ROXANNE Research Platform: Automate criminal investigations

Maël Fabien[1,2]*, Shantipriya Parida*[1]*, Petr Motlicek*[1]*, Dawei Zhu*[3]*, Aravind Krishnan*[3]*, Hoang H. Nguyen*[4]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Switzerland
[3]Saarland University, Saarland Informatics Campus, Germany
[4]L3S Research Center, Leibniz University Hannover, Germany

mael.fabien@idiap.ch

Abstract

Criminal investigations require manual intervention of several investigators and translators. However, the amount and the diversity of the data collected raises many challenges, and cross-border investigations against organized crime can quickly impossible to handle. We developed ROXANNE Research platform, an all-in-one platform which processes intercepted phone calls, runs state-of-the-art components such as speaker identification, automatic speech recognition or named entity detection, and builds a knowledge graph of the extracted information. Our aim for this work is to do a first step in the direction of an open research platform combining speech, text, and video processing algorithms with criminal network analysis for combating organized crime.

Index Terms: criminal investigations, platform, speaker identification, speech recognition

1. Introduction

Automating the process of criminal investigations to speed it up can have a large impact on the daily work of police practitioners, and on the resources needed to monitor criminal organizations. However, it remains a complex task, because of the multimodality of data (e.g. intercepted telephone calls, text mostly, but also CCTVs), the plurality of languages to handle, and the lack of realistic training data matching this domain. The knowledge that police practitioners can build in an investigation is also very broad. Knowing only the identities of who spoke to whom remains of very limited interest for them.

It is therefore essential for an automatic tool, to reflect the content of the conversations, their topics, the entities mentioned (places, names...), whether a conversation comprise a suspect or not (target and non-target calls), but also to identify the names of the speakers (potentially their age, gender, ...) and the names of the people they mention in their conversations, all of that being represented in a knowledge graph.

In the ROXANNE consortium, a European Union's Horizon 2020 research project leveraging real-time network, text, and speaker analytics for combating organized crime[1] (grant agreement No. 833635), we developed a tool that represents the building blocks of what we hope will become a standard platform in the investigation process. Several Law Enforcement Agencies (LEAs), member of the consortium, have already installed and use the platform to further improve its performance.

We developed this platform with the core idea of being easy to install and fine-tune, relying on Python and PyTorch scripts, with a graphical user interface, running locally only (for privacy reasons). Part of the aim of the ROXANNE project is also to leverage multimodal fusion of data. For example, in a prior work, we showed that the knowledge graph conveyed information on the links between speakers that could improve the performance of a speaker identification module [1]. All components communicate via intermediate files that are stored, and research experiments are therefore faster to implement.

[1]https://www.roxanne-euproject.org

2. Technologies

Figure 1 displays the workflow of technologies integrated in the first version of platform. Once a telephone call is intercepted and uploaded to the platform, a speaker diarization module is applied to segment the telephone conversation into speaker segments (i.e. if not available apriori). The speaker diarization module implemented relies on the Bayesian HMM clustering of x-vector sequences (VBx) introduced by Landini et al. [2].

Then, an open-set speaker identification module attributes the calls to existing or new speakers. The enrollment is done gradually (we enroll the first two speakers after the first call), in a multi-session fashion, and if the identity of a speaker is assessed with a given score, the model considers this new recording as being a second enrollment for this speaker, hence building robust speaker representations over time. The speaker identification system we integrated in the platform is state-of-the-art ECAPA-TDNN architecture, trained on VoxCeleb [3] and VoxCeleb 2 [4] datasets, available through SpeechBrain's library [5]. This step builds the core of the knowledge in the knowledge graph. Speakers are represented as nodes, and links between them as edges.

We also extract information related to the gender of the speaker, using conventional Gaussian mixture model on the MFCC features, and to the age of the speakers, and add these information on the nodes.

Automatic speech recognition (ASR) is performed on the intercepted telephone calls afterwards. The ASR system implemented in the platform is a Wav2Vec 2.0 XLSR architecture [6]. The ASR model was fine-tuned for several languages of interest for the LEAs of the consortium, including English, French and Hebrew. Transcripts are added as information to the edges, as well as metadata related to this specific call (date, time, telephone numbers...).

Various Natural Language Processing (NLP) tasks are then performed on top of the ASR output. First, entities are extracted from the text using a named entity recognition (NER) model. The model implemented is a fine-tuning of BERT [7] with an additional conditional random field layer, trained end-to-end, with a distant supervision for low resource languages, as pre-

