Multimedia Tools and Applications

Enriching Videos with Automatic Place Recognition in Google Map

--Manuscript Draft--

Manuscript Number:	MTAP-D-20-03790
Full Title:	Enriching Videos with Automatic Place Recognition in Google Map
Article Type:	1201 - Video on Demand over Over The Top Platform
Keywords:	Named Entity Recognition, Video Retrieval, Enriching Video, Geographic Information Retrieval, Education with Media
Corresponding Author:	Francesca Fallucchi Guglielmo Marconi University: Universita degli Studi Guglielmo Marconi Rome, ITALY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Guglielmo Marconi University: Universita degli Studi Guglielmo Marconi
Corresponding Author's Secondary Institution:	
First Author:	Francesca Fallucchi
First Author Secondary Information:	
Order of Authors:	Francesca Fallucchi
	Rosario Di Stabile
	Erasmo Purificato
	Romeo Giuliano
	Ernesto William De Luca
Order of Authors Secondary Information:	
Funding Information:	

Noname manuscript No.
(will be inserted by the editor)

Enriching Videos with Automatic Place Recognition in Google Map

**Francesca Fallucchi · Rosario Di Stabile ·
Erasmo Purificato · Romeo Giuliano ·
Ernesto William De Luca.**

Received: date / Accepted: date

Abstract The availability of videos is growing rapidly in recent years. Finding and browsing relevant information which should be automatically extracted from videos is not an easy task, but today it is an indispensable feature due to the immense number of digital products available. In this paper, we present a system which provides an automatic information extraction process from videos. We show a system solution which uses a re-trained OpenNLP model to find out all the places and famous people mentioned in the video. The system queries the Google Knowledge Graph to get information about related relevant about Named-Entities like places and famous people. Furthermore, we present the Automatic Georeferencing Video (AGV) system developed by RAI (Radiotelevisione italiana, which is the national public broadcasting company of Italy, owned by the Ministry of Economy and Finance) Teche for the European Project "La Città Educante" (The Educating City: teaching and learning processes in cross-media ecosystem) Our system contributes to the Educating City project providing the technological environment for creating statistical models for automatic named entity recognition (NER), acting in the educational field using the Italian language. The presented system has been used in the innovative learning challenges that the world of education with media is taking forward showing how is useful education with thematic newscast with scientific content.

F. Fallucchi, R. Giuliano, E. W. De Luca
Guglielmo Marconi University
via Plinio 44 Rome, Italy
E-mail: f.fallucchi@unimarconi.it E-mail: r.giuliano@unimarconi.it

R. Di Stabile
RAI S.p.A. Viale Giuseppe Mazzini, 14 Rome, Italy E-mail: rosario.distabile@rai.it

F. Fallucchi, E. W. De Luca, E. Purificato
Georg Eckert Institute Cer Straellße 3 Braunschweig, Germany E-mail: erasmo.purificato@ovgu.de E-mail: deluca@gei.de

Keywords Named Entity Recognition · Video Retrieval · Enriching Video · Geographic Information Retrieval · Education with Media.

1 Introduction

Today, technology enables teaching to transcend the boundaries of traditional teaching, opening up new learning opportunities for students.

From smart touch screens to lecture capture technology, new tools have made learning more flexible and accessible, while uprooting some commonheld ideas of how education works. Nowadays, as stated by Alex Parlour (B2B Marketing and Communication Manager at Sony Professional Solutions Europe), 94% of educators recognise the power of technology in improving student engagement levels. Advances, in particular in video technology, have demonstrated how new tools can improve students' engagement, enabling them to learn remotely or through interactive content. Video technology has played a key role in meeting expectations of this generation; in fact, 93% of institutions believe that video has improved student satisfaction [1]. Just think of *VideoLecture.NET*[2] portal, the world's biggest free and open access educational video lectures repository, through which teachers, or scientists in general, can share their lessons or conference talks.

Artificial Intelligence (AI) will increasingly play a key role in the futureproofing of video tools. One of the most active fields of study in this sense is certainly the Video Retrieval, that aims to retrieve a set of videos as a result of queries performed by text, images or even other videos. Determining the content of the video is even more important if you want to use the videos to deliver that content for learning. Identifying places and famous people who are protagonists in a video can be a very useful activity to consider it suitable for education.

