An Integrated AMIS Prototype for Automated Summarization and Translation of Newscasts and Reports

Michal L Grega, Kamel Smaïli, Mikolaj Leszczuk, Carlos-Emiliano González-Gallardo, Juan-Manuel Torres-Moreno, Elvys Linhares Pontes, Dominique Fohr, Odile Mella, Mohamed Amine Menacer, Denis Jouvet

An Integrated AMIS Prototype for Automated Summarization and Translation of Newscasts and Reports

Michał Grega[1(✉)], Kamel Smaïli[2], Mikołaj Leszczuk[1], Carlos-Emiliano González-Gallardo[3], Juan-Manuel Torres-Moreno[3,4], Elvys Linhares Pontes[3], Dominique Fohr[2], Odile Mella[2], Mohamed Menacer[2], and Denis Jouvet[2]

[1] AGH University, al. Mickiewicza 30, 30-059 Krakow, Poland
grega@kt.agh.edu.pl
[2] Loria University of Lorraine, Nancy, France
[3] LIA Université d'Avignon et des Pays de Vaucluse, Avignon, France
[4] Ecole Polytechnique de Montréal, Montreal, Canada

Abstract. In this paper we present the results of the integration works on the system designed for automated summarization and translation of newscast and reports. We show the proposed system architectures and list the available software modules. Thanks to well defined interfaces the software modules may be used as building blocks allowing easy experimentation with different summarization scenarios.

Keywords: Integration · Video summarization · Speech recognition
Machine translation · Text boundary segmentation
Text summarization

1 Introduction

We live in a world in which information as abundant. It is extremely easy to access all kinds of data and news. The main problem is not the availability of information, but our possibility of the digestion. This calls for effective summarization techniques, which are able to extract the most vital information and present it in an effective way. An additional problem is the language barrier. We commonly speak and understand two to three languages, while there are tens of commonly used languages in the world.

In our research within the AMIS project we solve the problem of effective summarization and translation of video newscasts and reports. We aim at reducing the length of the newscasts and reports and, at the same time, provide automatically translated subtitles to fit the users requirements. At the time of writing of this paper we offer summarization and automated translation form Arabic and French to English.

In this paper we describe the software components used in the process of summarization of newscasts and reports as well as the proposed summarization

K. Choroś et al. (Eds.): MISSI 2018, AISC 833, pp. 415–423, 2019.
https://doi.org/10.1007/978-3-319-98678-4_42

architectures. We have created a flexible system, based on well defined interfaces between software module which allows us to experiment with different order of the building blocks of the summarization and translation system.

The problem of video summarization has been addressed before e.g by Zhang in [18,19]. Video summarization is an important research area in medicine [11] while for newscast and reports it was investigated by Gao [6]. Our initial research on the topic has been reported in [10], while the paper describing our research on the user's requirements is available at [3].

The rest of this paper is structured as follows. Section 2 describes the Content Database. Section 3 covers the different software architectures. In Sect. 4 we present software modules used in our system and the paper is concluded in Sect. 5.

2 Content Database

In order to gather the videos we have first identified a list of controversial Twitter hash tags, such as #animalrights or #syria. Than we have extracted all twits from the Twitter service that were identified with this hashtag for a given period of time. Further on we have filtered the twits – only the twits containing valid YouTube links were passed to the next stage in which the videos were downloaded and stored into the video database.

In total we have downloaded 310 h of video from 19 TV stations in 3 languages (English, French and Arabic). The total number of 5423 videos that vary from 1 to 64 min in length. We have downloaded all available image formats.

The metadata is stored in a MySQL database and the video files are stored in the file system with paths available in the database. Each video is identified by its unique ID – videoID (that is, at the same time, the video identifier used by YouTube).

3 Scenarios and Integration

The modular construction of the AMIS system allowed us to experiment with different summarization scenarios, denoted Sc1–Sc4. The integrated scenario has a form of a Python 3.5 script that accepts as an input information on the videoID of the source video and the desired summarization length (as number of seconds or percentage of the original video). The processing is fully automated.

3.1 Scenario 1 - Sc1

Figure 1 depicts the most basic approach to newscast summarization. In this approach in the first step we perform a Video Based Summary, which is based on Shot Boundary Detection. This step results in creation o a summarization recipe. In the subsequent step we employ video processing in order to generate a video summary in the source language. After this step we perform Speech Recognition (on the summarized video) and Machine Translation to the target language. Finally, we process the subtitles in order to generate a subtitle file in the target language for the summarized video.

