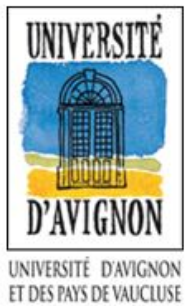
A first summarization system of a video in a target language



The key challenge and potential impact

- How to make the main idea presented in a video in a foreign language accessible and easy to understand by everyone?
- Accessing to information in foreign languages would permit to access to the other side of a story.
- Due to political, socio-cultural or religion reasons, divergence of opinions may exist within two medias from two different sources.

Main objectives

Objective 1: Understanding the main idea of a video by summarizing

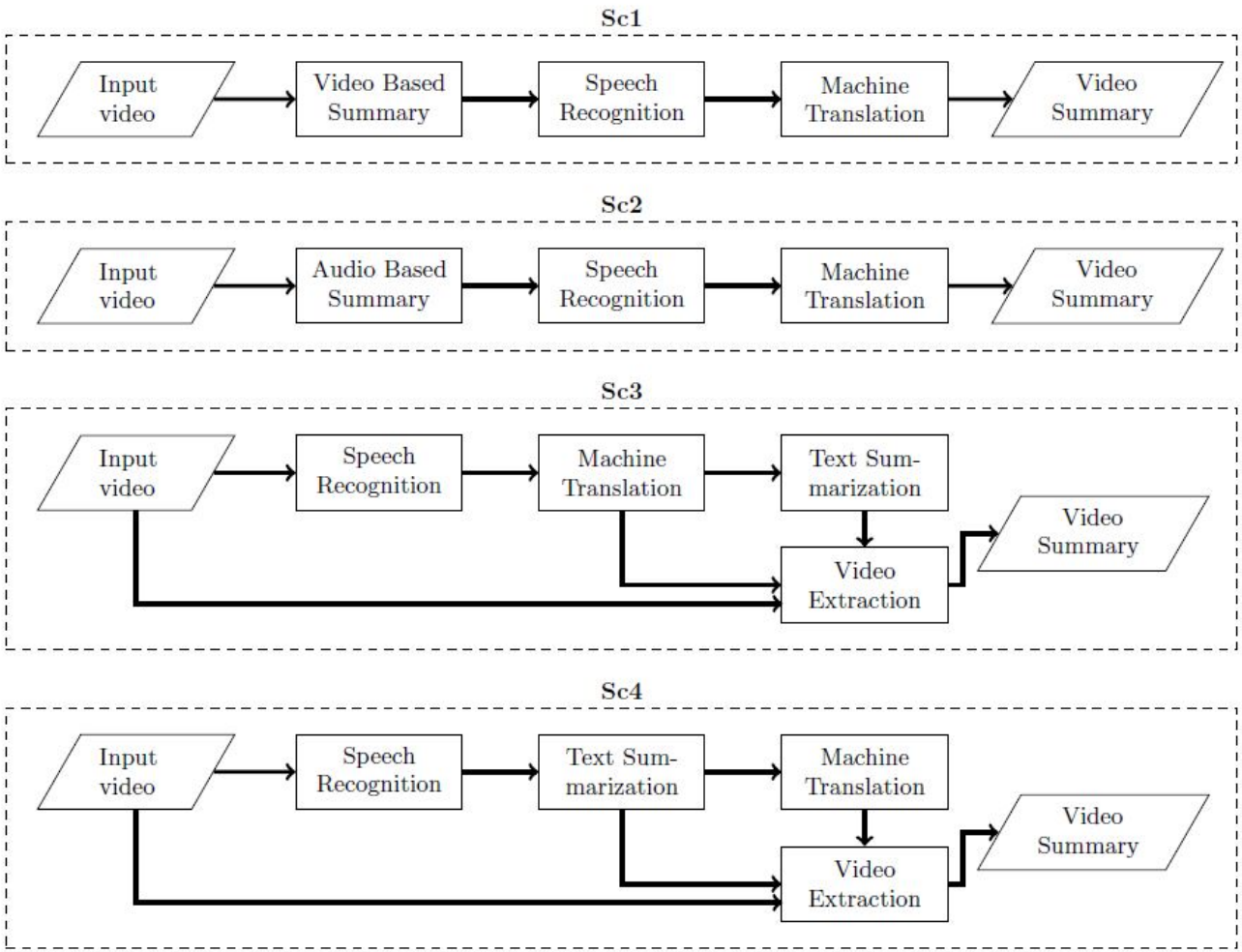
- Input: A video in Arabic or French.
- Output: A summary of the input video subtitled in English. This summary is supposed to capture the main idea of the source video.