In order to select the appropriate video, which corresponds to the text query, an important part is acted by text extraction: video-text extraction is identified as one of the key components of the video analysis and retrieval system [1]. Such extraction of textual information can be carried out with a combination of Speech Recognition and Named-Entity Recognition (NER) processes. The aim of a Speech Recognition system it to make a machine able to "hear, understand, and act upon spoken information" [2]. NER is usually defined as the extraction and the identification as well as the classification of information about people, locations, and several other entities within documents [3]. Although in the related literature there are many important studies about NER [4,5], works on real data are limited [6–8]. Furthermore, in order to perform at their best, NER systems need well-structured texts (including punctuation) and without orthographic errors [9]. This field is today still under

[1] `https://www.avinteractive.com/features/comment/ai-technology-can-impact-inspire-students-higher-education-22-10-2019/`. Last seen October 7, 2020.

[2] `http://videolectures.net/`. Last seen October 7, 2020.

research and most of efforts are made for English text, making it quite difficult to deal with other languages. In our case, major contributes come from EVALITA[3], a periodic evaluation campaign of Natural Language Processing (NLP) and speech tools for the Italian language.

The huge amount of videos owned by RAI (or Radiotelevisione italiana, which is the national public broadcasting company of Italy) inspired a number of projects having the objective of contributing to innovation in teaching. The European Project "La Città Educante" ("The Educating City: teaching and learning processes in cross-media ecosystem") aims to help the creation of a novel education model based on the reciprocity of the involved parties, i.e. an educator is also educated in the very act of teaching.

In this paper, we present our contribution to the project, the Automatic Georeferencing Video (AVG) system, that is the realisation of a video retrieval framework that exploits information about people and places extracted with automatic speech recognition transcriptions and enriched with geographic data. Results are obtained performing spatial queries. This system, developed by RAI Teche[4], is freely available for every Italian teachers who request access[5].

The main focus of our work is about "mentioned entities" (people or places), instead of "recorded entities". A user can select a city or, more general, an area on the map, and look at all the videos in which that place is mentioned by the speaker or in which famous people related to that city are mentioned. Effective and comprehensive annotations are needed in order to quickly and easily access those contents. For our purpose we decided to use NER and Google Knowledge Graph enrichment for the geolocalisation of the videos. In the presented paper, we show result about a case study on videos from TGR Leonardo, the RAI newscast focusing on Science and Environment.

The paper is organized as following: in Section 2 we describe the state of the art related to several different fields including Geographic Information Retrieval, Named-Entity Recognition, Speech Recognition, and Video Retrieval. In Section 3 we describe the approach and architecture of our system explaining all details related to the three phases used to extract entities from the videos. We apply the idea of using AI in education in a real case study. In Section 4 we explain how thematic newscast can be used for building linkage between the formal and informal learning contexts. We show how it is possible to significantly reuse existing videos to provide a more engaging way of teaching. Finally in Section 5 we draw the conclusions and present open problems and the future work.

[3] `http://www.evalita.it/`. Last seen October 7, 2020.

[4] `http://www.teche.rai.it/`. Last seen October 7, 2020.

[5] `https://almacloud.inet2.org/GeoreferencingProject-1.0/index.html`

2 Related works

Digital contents is becoming important in all fields: from domestic use to industry, from health care to education, from library to television channels. A huge number of contributions are describing how for example advances in the smart environment monitoring system [10] opening up related information retrieval problems from classification of environmental sounds [11] to image registration [12] or to the accurate indoor localization of people visiting a museum or any other cultural institution [13]. Improving the quality of education through the diversification of contents and methods and promoting experimentation, innovation, the diffusion and sharing of information is another example of a domain where digital content can make its contribution. This area has inspired our work.

In this paper we propose the automatic recovery of the places and people mentioned in videos using geospatial queries. This research involves well-known research areas in its process: Speech Recognition, Video Retrieval, Named-Entity Recognition, and Geographic Information Retrieval as we detailed in the following.

Speech Recognition (also known as Automatic Speech Recognition or, Speech-to-Text) is a particular pattern recognition case. Nowadays, most of the state-of-the-art speech recognition systems are developed by company as Google, IBM, or Amazon and used in their voice assistant software and are generally based on a combination of dense and convolutional Long short-term memory (LSTM) [14,15]. Open source speech recognition software solutions have limitations [16] and one challenge in automated speech recognition is to determine domain-specific vocabulary [17]. Especially in newscast where new names occur frequently. That's why in our system we preferred a RAI proprietary solution, with a domain-specific vocabulary, that works well on our newscast videos corpus.

During the last two decades many studies have been performed concerning state-of-the-art methods and applications in **Video Retrieval** [18–21]. The aim of this field is to retrieve a set of videos related to a specific multimedia or natural language query given in input to the system. More recently, with the spread of deep learning techniques, works are focusing on natural language video retrieval [22–24], on learning joint text-video embeddings [25,26], and on retrieving specific portion of a video from a given text query [27,28].

