Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 147-150 recibido 15-03-2016 revisado 13-04-2016 aceptado 3-05-2016

ASLP-MULAN: Audio speech and language processing for multimedia analytics

ASLP-MULAN: Procesado de audio, habla y lenguaje para análisis de información multimedia

Javier Ferreiros, José Manuel Pardo
GTH-Universidad Politécnica Madrid
Ciudad Universitaria s/n. Madrid
jfl@die.upm.es

Lluís-F Hurtado, Encarna Segarra
ELiRF-Universitat Politècnica València
Camino de Vera s/n 46022 Valencia
lhurtado@dsic.upv.es

Alfonso Ortega, Eduardo Lleida
Universidad de Zaragoza
C/ María de Luna 1. 50018 Zaragoza
ortega@unizar.es

María Inés Torres, Raquel Justo
Universidad del País Vasco UPV/EHU
Campus de Leioa. 48940 Leioa. Vizcaya
manes@we.lc.ehu.es

Abstract: Our intention is generating the right mixture of audio, speech and language technologies with *big data* ones. Some audio, speech and language automatic technologies are available or gaining enough degree of maturity as to be able to help to this objective: automatic speech transcription, query by spoken example, spoken information retrieval, natural language processing, unstructured multimedia contents transcription and description, multimedia files summarization, spoken emotion detection and sentiment analysis, speech and text understanding, etc. They seem to be worthwhile to be joined and put at work on automatically captured data streams coming from several sources of information like YouTube, Facebook, Twitter, online newspapers, web search engines, etc. to automatically generate reports that include both scientific based scores and subjective but relevant summarized statements on the tendency analysis and the perceived satisfaction of a product, a company or another entity by the general population.
Keywords: Audio processing, speech recognition, language processing, sentiment analysis, emotional speech synthesis, multimedia social Web, information retrieval.

Resumen: Nuestra intención es generar la mezcla ideal de tecnologías del audio, el habla y el lenguaje con las de *big data*. Algunas tecnologías automáticas del procesado de audio, habla y lenguaje están adquiriendo suficiente grado de madurez para ser capaces de ayudar a este objetivo: transcripción automática del habla, métodos de búsqueda por habla, recuperación de documentos hablados, procesado del lenguaje natural, transcripción y descripción de contenidos multimedia no estructurados, resumen de ficheros multimedia, detección de emoción en el habla y análisis de sentimientos, comprensión de texto y habla, etc. Parece que merece la pena unirlos y ponerlos a trabajar sobre secuencias de datos obtenidos automáticamente procedentes de diversas fuentes de información como YouTube, Facebook, Twitter, periódicos digitales, buscadores de internet, etc. para generar informes que incluyan tanto puntuaciones basadas en análisis cuantitativo como expresiones resumidas subjetivas pero significativas sobre el análisis de tendencias y la satisfacción percibida sobre un producto, una empresa u otra entidad.
Palabras clave: Procesamiento de audio, reconocimiento del habla, procesamiento del lenguaje, análisis de sentimientos, síntesis de emociones, Web social multimedia, recuperación de información.

ISSN 1135-5948

1 About the project

This Project is founded by the "Ministerio de Economía y Competitividad" TIN2014-54288-C4 and there are four reseach groups involved: ELiRF (Universitat Politècnica de València), ViVoLab (Universidad de Zaragoza), SPIN (Universidad del Pais Vasco), GTH (Universidad Politécnica de Madrid).

2 Introduction

Society moves motivated by a lot of influences from fashions to established tendencies. Moreover, nowadays this movement is also highly modulated by the instant exchange of information fostered by social media far beyond TVs and radios. Internet social media sharing opportunities have reached a high percentage of the population and it is crucial, not only for the companies but for all economic and administration drivers in general, to know about the opinions, reputation feeling, political polarities and tendencies auto induced inside the social media. Having this information is relevant to drive new marketing policies and also have high relevance for security and defense in other contexts.

This relevance has been already detected by some companies that offer market surveys and reputation studies acquired in the social media by products, companies and other entities as political parties and administrations. The study is mostly based on costly polls and superficial hand analysis (as opposed to an automatic one) on a sampling on some limited sources using simple criteria in the analysis and we believe that they need a technological impulse to improve their capacities.

Big data science community has begun to apply their specific abilities to these data content analysis. In parallel, some audio, speech and language automatic technologies are now available or gaining enough degree of maturity as to be able to help to this objective. Some of these technologies are: automatic speech transcription, query by spoken example, spoken information retrieval, natural language processing, unstructured multimedia contents transcription and description, multimedia files summarization, spoken emotion detection and sentiment analysis, speech and text understanding, etc. They seem to be worthwhile to be joined and put at work on automatically captured data streams coming from several sources of information like YouTube, Facebook, Twitter, online newspapers, web search engines, etc. Out of this analysis, we will automatically generate reports that include both scientific based scores and subjective but relevant summarized statements on the tendency analysis and the perceived satisfaction of a product, a company or another entity by the general population.