Objetcive 2: Cross-lingual opinion Analysis

- Input: An Arabic video will be compared to French or English video.
- Output: A review concerning the degree of divergence between the two videos .

Global architectures

To reach the first objective several components are necessary. We developed several scenarios.



What is necessary to develop AMIS?

1. Collecting video data
2. Video Summarization
3. Speech recognition system
4. Machine Translation
5. Text and Audio summarization
6. Making multilingual corpora comparable
7. Labeling (sentiments and opinions) the multilingual corpora
8. Evaluation (objective and subjective)

1- Video database

Data on conflicting topics for studying the cross-lingual opinions in Arabic, English and French.

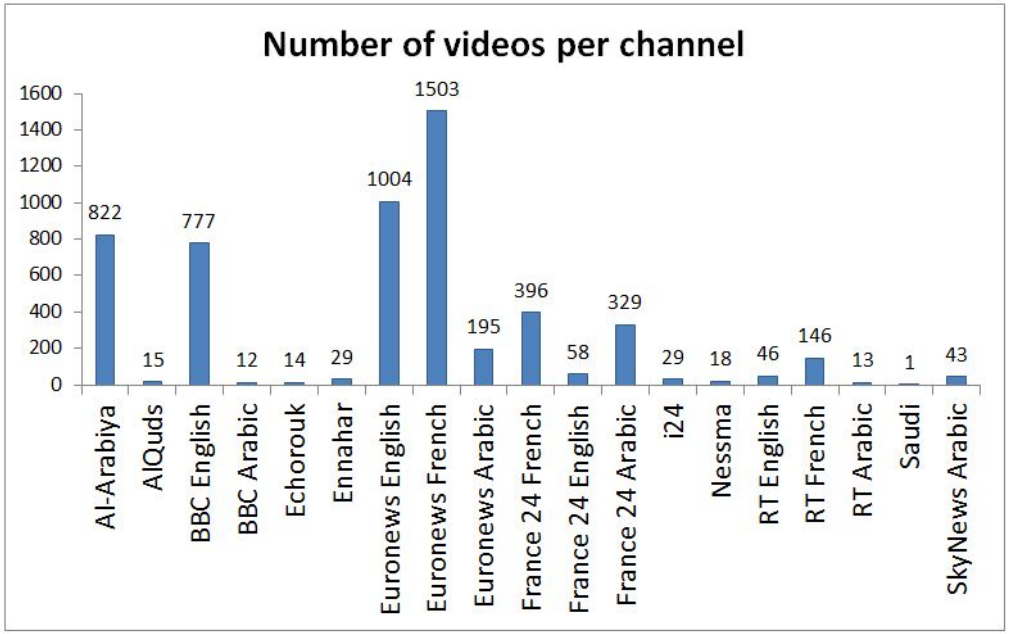
Syria	Real Madrid - FC Barcelona	Animal rights
Women's rights	Homosexual marriage	Drug liberalization
Death sentence	Occupied territories	Trump