In the context of extraction and identification, as well as classification, of information about people, locations, and several other entities within textual documents **Named-Entity Recognition** (NER) plays a crucial role [3]. In most languages, there is only a very small amount of labeled datasets usable for supervised learning models, and these systems are strongly anchored on manually annotated corpora yet [29]. For NLP and NER research on Italian language, great results are obtained from EVALITA, whose aim is "to promote the development of language and speech technologies for the Italian language, providing a shared framework where different systems and approaches can be evaluated in a consistent manner". Significant works on NER tasks are made in any edition of this biannual campaign [30,31].

When there are entities about places, their location is crucial. **Geographic Information Retrieval** (GIR) is generally considered as the specialization of Information Retrieval (IR) and Geographic Information Systems (GIS) that provide access to georeferenced information sources [32]. In the same paper, Larson defines *spatial queries* (also known as geographic queries or geospatial queries) as type of queries with spatial relationships to entities geometrically defined and spatially localized. GIS are available to common users only since last fifteen years with Google and Microsoft that introduced, in 2005, the most used on-line GIR services, i.e. respectively, *Google Maps*[6] and *Live Search Maps* (today *Bing Maps*[7]). The huge growth of Internet and available data lead to many possibilities for GIR research and applications, such as georeferenced media or geotagged pictures posted on social networks. Two comprehensive surveys [33,34], in addition to outline progress and challenges in spatial search, highlighted significant aspects of GIR systems: storing and browsing georeferenced data, mining semantic information from geographic data, and discover geographic location of pictures. In the last years, efforts about GIR systems have been made on extracting spatial information from the internet searches of users and from social media [35], and on integrating heterogeneous data sources to develop efficient multimodal systems [36].

3 Our Approach and System Architecture

In the following, we present our approach and the related architecture, which allows the automatic information extraction of Named Entities from videos. The implemented system is very complex and the main goal is to realize an annotated corpus which relies on a real dataset of raw videos. Therefore, the process starts with the video processing, extracting the text and recognizing the entities, then validating the results and presenting them with a user interface. We extract the related metadata associated to the videos using 5 components show in Figure 1. The first is the ***FFmpeg***[8], a standalone library, which is able to extract audio from video. We use ***Ants*** [37], a speech recognition system developed by RAI, to obtain transcripted information which has been uploaded. On the extracted text we use ***Apache OpenNLP***[9] for processing the named-entities and for providing them as a solution for further use. After that, the system queries the ***Google Knowledge Graph*** to get information about related relevant entities extracted. At the end, the user interface ***Automatic Georeferecing Video (AGV)*** shows the video with the extracted relevant information. AGV consists in a web service where with

[6] `https://www.google.com/maps`. Last seen October 7, 2020.

[7] `https://www.bing.com/maps`. Last seen October 7, 2020.

[8] `https://ffmpeg.org/`. Last seen October 7, 2020.

[9] `https://opennlp.apache.org/`. Last seen October 7, 2020.

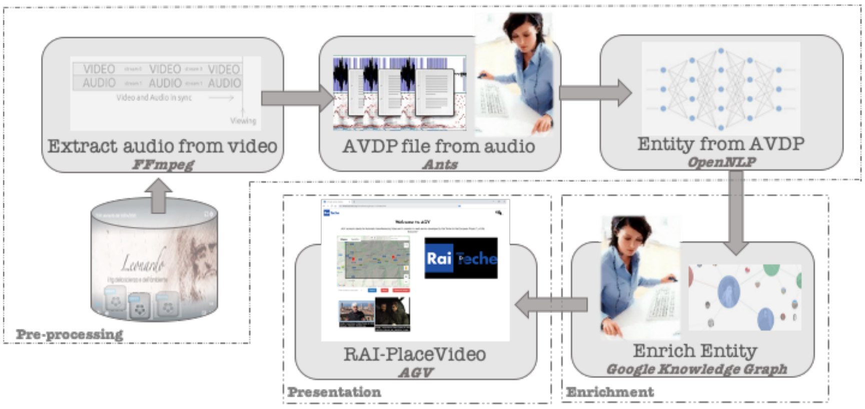


Fig. 1 Automatic Georeferecing Video (AGV) system architecture

a customized Google Map is possible to find places, cities, person which are mentioned in the video.

