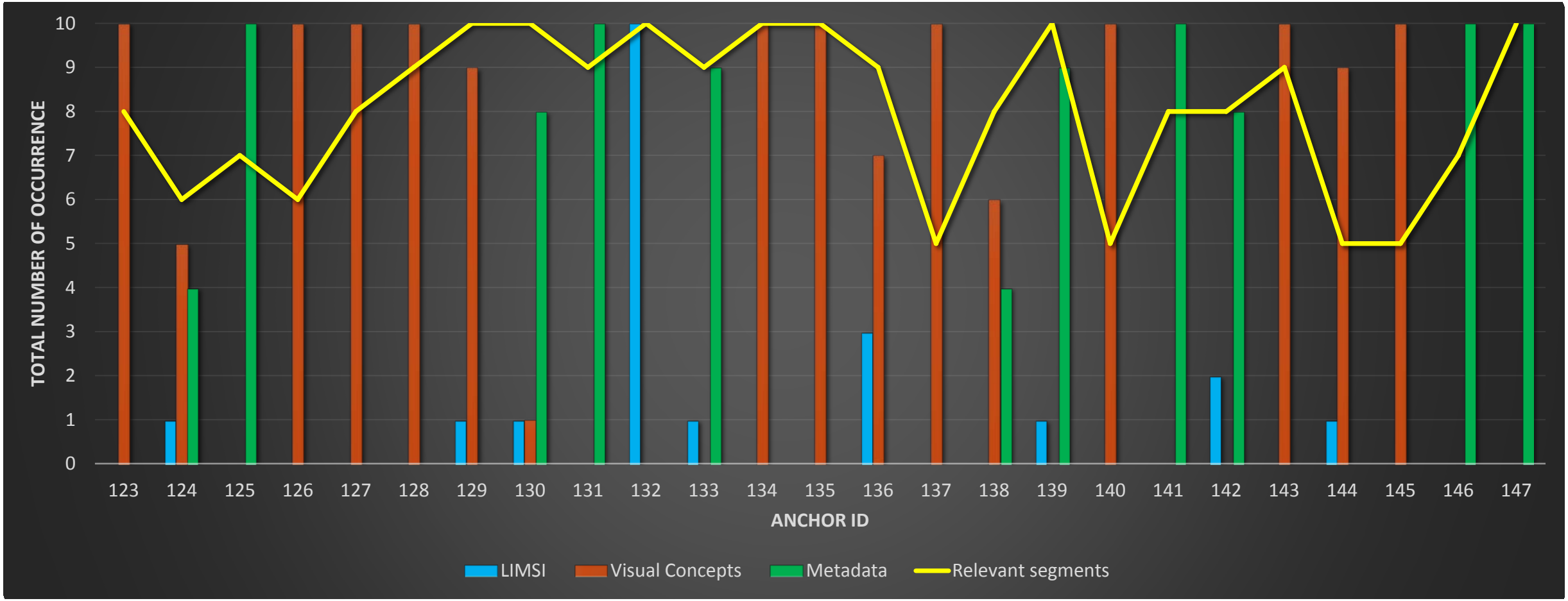
✓ All the modality tested are relevant and accuracy is rather strong.

✓ Using visual concepts only would not win over the multimodal approach.

Feature	% of occurrence in top 10 segments	% average performance per modality over all the queries
LIMSI	8.4 %	76.2 %
Visual concepts	54.8 %	75.2 %
Metadata	36.8 %	89.1 %

RESULTS - Automatic Feature selection (AFS)

- The chart below shows the total number of occurrence of each modality for each anchor and for the 10 segments



FUTURE WORK AND CONCLUSION

- The proposed system has explored the use of textual and visual features for solving the Hyperlinking task.
  - ❖ Specifically, we have considered the LIMSI transcripts, visual concepts and Meta-data.
  - ❖ Moreover, named-entity recognition and a concept mapping technique have also been considered.

- The achieved results show that the proper combination of several features performs better than single features.

- For the future work:
  - ❑ On the AFS approach, features like OCR would be added to the algorithm for further analysis
  - ❑ On the pipeline approach, the intersection of the various modalities (pairs and also the triplet) will be analyzed

Thank you for your attention

Authors are awaiting questions by email

Benoit Huet (huet@eurecom.fr)

Elena Baralis (elena.baralis@polito.it)

Paolo Garza (Paolo.garza@polito.it)

Mohammad Reza Kavoosifar (mohammadreza.kavoosifar@polito.it)

EURECOM
Sophia Antipolis

POLITECNICO DI TORINO