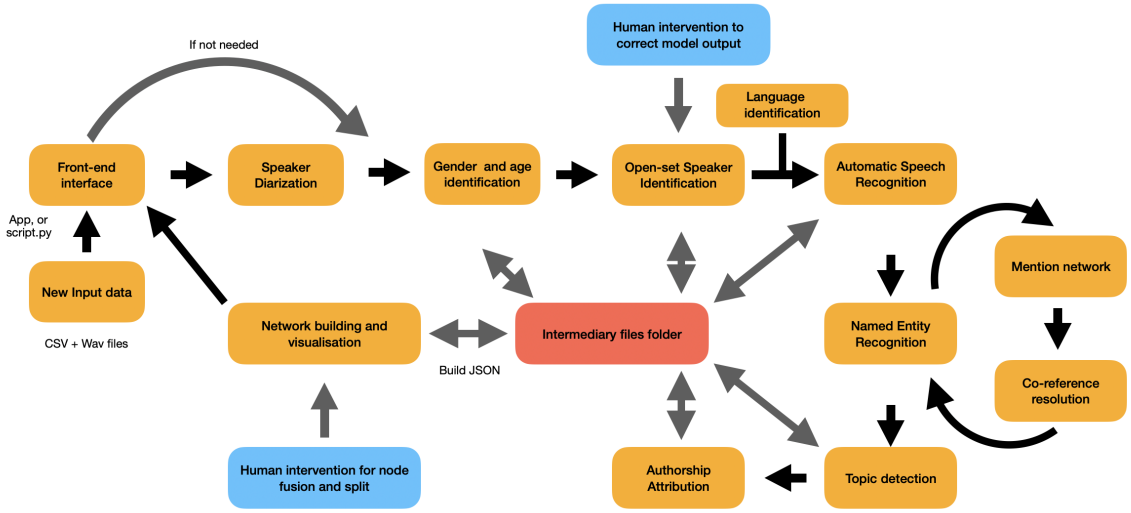

Figure 1: *Data processing in the platform.*

sented in [8]. Based on extracted entities, a mention network is added to the knowledge graph. The idea behind the mention network is that when two speakers, A and B, mention another person C, an edge should be established between A and C, and B and C. This is done by filtering only the "person" class from the named entities. However, if two speakers mention each other's names, it might lead to a confusing situation where edges are drawn between the speaker identification cluster and the real name of the speaker, suggesting that they are two different entities, which should be avoided. To do so, we improved the NER module and added NeuralCoref, a co-reference resolution module, that can identify the links between the names mentioned and the pronouns used. Places, dates and money amounts are also captured as metadata related to the conversation.

Topic detection is also an important component of the platform we developed. Typically, topic detection models rely on Latent Dirichlet Allocation and require certain amount of annotated data for training. As each criminal investigation is different, the topics of interest differ every time. To overcome the burden of having to annotate thousands of conversations each time, we proposed a method to LEAs in which they can provide their own topics, and the classification is performed in a zero-shot fashion, using pre-trained BERT language models in specific languages.

Finally, we provide a machine translation (MT) module relying on mBART 50, a sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages using the BART objective [9]. This allows investigators to perform all the NLP aforementioned task in a different language from the intercepted one, but also to read the transcripts. Two additional components are also displayed in Figure 1, the language identification and the authorship attribution, but they are still being developed.

3. Human in the loop

Automating the investigation process involved combining several technologies that can all produce errors. The lack of relevant training data might also produce domain mismatch, although part of the aim of ROXANNE is to fill this gap by identifying or generating relevant data. In order for LEAs to trust the tool, human intervention might be required. This is done through our graphical user interface, at several levels:

- LEAs can modify the output of the speaker identification model, and attribute the utterance to another speaker. The utterance is therefore used as an additional enrollment data for the corrected speaker.
- LEAs can also modify the layout of the knowledge graph, by merging nodes for example, if two nodes have been identified by the investigator as being the same speaker. At the next iteration, the two speaker models are merged into a single one.

Future works will focus on allowing a wider range of prior knowledge inputs from investigators, and adapting the technologies afterwards. For example, LEAs could annotate some new named entities, and the NER module should be able to handle these new entities afterwards.

4. Fusion tasks

Research we conduct as part of the ROXANNE project is done on fusion of various tasks. Leveraging prior knowledge from the knowledge graph in the speaker identification task, doing a joint speech and text-based diarization or developing fusion methods for metadata such as age, gender, places are research directions currently considered.

This tool is used as a research platform by technical partners of the consortium, including LEAs benchmarking models on their proprietary data.

5. Conclusions

The tool introduced in this paper aims to set a standard pipeline for criminal investigation automation. We raise come challenges regarding training data and the need for human inputs, and presents some of the considered research directions.

6. References

[1] M. Fabien, S. S. Sarfjoo, P. Motlicek, and S. Madikeri, "Graph2Speak: Improving Speaker Identification using Network Knowledge in Criminal Conversational Data," *arXiv:2006.02093 [cs, eess]*, Sep. 2020, arXiv: 2006.02093. [Online]. Available: http://arxiv.org/abs/2006.02093

[2] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks," *arXiv:2012.14952 [cs, eess]*, Dec. 2020, arXiv: 2012.14952. [Online]. Available: http://arxiv.org/abs/2012.14952

[3] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," *arXiv:1706.08612 [cs]*, May 2018. [Online]. Available: http://arxiv.org/abs/1706.08612

[4] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," *arXiv:1806.05622 [cs, eess]*, Jun. 2018. [Online]. Available: http://arxiv.org/abs/1806.05622

[5] M. Ravanelli, T. Parcollet, A. Rouhe, P. Plantinga, E. Rastorgueva, L. Lugosch, N. Dawalatabad, C. Ju-Chieh, A. Heba, F. Grondin, W. Aris, C.-F. Liao, S. Cornell, S.-L. Yeh, H. Na, Y. Gao, S.-W. Fu, C. Subakan, R. De Mori, and Y. Bengio, "Speechbrain," https://github.com/speechbrain/speechbrain, 2021.

[6] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-lingual Representation Learning for Speech Recognition," *arXiv:2006.13979 [cs, eess]*, Dec. 2020, arXiv: 2006.13979. [Online]. Available: http://arxiv.org/abs/2006.13979

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, May 2019. [Online]. Available: http://arxiv.org/abs/1810.04805

[8] D. I. Adelani, M. A. Hedderich, D. Zhu, E. v. d. Berg, and D. Klakow, "Distant Supervision and Noisy Label Learning for Low Resource Named Entity Recognition: A Study on Hausa and Yor\'ub\'a," *arXiv:2003.08370 [cs]*, Mar. 2020, arXiv: 2003.08370. [Online]. Available: http://arxiv.org/abs/2003.08370

[9] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual Denoising Pre-training for Neural Machine Translation," *arXiv:2001.08210 [cs]*, Jan. 2020, arXiv: 2001.08210. [Online]. Available: http://arxiv.org/abs/2001.08210