Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 147-150 recibido 20-03-2018 revisado 29-04-2018 aceptado 03-05-2018

AMIC: Affective multimedia analytics with inclusive and natural *communication*

AMIC: Análisis afectivo de información multimedia con comunicación inclusiva y natural

Alfonso Ortega[1], Eduardo Lleida[1], Rubén San-Segundo[2], Javier Ferreiros[2], Lluis Hurtado[3], Emilio Sanchís[3], María Ines Torres[4], Raquel Justo[4]
[1] ViVoLab-Universidad de Zaragoza C/ María de Luna 1. 50018 Zaragoza
[2] GTH-Universidad Politécnica Madrid Ciudad Universitaria s/n. Madrid
[3] ELiRF-Universitat Politècnica València Camino de Vera s/n 46022 Valencia
[4] SPIN-Universidad del País Vasco, Campus de Leioa. 48940 Leioa
ortega@unizar.es, ruben.sansegundo@upm.es, lhurtado@dsic.upv.es, manes.torres@ehu.eus

Abstract: Traditionally, textual content has been the main source of information extraction and indexing, and other technologies that are capable of extracting information from the audio and video of multimedia documents have joined later. Other major axis of analysis is the emotional and affective aspect intrinsic in human communication. This information of emotions, stances, preferences, figurative language, irony, sarcasm, etc. is fundamental and irreplaceable for a complete understanding of the content in conversations, speeches, debates, discussions, etc. The objective of this project is focused on advancing, developing and improving speech and language technologies as well as image and video technologies in the analysis of multimedia content adding to this analysis the extraction of affective-emotional information. As additional steps forward, we will advance in the methodologies and ways for presenting the information to the user, working on technologies for language simplification, automatic reports and summary generation, emotional speech synthesis and natural and inclusive interaction.
Keywords: Audio, Speech, Language, Multimedia Analytics, Affective, natural inclusive communication

Resumen: Tradicionalmente, el análisis de los contenidos textuales ha sido la principal fuente de extracción y catalogación de contenidos multimedia y a él se han ido sumando tecnologías que son capaces de extraer información del audio y del video. Un nuevo eje de análisis es la vertiente emocional-afectiva intrínseca en la comunicación humana. Esta información de emociones, posiciones, preferencias, lenguaje figurativo, ironía, sarcasmo, etc. Es fundamental para una comprensión total del contenido de conversaciones, discursos, debates, etc. El objetivo de este proyecto se centra en avanzar en el desarrollo y mejora de prestaciones de las tecnologías del habla, el lenguaje, la imagen y el vídeo para el análisis de contenidos multimedia y añadir a este análisis la extracción de información afectiva-emocional. Como pasos adicionales, se avanzará en los métodos de presentación de resultados al usuario, trabajando en tecnologías de simplificación del lenguaje, generación automática de resúmenes e informes, síntesis de voz emocional e interacción natural e inclusiva.

Palabras clave: Audio, voz, lenguaje, análisis multimedia, afectivo, comunicación natural inclusiva.

1 Project consortium

In this project are involved four partners:
- Universidad de Zaragoza (ViVoLab).
- Universidad Politécnica de Madrid (GTH).
- Universidad Politécnica de Valencia (ELiRF).
- Universidad del País Vasco (SPIN).

The project is founded by the Ministerio de Economía y Competitividad under the grant TIN2017-85854-C4-1-R and lasts for three

ISSN 1135-5948. DOI 10.26342/2018-61-19

years. It can be considered a natural extension of the previous one ASLP-MULAN project (Ferreiros et al., 2016), where the same partners were involved.

2 Introduction

The scientific community is highly interested in developing new effective and efficient content-based indexing and retrieval methods and techniques to extract all the information from multimedia data. The AMIC project aims to advance on audio, speech and language technologies as well as image and video technologies in the analysis of multimedia content adding to this analysis the extraction of affective-emotional information providing to the user a natural and inclusive interaction.