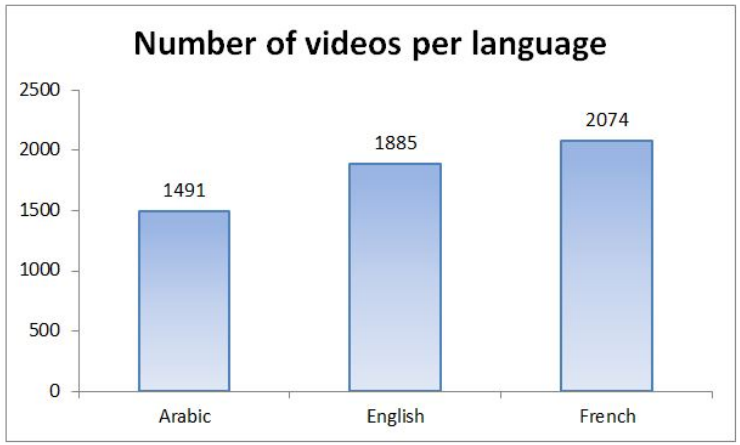
Arabic channels

English channels

French channels

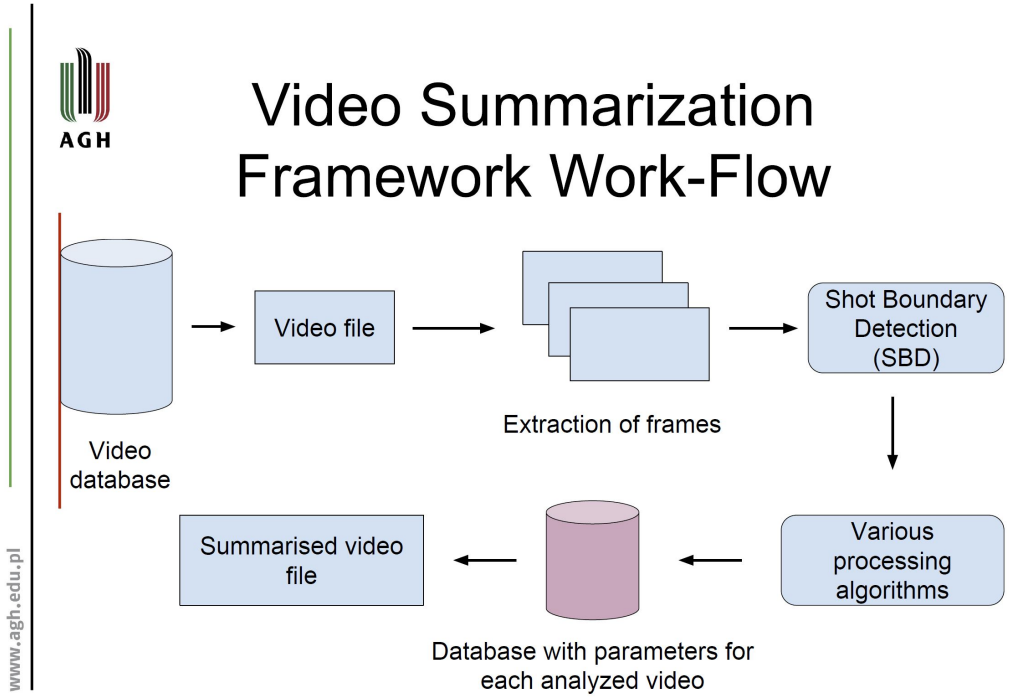
Some statistics





- 100 hours of videos have been collected for each of the three languages.
- For each video in a language, videos in the 2 other languages correspond to the same topic.

2- Video Summarization



- Anchor-person.
- Recognition of day and night shots.
- These indicators are used for calculating coefficient for activity.
- The coefficient of activity is calculated for each frame and then as an average per each shot.

Jump Cuts and Dissolve (Fade) to White

Jumpcut – cut in film editing in which two sequential shots of the subject taken from camera positions varying only slightly if at all

Dissolve (fade) to white used to soften jump cuts startling viewer

3- Arabic Automatic Speech Recognition System

Arabic is considered as the foreign language. Development of an Arabic automatic speech recognition system with the following data:
Acoustic corpus (hours): 63 hours.

- 35 acoustic models have been trained
- The emission probabilities of the HMM models are estimated by DNN.
- The number of parameters to estimate is about 30 million.

Text corpus

- GigaWord corpus was collected from 9 sources of information with a total of 10^9 words.
- 315K words from the transcription of the acoustic training data
- A vocabulary of 95K words with an average of 5.07 pronunciations for each entry

4- Machine Translation

- A statistical Machine Translation has been developed Arabic-English.
- For the training process, we use a parallel corpus extracted from UN proceedings of 9.7 million of sentences.
- For the target language we used a 4-gram language model
- The vocabulary contains 224000 entries.

5- Text summarization

- Several techniques of text summarization do exist in the literature.
  - Summarization by extraction
  - summarization by abstraction
  - summarization by sentence compression
- Our purpose is to summarize the result of a noisy text. We decided mainly to adopt the extractive summarization paradigm in AMIS project.