Our intention is working in this direction and generating the right mixture of audio, speech and language technologies with big data ones as to be able to offer it to both the analytics companies interested in this information, improving their capacity to offer their services with increased quality, accuracy and usability of their reports. Also directly to companies or administrations willing to gain this information on their own via deploying our new solutions in their marketing or intelligence departments.

Relevant information about the feelings of the general public about products, companies and institutions can be distilled from multimedia information which is spread on the Internet on several content sharing applications like YouTube, Facebook, twitter, etc. Additionally, a lot of people is interested in finding the most appropriate documents related to a personal need of informative data connected with their interests or on products to be able to compare them to choose the best one under their own criteria.

From several potential ways to make use of social digital multimedia data we choose a couple of applications for the motivation and demonstration of this proposal: the retrieval of the most appropriate multimedia documents under a certain interest and the analysis of the opinion on a brand, person or event making use of multimedia files available in social digital media.

The problem is that most of this information is presented in an unstructured format: it may be directly in text or interleaved in the audio of a video or in the image of the video itself.

A lot of analytics effort is based on text or video, but audio, speech and language is not fully considered yet. We will use several audio, speech and language technologies to extract as much information as possible from these sources. First of all we face a problem of selecting the messages applicable to the specific search we are interested in. Information retrieval techniques including query by example will help us score and select the most appropriate. Then, we have to use techniques able to extract the

messages from the unstructured sources. After extracting the message, natural language processing has to be employed to obtain the semantics behind. Then, other kind of analysis have to be applied to analyze the emotions and polarities represented by these messages.

Afterwards, statistical analysis must be made in order to know the relevance of each analyzed tendency, polarity or sentiment. Finally, summarization techniques help us compose the report in the most usable format.

3 Project objectives

3.1 Strategic objectives

The main goal of this proposal is to explore the maturity of diverse technologies, progress or develop them when necessary as well as use them to deal with multimedia document retrieval and analysis, focusing on the information provided by audio and speech. We thus contribute to multimedia information retrieval and other use cases, for example, the building of automatic market and reputation analysis from social media sources. In this context our two main strategic objectives are:

- **Developing audio, speech and language technologies devoted to**
    1. Speech and audio information retrieval.
    2. Multimedia information analytics.
    3. Automatic output generation.
- **Transferring the acquired knowledge to the society trough dissemination and tec**hnology transfer actions.

3.2 Scientific-Technological objectives

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

- **To develop technologies for audio and speech processing intended to**
    1. Transcribe audio files, sometimes coming from videos, into text.
    2. Detect and Classify multimedia events from the acoustics
    3. Language and speaker identification and diarization
    4. Label emotions directly from the acoustics.
- **To develop technologies for audio and speech processing aimed to**
    1. Key term detection and concept discovery.
    2. Multilingual language understanding.
- **To develop technologies for multimedia analytics devoted to**
    1. Integrate developed audio, speech and language processing techniques for Multimedia information retrieval purposes
    2. detect, analyze and classify emotional states and language
    3. Analyze reputation, polarity and tendencies from multimedia social web.
- **To develop technologies for output generation and results presentation dealing with**
    1. Automatic report and summary generation
    2. Natural language generation
    3. Emotional speech generation

3.3 Transferring knowledge objectives

To deal with the second strategic objective we propose three main objectives related to the knowledge transfer to the society:

- **To develop and evaluate application demonstrators related with two use cases:**
    1. Multimedia information retreival
    2. Polarity and tendencies report
- **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.

4 Acknowledgments

This work is founded by the “Ministerio de Economía y Competitividad” TIN2014-54288-C4.

References

García F., L. Hurtado, E. Segarra, E. Sanchis, and G. Riccardi, “Combining multiple translation systems for Spoken Language Understanding portability,” in Proc. of IEEE Workshop on Spoken Language Technology (SLT 2012), 2012, pp. 282–289.

Hurtado L.F., J. Planells, E. Segarra, E. Sanchis, D. Griol (2010): “A stochastic finite-state

transducer approach to spoken dialog management". Proc. of Interspeech, pp. 3002-3005

Justo R., T. Corcoran, S. M. Lukin, M. Walker: "Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web". Knowledege-Based Systems. 2014

Justo R., M. I. Torres "Integration of complex language models in ASR and LU systems". Pattern Anal. Appl. 18(3): 493-505, 2015

Martinez F.F., J. Ferreiros, R. Cordoba, J.M. Montero, R. San-Segundo and J.M. Pardo " A bayesian networks approach for dialog modeling: The fusion bn". Proceedings of ICASSP 2009, pp. 4789-4792

Miguel A., J. Villalba, A. Ortega, E. Lleida, C. Vaquero "Factor Analysis with Sampling Methods for Text Dependent Speaker Recognition". INTERSPEECH 2014

Pla, F., L.F. Hurtado. "Political tendency identifcation in twitter using sentiment analysis techniques". Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics

Planells J., L.F. Hurtado, E. Sanchis and E. Segarra (2012): "An online generated transducer to increase dialog manager coverage". Proc. of Interspeech, pp. 1-4

Vaquero C., A. Ortega, A. Miguel, J. Villalba, E. Lleida "Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification". Interspeech, Makuhari, Japan. 2010