The challenges and goals of affective computing and multimedia analytics must merge to accomplish the task of providing a comprehensive analysis of each topic of interest. Automatic extraction of the information contained in multimedia data including those pieces of information coming from social media is essential nowadays for multiple purposes. Affective information plays a key role in audiovisual and textual information that must be taken into account in order to assist multimedia computing (processing, indexing, retrieval …). Despite of the progress in multimedia content analysis there is much work to be done in the field to provide effective ways of integrating all possible sources of information in this analysis: emotional cues, figurative language, stance, preference, reputation…

3 Technologies involved in the project

The main technologies involved in the AMIC project are:

- *Machine learning*: Recent advances in deep learning (Hinton, Simon & Teh, 2006) have provided a great progress in the field of machine learning, especially in tasks where we have to deal with raw signals like audio processing (Hinton et al, 2012), image classification (Krizhevsky, Sutskever, & Hinton, 2012), natural language processing (Mikolov et al., 2013), but there are new challenges when dealing with multimedia content, and these techniques must be adapted to different domains and contexts.
- *Speech technologies*: In the last years, deep learning has had a great impact on ASR core technology and its applications (Deng, 2016), (Amodei et al., 2016), but there are new challenges when dealing with multimedia content: open and dynamic vocabularies, adaptation to new speakers, and adaptation to new acoustic environments.

In the field of language and speaker identification/diarization also Deep Neural Networks (DNNs) have been recently used, mainly as a bottleneck feature extraction, in order to improve the behavior of the classical i-vector approaches (Miguel et al., 2017), (Viñals et al., 2017), (Martínez-González et al., 2017).
On the other hand, emotion analysis can be considered from the point of view of speech recognition, by means of classifiers based on features associated to emotions, or speech synthesis using generative models that can be adapted to different emotions, styles and expressivity intensities (Lorenzo-Trueba et al, 2015).

- *Natural language technologies*: Two main areas can be considered in the scope of this project. On the one hand, information about the emotional states of the speaker can be extracted from the speech transcriptions by using statistical multilayer classifiers or deep learning methods. The intentionality of the text can be also identified by detecting figurative language forms, like sarcasm, irony, nastiness, or humor (Justo et al, 2014) (Hurtado et al, 2017). This information is very useful for the Sentiment Analysis problem, where phenomena such as irony, sarcasm and some types of humor like pun, limits the accuracy of the systems and new methods and resources must be used.

On the other hand, a deeper analysis of the text, based on its semantics, is necessary to tackle with important task as Social Media Analytics. Sentiment Analysis at different levels, user profiling, and stance detection are some interesting problems that can be addressed for monitoring social networks as Twitter, Facebook, or Instagram. One of the successes of the neural networks applied to the Natural Language Processing (NLP) area is to model complex relationships between the words of a document generating continuous representations rich in syntactic/semantic information. Unsupervised generation of these continuous of words and phrases (word

embeddings) take into account the complex relationships between the words of a document. Knowledge graphs could be used to enrich the representation of the semantics of words.

- *Audiovisual technologies*: The analysis of visual content can help to catalogue the type of scene, to segment the video in clips or to search for specific concepts. This information can be used for video retrieval, or to improve the video content analytics.

  Multimedia content summarization refers to text, audio and video-based summarization. The aim is to produce a condensed representation that captures the core meaning. Recent advances of DNNs have given promising results by using convolutional neuronal networks for feature extraction and Long-Short Term Memory (LSTM) to model temporal dependency among video frames (Zhang et al, 2016).

- *Inclusive communication technologies*: People can have difficulty in communication for many different reasons. Physical disabilities, motor co-ordination problems or learning difficulties can make hard to produce speech or handle spoken language. Augmentative and alternative communication (AAC) includes all forms of communication (other than oral speech) that are used to express thoughts, needs, wants, and ideas. Recently there are some approaches for text-to-pictogram translation using simple NLP tools at the sentence level (García et al, 2015). However, there is still a lot of research to be done to make more useful document-to-pictograms translation, where keyword extraction, text summarization and language simplification are key NLP tools.

4 Project objetives

4.1 Strategic objectives

The main strategic goal is to progress a set of diverse technologies and use them to deal with affective analysis on multimedia documents and affect-aware person-computer interactive systems. We are committed to contribute to an improved study on all kind of sources of information including traditional broadcast media and new massive and heterogeneous social media. We aim at proposing novel technological solutions to support a comprehensive information extraction of multimedia sources that includes developing audio, image, speech and language technologies devoted to: multimedia information extraction and processing, affect aware multimedia analytics, and natural, affective and inclusive communication. Additionally, we are committed to find efficient methods to manage and integrate these new sources of information into multimedia analytics systems and to provide effective ways of including affective aspects into natural and inclusive human machine interaction systems.