The process has three main phases: pre-processing phase, enrichment phase, presentation phase. In the pre-processing phase we extract both video and audio files. Then, we can extract text from speech and it presents as output a file XML in AVDP schema. They will be the input of the enrichment phase for the OpenNLP module, that extracts entities using its NER algorithms. In order to have each entity repeated once, multiple instances of the same entity are collected in a single object containing all the different timestamps indicating the position where the entity is mentioned in the video. The timestamp array will be used to perform the seek on the video played in the user interface. Each entity is enriched using Google Knowledge Graph API[10], to get: a representative image for the entity, a brief description, and the Wikipedia URL. Only for entities regarding places, we also retrieve information about their geolocalization, i.e. latitude and longitude. Finally, in the last phase we save data in a PostgreSQL[11] database and show to user on demand using a user interface. More details related to each phase are given in the following subsections: pre-processing phase (section 3.1), enrichment phase (section 3.2), presentation phase (section 3.3).

3.1 The Pre-processing Phase: the NE extraction

In the pre-processing phase we extract relevant content from the video. The pre-processing phase could be split in 3 actions: Extracting audio from video using FFmpeg (section 3.1.1), speech recognition and validation phase (section 3.1.2), the creation of AVDP files using Ants and the usage of OpenNLP to get entities in video (section 3.1.3).

[10] `https://developers.google.com/knowledge-graph`. Last seen October 7, 2020.

[11] `https://www.postgresql.org/`. Last seen October 7, 2020.

3.1.1 Extracting audio from video phase

In order to recognise entities from video we must first extract audio signals from video. FFmpeg is a complete command line system for this purpose and let us record, convert and play audio and video files.

Please note that sometimes the video file may be missing the audio track. In this case our system will not be able to find the entities. For our experimentation, we have used only the mapped audio tracks-videos information.

3.1.2 Speech Recognition and Validation Phase

In this phase we extract the text from the audio files and we create AVDP files using Ants. The output text from Ants has the AVDP schema, because it is simple to integrate for producing a good input file for the OpenNLP API. The Ants controls waveforms for audio track, identifies the video scene changes and associates the sentence to the scene. Both audio and video files, returned by FFmpeg, are used as an input for Ants, which returns an XML file containing the mentioned words. By Ants we can obtain a text with sentences, speakers, video shots, with the related timestamps. The output format respects the AVDP schema, which is part of the MPEG-7 standard. The Ants system can understand sound and returns it as a text in a transcripted file extracted from the audio stream. Different transcription errors may emerge during this process. Some examples are described in the following:

- Two or three words could be merged to one due to the too fast speech recognition;
- Organization names cannot be recognized because the system does not include them, because they are too new (it was created in years were these organization did not exist);
- Organization names cannot be recognized because the name of the organization has another name in Italian than the original one;
- The end of a phrase can not be recognized because there aren't:
  - the correct speaker's tone, or
  - the correct timing for punctuation elements (it happens often in flash news);
- The recognized person can not be recognized because it is too similar or related to a city (for example **Leonardo Da Vinci**)

For these reasons we use a manual validation phase to remove errors, made by a domain experts team.

3.1.3 Entity Extraction and integration in Apache OpenNLP

The last operation necessary for concluding the pre-processing phase is the entity extraction. In this last phase we identify specific information, the entities in the sentences extracted from the audio files and save them in a file XML in AVDP schema after it has been cleaned up by errors as described in the previous phase. We use the NER package of OpenNLP API to perform the task. In our work, due to the project requirements, we denote as valid entities regarding places (including every point of interest, not only cities) and people. After obtaining the entities from the developed service named EntitySearcher, we must remove duplicate entity and we use the developed service named JsonEntityCreator to merge the entities into a single Java object. For each entity, a set of timestamps will be created. The array of timestamps is used to perform the retrieval process on the video played in the user interface. For each entity, we send a query to the Google Knowledge Graph. The Google service creates a series of nodes related to the performed search. To improve NER we re-train an OpenNLP-based model. We first used an OpenNLP module without training, but it did not recognize any italian texts. For this reason, we created several examples for places using transcripts obtained from the previous phase. The OpenNLP system has a particular syntax, as we can see in dedicated site. An example of an input text is given in the following:

> Alla metà del 500 <START:person> I Medici <END> erano una casata andata in estinsione. L'ultima erede era <START:person> Caterina dei Medici <END>. Ella costruì la sua Dimora a <START:place> Parigi <END>. Quindi andò a <START:place> Venezia <END>.