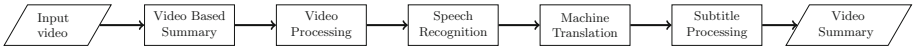

Fig. 1. Scenario 1 - the most basic approach to newscast summarization

3.2 Scenario 2 - Sc2

Figure 2 depicts an audio based approach to video summarization. In this scenario we generate a video summary recipe based on the audio signal, rather than visual cues as in the Sc1. The subsequent processing is identical as in Sc1. Based on the recipe file we generate a summarized video in the source language, which is than processed by the Speech Recognition module. The resulting transcription is automatically translated to the target language and finally the subtitle file is generated. At the moment of writing of this paper this scenario was not fully integrated, as work on the Audio Based Summary module was still ongoing.

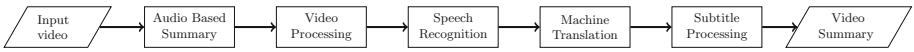

Fig. 2. Scenario 2 - audio based approach to newscast summarization

3.3 Scenario 3 - Sc3

Figure 3 shows a more advanced approach to video summarization. In this architecture first Speech recognition module is employed in order to obtain the transcription in the source language. Than the transcription is translated to the target language and summarized. Based on the recipe generated by the Text Summarization module a video summary is generated. Also, a subtitle file in the target language is provided.

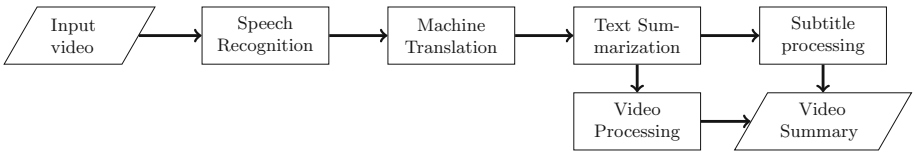

Fig. 3. Scenario 3 - text based summarization in target language

3.4 Scenario 4 - Sc4

This scenario, as depicted in Fig. 4, is a variation of Sc3. The core difference is that we perform text summarization on the source language, rather than on the target language, as in Sc3.

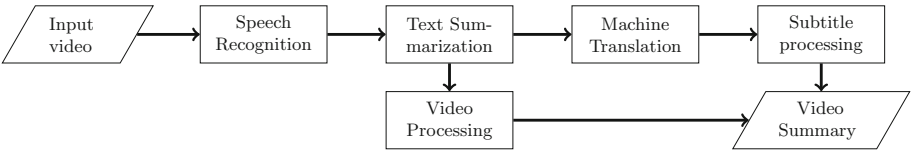

Fig. 4. Scenario 4 - text based summarization in source language

4 Available Modules

For the purpose of experimentation the whole AMIS system consists of seven separate modules. This modules, thanks to well-defined interfaces, may be connected in different configurations. This allows us to experiment with architectures in pursuit of the most effective one.

4.1 Video Based Summarization

The Video Based Summarization module role is to summarize a video based on its spatial and temporal activity. On input this module requires the video file to be summarized and the desired length of the video. The module fetches from the database a pre calculated information on the shot boundaries. It also fetches a pre calculated information on spatial and temporal activity. Based on the spatial and temporal activity the module selects most active shots, sorts them in the order of appearance in the original video and selects a number of shots to fill the required length of the summary. On the output the module provides textual description which is called 'summarization recipe' (explained later on).

4.2 Audio Based Summarization

The Audio Based Summarization module goal is to create a summary based only on the audio signal. This is a difficult task because in the audio signal there is no linguistic information (words, sentences, etc.) that would help to choose the informative parts of the video. We explore several neural architectures using deep learning in order to find the best features in the hidden layer. These features will allow us to capture the most important abstract structure (linguistic level) from a low level resource (signal).

4.3 Text Based Summarization

The Text Based Summarization module aims to select the most important information from a source transcript generated by the ASR module in order to produce an abridged and informative version of the original video. The input video transcript could be in any of the three languages involved in our research (French, English and Arabic).

We opted for an extractive summarization approach with the idea of finding the video segments that contain the most pertinent information based on the transcript. The main idea of Extractive Text Summarization (ETS) applied to video transcripts is to choose the most pertinent segments based on different criteria (information content, novelty factor and relative position) and arrange them in order of appearance to generate a shorter and informative version of original video. The Text Based Summarization module has been deployed based on ARTEX (Autre Résumer de TEXtes), an ETS system originally created by Torres-Moreno *et al.* for French, English and Spanish [17]. A Modern Standard Arabic (MSA) extension has been developed and added to ARTEX.

Sentence Boundary Detection. Before any ETS process, a Sentence Boundary Detection (SeBD) phase is needed to be performed in order to separate the unsegmented transcript produced by the ASR system.