5- Audio Summarization

- Selecting segments within the audio signal that are considered more relevant.
- Informativeness of each segment is obtained by mapping a set of audio features issued from its Mel-frequency Cepstral Coefficients and their corresponding Jensen-Shannon divergence score.

Objective 2: Cross-lingual sentiment analysis

First we need to make all the database comparable.

- To compare two videos in two different languages in terms of sentiment, we need first to make them comparable.
- We tested several based-dictionary methods for the comparability.
- Each document is represented by a TFIDF vector.
- The dimension n of a vector is fixed in terms of the most frequent words in Arabic.
- An English document is represented also by a vector of dimension n corresponding to the same words such as in Arabic.

Making corpora comparable

- The evaluation is done on 123 videos of Euronews that we know that they are comparable.

	Max	**Min**	**Average**
Arabic	359	39	160
English	432	38	186

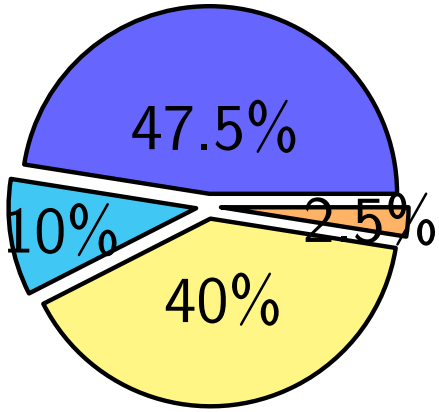
Table: Statistics in terms of words of the test corpus

- The method achieved a score of a Recall of 0.73 and a Precision of 0.62

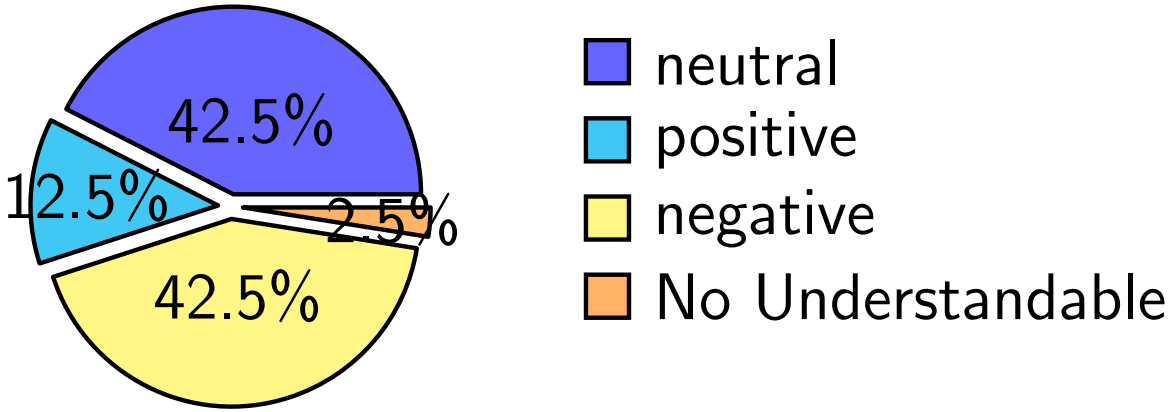
Building a Reference Corpus for cross-lingual sentiment analysis

- We asked 12 annotators to label videos in terms of sentiments (polarity).
- They label 40 videos.
- Each transcribed document has been segmented into phrases.
- The used labels are the following:
  - 1: If the phrase is Positive
  - -1: If the phrase is Negative
  - 0: If the sentence is Neutral
  - ∞: No Understandable

Results of Labeling in terms of videos (40)



(a) Arabic



(b) English

Evaluation of the different components

In the following table, each component of the system has been evaluated.

Compo	**R**	**P**	**WER**	**BLEU**	**Test Corpus**
Video Sum	0.13	0.36	X	X	50 Seq
ALASR	X	X	14.02	X	31K Sent
MT	X	X	X	0.39	3K Sent

These results have been achieved on data not extracted from our video database except for the video summarization.