As we can see we have start and end tags. With this tags OpenNLP understands which are the entities to be used:

- Persons
  - I Medici
  - Caterina dei Medici
- Places
  - Parigi
  - Venezia

Then, we create also a wrapper for the OpenNLP response. This wrapper gets a text as an input and returns a JSON array object. In the following we reported an extract of JSON referred to the previous example:

```
[
  {
   "entity":"Giovanni",
   "from":11,
   "tag":"person",
   "probability":0.87728355595
  },
  {
   "entity":"Aosta",
   "from":41,
   "tag":"place",
   "probability":0.94728355595
  }
]
```

Each JSON object presents an entity enriched by the OpenNLP model. For each entity we can see the following keys:

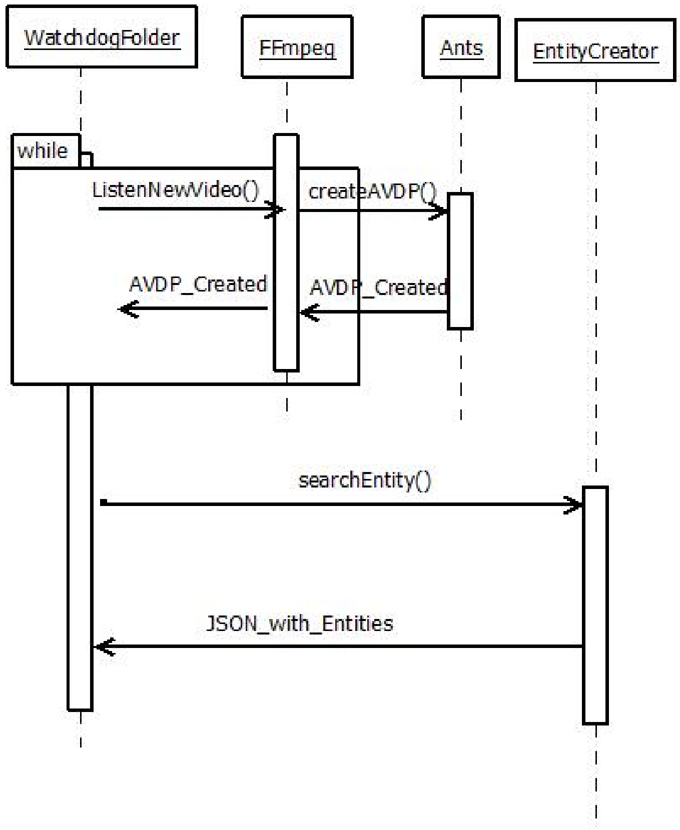


Fig. 2 Sequence diagram for Speech Recognition and validation phase.

- Entity represents name entity
- From represents index from entity starts in text. In our example, we can see Giovanni after eleven characters
- Tag could to have two choices:
  - person if the entity is a person
  - place if the entity is a place
- Probability. In this work we use a statistical model which included probabilities for every categorized word. We consider only entities with probability higher than 0.7.

In Figure 2 it is reported the sequence diagram to create JSON with entities related to the video file.

3.2 The Enrichment Phase: Google Knowledge Graph and Validation

The enrichment phase consists in using the queries of the Google Knowledge Graph to get some information about entities identified in previous phase.

It is clearly possible that errors occur due to NER and Google Knowledge Graph. For this reason at the end of this phase we insert also a validation phase. It allows correcting possible problems on the entities generated by the two previous processes before the information is displayed in the presentation phase. In the rest of the section, we first introduce how we enrich entities using Google Knowledge Graph API (section 3.2.1) and then how we create a user interface to allow manual validation for domain experts (section 3.2.2).

3.2.1 Enrichment using Google Knowledge Graph

After the NER process, the Google Knowledge Graph is used to enrich, with the additional information, the entities extracted from the video. As first step of this phase, we extract the timing where each word related to the entity is used in video. So, we obtain an array sorted by the timing for each entity recognized in extracted transcript text. After the assignment of the timing, we use Google for searching for an entity. In this phase, we could encounter some ambiguity errors. We solve this ambiguity errors querying Google only for places and people. Other ambiguity issues are solved in the validation phase (section 3.2.2). In order to discover some information about entity, we used a query like this insert in the following URL: `https://kgsearch.googleapis.com/v1/entities:search?query=celine+dion&key=API_KEY&limit=1&indent=True`. With this query we retrieve the results from the Google Knowledge Graph search for the relevant entities. For instance, we specify the query "Celine Dion" and we use an API_KEY. We add some parameters to limit the specification and retrieve only 1 result. Using indent parameter, we are specifying we would an indented response. Google Knowledge Graph returns a JSON-LD response like this:

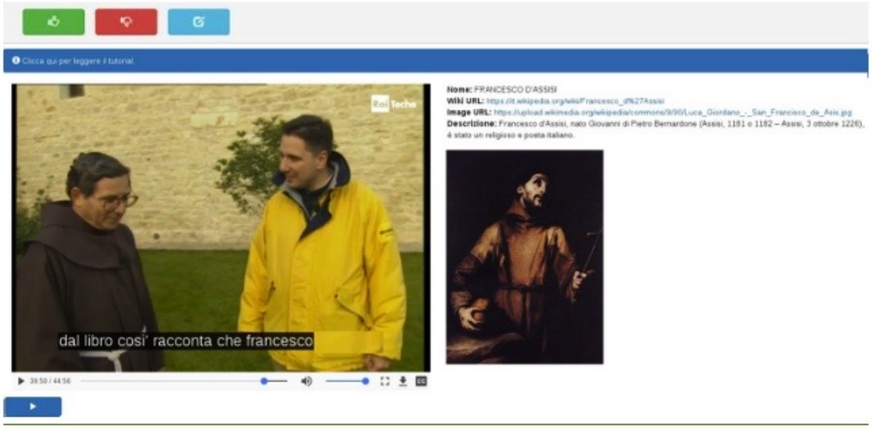
```
{
"@context":{some google context information},
"@type":"itemList","itemListElement":
 [
   {
    "@type": "EntitySearchResult",
    "result":
    {
     "@id":"kg:/m/01cwhp",
     "image":
     {
      "contentUrl":"https://urlImage",
      "url":"https://urlImageWikipedia",
     },
     "name":"Celine Dion",
     "@type":[ "person"],
     "description":"Canadian singer",
     "detailedDescription":
         {
      "license": some information license,
      "url":"https://wikipediaUrl/CelineDion",
      "articleBody":"some information"
     } "resultScore":2793.68
```

```
      }
    },{another result for string searched}
  ]
```

Note that we could have more result for a search query because, for example, we could search using only surname, thus Google will retrieve us all famous persons having this searched surname. In the JSON-LD we see these information: some information for Google, information regarding image related entities, complete name of the searched entity, type of the entity (google knowledge graph uses a lot of type of entities), the page where we can find information about the entities, a brief description of the searched entity, and result score, a score indicating how much this entity is pertinent for our search. High score means high pertinence. In our case, for example, a place has a JSON with coordinates. Using a JSON-LD for place type, we could geo-reference a video because, from a word place, we can save in our database a point with its coordinate. We use this coordinate point to retrieve a video and drawn a rectangle in the user interface. After we retrieve information of an entity presented in each video, we can create a JSON for each video saved in our repository. In this JSON we can recognize some information for the video such as: the path where video is saved, an image representing video, the video's id in database where we can see video information, a JSON with two arrays: one for people information and one for place information. Examples of information are: an URL, the entity name; an image, the timings where the speaker pronounces the entity, a brief entity's description. Only for the place entity, latitude and longitude where place is located. This JSON is returned and saved in database and it is used to show information for places and people in the player's layers, show in Figure 4.

3.2.2 Validation of the entity and its enrichment

Before showing the data to the user, we perform the validation of the entities and their enrichment. In this phase the errors due to NER and the enrichment due to Google Knowledge Graph API are solved. This is the most important phase, because it ensures that only real and correct entities will be shown to the end user. In this phase you will not discover new entities, unless they are part of the time slot of an incorrect entity. For example, if the system failed to retrieve the entity, it will be lost and you won't be able to find it. If the software has recovered an incorrect entity in a certain time slot, then the human validator can modify it appropriately. Let's suppose that the software has recovered the character Pope Francis, while in the video we talked about Francis of Assisi. In this case, the validator will notice that the wrong entity has been found and will perform the correct steps in order to eliminate the false positive, replacing it with the correct entity. At this stage, a user interface will propose to the validator an entity detected by the component. The validator will confirm if the entity has been detected correctly and will save it in the database or modify the entity. When the validation phase is finished and therefore all the entities will have been validated manually by the validator, then we will proceed to the editing phase of the JSONs related to the videos. In this phase, for each video, you will check which entities are connected to it and which timestamp each entity has. So, we will create the new JSON as it was previously presented in this document. For the validation phase we used the validator object. We developed a dedicated user interface for this purpose. The user interface for the validation of people is shown in Figure 3. In it you can view from top to bottom and from left to right button to approve the person in the video, button to discard the person, button to change the person, information concerning the person taken into consideration, etc.. By clicking on the green button on the top banner you will evaluate a person as valid. It will express the correct relationship between the one said in the video and the entity detected by NamedEntitySearcher. In addition, all initialized fields will be considered valid. For example, the button will be clicked when the entity Francis of Assisi is pronounced in the video and all the fields on the left of the interface are correctly initialized. By clicking on the red button we'll discard the person-type entity. The button will be clicked when no person type entity has been listed in the proposed time interval. This could happen due to ambiguities. An example could be the pronunciation of the words Santa Maria. In this case maybe in the video you are referring to the location in California, while our NamedEntitySearcher may have detected the entity Maria, the virgin of the Catholic religion. By clicking on the blue you will be able to manage the person to be validated. The click on this button will be done when the proposed entity is really a person, but maybe it has not been correctly initialized by our system. In a similar way, the human validator correct errors in places. We use this validation phase also to create an annotated corpus that can be use to train again the neural network to recognize the entity and to to measure system performance.