We developed a Convolutional Neural Network (CNN) SeBD system based on textual features. In this approach, the SeBD problem is modeled as a classification task which goal is to predict if the middle word of a 5-word window is (or not) a sentence boundary [7]. During the training and testing phases we used subsets of the French, English and Arabic Gigaword corpora. The size and boundaries ratio of each corpus can bee seen in Table 1.

Table 1. Train/test Gigaword corpora

Language	Words	Boundaries	Ratio
French	587 M	56 M	9.54%
English	702 M	85 M	12.13%
Arabic	62 M	10 M	16.13%

Table 2. Performance of the AMIS SeBD

Language	Class	Precision	Recall	F1
French	NO_BOUND	0.976	0.984	0.980
	BOUND	0.838	0.768	0.801
English	NO_BOUND	0.969	0.983	0.976
	BOUND	0.856	0.762	0.806
Arabic	NO_BOUND	0.928	0.963	0.945
	BOUND	0.782	0.638	0.700

The performance in terms of Precision, Recall and F1 concerning the SeBD system over the Gigaword test datasets is shown in Table 2. For all languages, the "no boundary" (NO_BOUND) class reaches Precision and Recall scores over 0.92. Lower scores are reached for the "boundary" (BOUND) class, being the

lowest the Recall for Arabic (0.638). This behavior can be explained given the sample disparity between the "boundary" and "no boundary" classes.

4.4 Automatic Speech Recognition

As the objectives of AMIS are to summarize Arabic or French videos in English and to compare the opinion of these videos with the opinion of English videos dealing with the same topic, three Automatic Speech Recognition (ASR) modules were designed. For each language, an ASR system needs at least three components: an acoustic model, a language model and a lexicon with the different pronunciations of each recognizable word. The acoustic models are based on Deep Neural Networks (DNN) - more precisely on HMM-DNN models - and their development is based on the Kaldi recipe [15]. The language models are statistical n-grams models.

For acoustic model, the French ASR uses 40 MFCC (Mel-Frequency Cepstral Coefficients) acoustic parameters calculated on 25 ms windows every 10 ms. An i-vector of size 100 is added to them. The TDNN (Time Delay Neural Network) estimates 4000 senones (contextual states of Markov models) with a network composed of 6 hidden layers. The main advantage of the TDNN is its ability to take into account a broad context to estimate the probability of senones. In our implementation, we use a context of 29 frames (290 ms). The total number of parameters is about 11 million and the 33 acoustic models were trained on a TV and radio French corpus of 250 h. The 3-gram language model contains a total of 1 million grams trained from a text corpus of 1.5 billion words and the lexicon is composed of 96000 words. The French ASR achieved a Word Error Rate (WER) of 17.2% on the Ester2 development corpus [5].

The topology of the 40 English acoustic models is the same as for French and they were trained on 212 h of TED talks (www.ted.com). The 4-gram language model contains a total of 2 million grams trained from a text corpus of 140 million words and the lexicon is composed of 98000 words. The English ASR achieved a Word Error Rate (WER) of 12.6% on the TED-LIUM development corpus [16].

For the Arabic ASR, the topology of the neural network is different: 440-dimensional input layer (40×11 fMLLR vectors), 6 hidden layers composed of 2048 nodes each and a 4264-dimensional output layer, which represents the number of HMM states. The total number of weights to estimate is about 30.6 millions. The 35 acoustic models were trained on 63 h of broadcast news after a step of parameterization with 13-dimensional Mel-Frequency Cepstral Coefficients (MFCC) features and their first and second order temporal derivatives. The 4-gram language model contains a total of 1 million grams trained from a text corpus of 1 billion words and the lexicon is composed of 95000 words. The Arabic ASR achieved a Word Error Rate (WER) of 14.4% on a test corpus [12].

4.5 Machine Translation

The Machine Translation (MT) module role is to translate an Arabic word sequence into its English corresponding one. For integration, this module needs sequence of sentences, and sentences are translated one after the other. The module outputs translated sentences.

The MT module for AMIS has been developed for the direction Arabic–English, since Arabic is considered such as the foreign language of the video to translate to a summarized video in English. The MT system has been developped using the Moses [8] and Giza++ [13] toolkits, references of the statistical phrase-based approach [2,9] in machine translation.

The statistical approach requires a parallel corpus for training the translation and language models: we chose the Arabic—English United Nation corpus [4] (UN). The training corpus is made up of 9.7 million parallel sentences extracted from UN, concerning the period from January 2000 to September 2009. This corpus has been used to train the translation model. The language model has been trained on the target language of this corpus. The vocabulary contains 224,000 words. The development and the test corpus are composed of 3,000 parallel sentences. The evaluation on the test corpus leads to a BLEU [14] of 39.