4.2 Scientific-Technological objectives

Following the project structure, our scientific-technological goals are:

- To develop technologies for audio, video, speech and text processing intended to
  1. Transcribe the speech content of multimedia documents into text.
  2. Use Web of Data as a source of knowledge to improve language, understanding, and aspect-based polarity models.
  3. Identify the language and the speaker automatically from the audio.
  4. Analyze the video of each multimedia document to extract useful information such as scenarios or characters.
- To develop technologies for affective analysis intended to
  1. Extract emotional cues from video, text and audio documents
  2. Study the impact of multimedia content on users while they are watching or listening to this content.
  3. Process figurative language, detecting and interpreting pun, irony and sarcasm.
  4. Automatically detect the stance of people involved in conversations, identify the reputation of an institution or company and track trends in social media.
- To develop technologies for natural, affective and inclusive communication devoted to
  1. Automatically generate reports and summaries out of the information extracted from multimedia documents using simple language.
  2. Synthesize speech with affective aspects such as expressivity control through emotion and style transplantation.

3. Develop person-computer interactive systems taking into account emotional and inclusive aspects including alternative and augmentative communication.

4.3 Transferring knowledge objectives

We propose three main objectives related to the knowledge transfer to the society:

- To develop and evaluate an application demonstrator: Affective Multimedia analysis platform with inclusive and natural interface
- To develop multimedia annotated resources and software tools freely available.
- To train experts in the developed technologies that may be employed by companies interested in our results.

Acknowledgments

This work is supported by Ministerio de Economía y Competitividad under the grants TIN2017-85854-C4-(1, 2, 3, 4)-R.

References

Amodei, D., S. Ananthanarayanan, R. Anubhai, , J. Bai, E. Battenberg, C. Case and J. Chen. 2016. Deep speech 2: End-to-end speech recognition in English and mandarin. In Int. Conf. on Machine Learning, pp. 173-182.

Deng, L. 2016. Deep learning: from speech recognition to language and multimodal processing. APSIPA Transactions on Signal and Information Processing.

Ferreiros, J., J.M. Pardo, L.F. Hurtado, E. Segarra, A. Ortega, E. Lleida, M.I. Torres, and R. Justo, 2016. ASLP-MULAN: Audio speech and language processing for multimedia analytics. Procesamiento del Lenguaje Natural, Vol 57, pp.147-150.

García P., E. Lleida, D. Castán, J.M. Marcos, and D. Romero, 2015. Context-Aware Communicator for All. In Universal Access in Human-Computer Interaction. Lecture Notes in Computer Science, vol 9175. Springer.

Hinton, G. E., O. Simon and Y.W. The, 2006. A fast learning algorithm for deep belief nets. Neural computation 18.7, pp. 1527-1554.

Hinton, G., L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen and T.N. Sainath 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups, Signal Processing Magazine, IEEE, vol. 29, no. 6, p. 829.

Hurtado, L., E. Segarra, F. Pla, P. Carrasco and J.A. González 2017. ELiRF-UPV at SemEval-2017 Task 7: Pun Detection and Interpretation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).

Justo, R., T. Corcoran, S. Lukin, M. Walker and M.I. Torres 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web, Knowledge-Based Systems 69.

Krizhevsky, A., I. Sutskever and G. Hinton 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems.

Lorenzo-Trueba J., R. Barra-Chicote, R. San-Segundo, J. Ferreiros, J. Yamagishi and J.M. Montero 2015. Emotion Transplantation through Adaptation in HMM-based Speech Synthesis. Computer Speech and Language. Volume 34, Issue 1, pp. 292–307.

Martinez-González, B., J.M. Pardo, R. San-Segundo, and J.M. Montero 2016. Influence of Transition Cost in the Segmentation Stage of Speaker Diarization. In Proc of Odyssey, Bilbao-Spain.

Miguel, A., J. Llombart, A. Ortega, and E. Lleida 2017 Tied Hidden Factors in Neural Networks for End-to-End Speaker Recognition. In Proc. of Interspeech.

Mikolov, T., K. Chen, G. Corrado, and J. Dean 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Viñals, I., A. Ortega, J. Villalba, A. Miguel and E. Lleida 2017. Domain Adaptation of PLDA models in Broadcast Diarization by means of Unsupervised Speaker Clustering. Proc. Interspeech 2017.

Zhang, K., W.L. Chao, F. Sha and K. Grauman 2016. Video Summarization with Long Short-term Memory, arXiv:1605.08110v2.