Clicca qui per leggere il tutorial
Rai Teche
dal libro cosi' racconta che francesco

Nome: FRANCESCO D'ASSISI
Wiki URL: https://it.wikipedia.org/wiki/Francesco_d%27Assisi
Image URL: https://upload.wikimedia.org/wikipedia/commons/9/90/Luca_Giordano_-_San_Francisco_de_Asis.jpg
Descrizione: Francesco d'Assisi, nato Giovanni di Pietro Bernardone (Assisi, 1181 o 1182 – Assisi, 3 ottobre 1226), è stato un religioso e poeta italiano.

Fig. 3 User interface to validate a person

3.3 The Presentation Phase: the User Interface

The presentation phase simply shows in a user interface entities with information obtained in previous phases. In this section we will show how we present the data extracted from video and enriched with previous phases. This component was developed for supporting the users (i.e. students, university professors and teachers) in recognizing places in videos. This component allows you to show video about a city and to present them in Google maps. So, with our system we can search, e.g. “Milano” and see all videos related Milano. Furthermore, we could find out which persons are relevant given the searched city. This is what we consider as a Geo-referencing video task. In this paper we use four objects given from Google maps:

- DrawingManager, for drawing a rectangle on the maps
- Searchbox, for searching a city on maps
- Pegman and Google Street View function, to view street and places as some image and walk on street like we are in this place
- Marker Place, to marker where the place is on maps.

All is presented using the Google Maps system and integrated in our system (Figure 4).

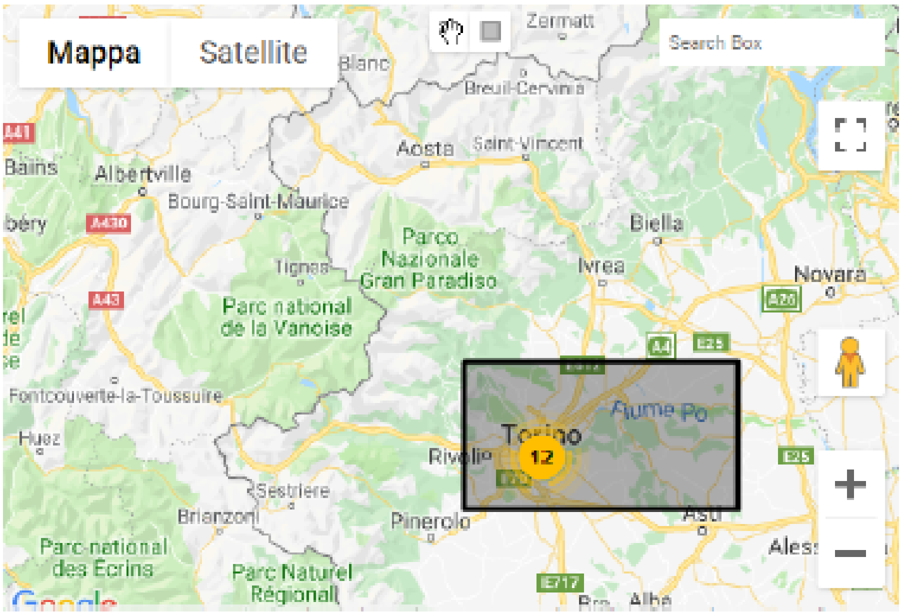
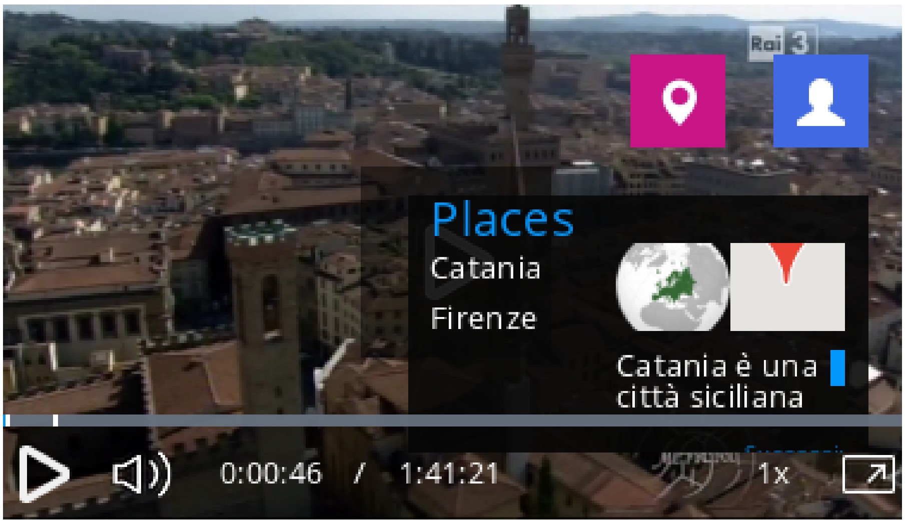