4.6 Subtitle Processing

Before presenting video sequences with translated subtitles, they have to be converted to the format acceptable commonly by video players. We have decided to choose a widely used and broadly compatible *SubRip (SubRip Text, SRT) file format* with the extension `.srt`. SRT files contain formatted lines of plain text in groups separated by a blank line.

A dedicated Python script converts subtitles from an internal AMIS format into the SRT format. Furthermore, the script has an ability to split long subtitles (occupying more than 2 lines of text) into a series of shorter subtitles meant to be displayed one-by-one. The maximum number of words in a single subtitle has been empirically set to 17. Any longer subtitle will be split into a required number of shorter ones and displayed for a fraction of time being relative to the subtitle length.

4.7 Video Processing

Each Summarization module produces a recipe file. This recipe file contains information on shots and their order (defined by their start and end frame number) which are supposed to be included into the video summary. Based on this information the video processing module generates the final video summary. Using the popular ffmpeg software it strips audio of the original video, splits the original video and reconstructs the summary based on the recipe.

5 Summary

In this paper we have proposed a newscast and reports summarization system architectures together with a description of the software components used in the system. We have described components that perform speech recognition, video, audio and text based summarization, video processing, machine translation and subtitle processing. The system architectures were integrated in the Python language and are currently used for experimentation on video newscasts and reports summarization and translation. We foresee to work further on the system by both expanding the amount of available meta data (e.g. by incorporating face recognition [1]), modules and working further on optimizing the architecture of the system.

Acknowledgements. Research work funded by the National Science Centre, Poland, conferred on the basis of the decision number DEC-2015/16/Z/ST7/00559 under the Chist-Era AMIS project.

References

1. Baran, R., Rudzinski, F., Zeja, A.: Face recognition for movie character and actor discrimination based on similarity scores. In: 2016 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1333–1338, December 2016
2. Brown, P.F., Della Pietra, V.J., Della Pietra, S.A., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. Comput. Linguist. **19**(2), 263–311 (1993)
3. Derkacz, J., Leszczuk, M., Grega, M., Koźbiał, A., Hernández, F.J., Zorrilla, A.M., Zapirain, B.G., Smaïli, K.: Definition of requirements for accessing multilingual information opinions. Multimedia Tools Appl. **77**(7), 8359–8374 (2018)
4. Eisele, A., Chen, Y.: Multiun: a multilingual corpus from united nation documents. In: LREC (2010)
5. Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., Gravier, G.: The ester phase 2 evaluation campaign for the rich transcription of French broadcast news. In: Interspeech (2005)
6. Gao, X., Tang, X.: Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing. IEEE Trans. Circuits Syst. Video Technol. **12**(9), 765–776 (2002)
7. González-Gallardo, C.-E., Torres-Moreno, J.-M.: Sentence boundary detection for French with subword-level information vectors and convolutional neural networks. arXiv preprint arXiv:1802.04559 (2018)
8. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 177–180. Association for Computational Linguistics (2007)
9. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pp. 48–54. Association for Computational Linguistics (2003)

10. Leszczuk, M., Grega, M., Koźbiał, A., Gliwski, J., Wasieczko, K., Smaïli, K.: Video summarization framework for newscasts and reports – work inprogress. In: Dziech, A., Czyżewski, A. (eds.) Multimedia Communications, Services and Security, pp. 86–97. Springer, Cham (2017)
11. Leszczuk, M.I., Duplaga, M.: Algorithm for video summarization of bronchoscopy procedures. BioMed. Eng. OnLine **10**(1), 110 (2011)
12. Menacer, M.A., Mella, O., Fohr, D., Jouvet, D., Langlois, D., Smaïli, K.: Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect. In: ACLing 2017 - 3rd International Conference on Arabic Computational Linguistics, Dubai, UAE, pp. 1–8, November 2017
13. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. **29**(1), 19–51 (2003)
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
15. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, December 2011
16. Rousseau, A., Deléglise, P., Estève, Y.: Enhancing the TED-LIUM corpus with selected data for language modeling and more ted talks. In: 9th International Conference on Language Resources and Evaluation (LREC 2014), Interspeech (2014)
17. Torres-Moreno, J.-M.: Artex is another text summarizer. arXiv preprint arXiv:1210.3312 (2012)
18. Zhang, H.J., Low, C.Y., Smoliar, S.W., Wu, J.H.: Video parsing, retrieval and browsing: an integrated and content-based solution. In: Proceedings of the Third ACM International Conference on Multimedia, MULTIMEDIA 1995, pp. 15–24. ACM, New York (1995)
19. Zhang, H.J., Wu, J., Zhong, D., Smoliar, S.W.: An integrated system for content-based video retrieval and browsing. Pattern Recogn. **30**(4), 643–658 (1997)