Evaluation on AMIS data

- ASR: Since there is no reference, we considered the transcription of Youtube and Euonews as reference. On a corpus test of 1300 sentences we get a WER of 36.5.
- MT: There is no reference, we translated the output of the ASR with Google and we considered it as a reference. The value of BLEU on a test set of 197 videos is 26.7.

End User Evaluation of Improved Components for Speech Recognition, Machine Translation and Video/Audio/Text Summarisation

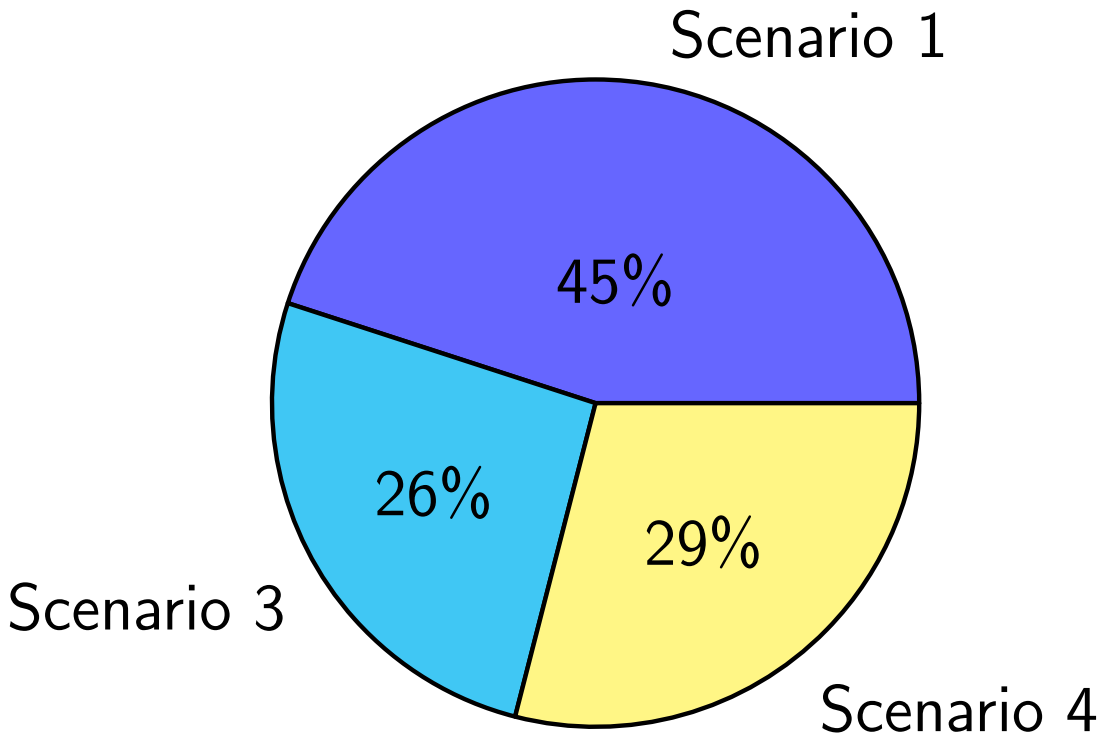


Figure: Subjective evaluation losers among 11 professionals

Evaluation of the quality of an Audio summary

Five people were asked to evaluate the quality of a summary as a whole and the quality of each segment independently from the whole summary.

Score	**Explanation**
5	Full informative
4	Mostly informative
3	Half informative
2	Quite informative
1	Not informative

Table: Evaluation scale

Results

Sample	Length	Segments	Full Score	Average Score
1	3m19s	8	4.20	2.90
2	5m21s	13	3.50	2.78
3	2m47s	5	3.80	3.76
4	1m42s	5	3.60	2.95
5	8m47s	22	4.67	3.68
6	9m45s	30	4.00	2.49
7	5m23s	8	3.20	3.75
8	6m24s	20	3.75	2.84
9	7m35s	18	3.75	3.19
10	2m01s	4	2.75	2.63

Table: Audio summarization performance over complete summaries and summary segments

Conclusion

- The initial idea becomes a reality and we developed a prototype
- New scientific challenges arise: Audio summarization, Code-switched signal (speech recognition and machine translation), several people speaking simultaneously, etc.
- Subjective evaluation is under progress.
- Deployment of other architectures.
- Comparison of opinions.