Fig. 4 Google Maps Objects in our system

We can see that on the left of Figure 4 buttons are used for visualizing the map. These buttons just present maps in different ways: satellite and cartographic maps. On the up-center we can see a drawing manager, which is able to draw rectangles on the map and a button with hand to move maps inside. In the right up corner, we have a searchBox, which can be used to easily change position on the map. Last but not least, we can see the pegman icon to start the street view process. When we draw a rectangle on the maps in our system, the user interface shows the videos describing the related places in rectangle area using a markerPlaces and clusters markerplaces. Clicking on a markerPlace the videos are started and the user can hear information about the related places. We can see how many occurrences are available for the places presented in the video. We can click on this result image to show video loaded in player which has been designed proprietary by our company. In our example we have just 1 occurrence for the city "Pavia" in the video. So, in the previous step, a user drew a rectangle around Pavia location and this is the resulting video. The player has two layers: the people and places layers. Clicking the places button, the user can see detail of the city of interest.



Catania è una città siciliana

Fig. 5 Places layer with the information related to the cities in the video

Figure 5 shows the work of the layer. In this layer we see, on the left side a list of places found in video we are playing. Zooming in place layer, we can see Information related to the cities in the video, such as: representative place images, markerplace on the map, brief place's description, link for a long place's description, and buttons to go previous and next timing where speaker says about selected place.

4 Case Study: Leonardo TGR RAI video

In the previous sections we described the proposed system, which extracts, enriches and shows information in videos. In this section we present the performance evaluation of this system on a real case study. It supports users with an innovative technological tool allowing the linkage between the formal and informal learning contexts.

For this aim, we created a dataset using more appropriate scientific contents in order to respond to the "La Città Educante" project that inspired the system. Then, we extracted randomly 10 minutes for each video of Leonardo TGR dataset. Leonardo TGR is a thematic newscast[12], which combines attention to current events with rigorous documentation and in-depth analysis and for this reason it provides a unique experience in Europe. Among the topics covered there are not only science and technology, but also health, environment, economy and society. In this way we created a video collection with scientific contents, that are more appropriate to respond to the considered project. To test our system we restrict the analysis on two entities person and places. In order to extract the geolocalization information on the corpus of videos, we produce two different neural networks: one for person (PER) and one for places (LOC). Then, we used the same extracted text (i.e. the transcription) as input, but with different training datasets. The training set was created as we discussed in previous paragraphs thanks the human domain expert validation, that manually annotate video corpus using a customize user interface (see section 3.3). In the transcripts we removed punctuation and used uppercase. The NER process was measured in term of precision. We do not calculate recall performance because it is too time consuming. For recall we need a human to read all transcripts and so he needs to count number of entities recognized from our system. For our purpose we need a high precision of the categorization, since it is important to have that all name entities recognized from system are correct (i.e. pertinent to the topic) rather than increasing the irrelevant entities. In the analysis, our dataset is composed by 6600 videos taken from Leonardo newscast. Each video has a time of about 10 minutes and we transcript it using Ants. This dataset is randomly divided in four subsets, choosing for each transcript a random subset. Experimentally, this number of subsets is the right trade-off between number of samples and number of examples. The precision for each subset was manually evaluated counting occurrences. It can happen the following cases:

- Correct identification and classification: entities are correctly identified in the text and they are classified in the correct category, too. For example the name entity "Celine Dion" is placed in the person category (instead of Place category);
- Correct identification but incorrect classification: entities are correctly identified in text but they are placed in a wrong category. For example "Celine Dion" is placed in the Place category;

[12] https://www.rainews.it/tgr/rubriche/leonardo/

	Tot	Errs	Category Errs	Tot Errs
Dataset 1				
LOC	2564	8.46%	1.64%	10.1%
PER	1056	22.9%	1.42%	24.3%
Dataset 2				
LOC	795	8.93%	0.25%	9.18%
PER	292	28.4%	0%	28.4%
Dataset 3				
LOC	1081	11.1%	1.11%	12.2%
PER	958	29.0%	1.15%	30.1%
Dataset 4				
LOC	1297	8.56%	0.62%	9.18%
PER	539	26.9%	0.37%	27.2%

Table 1 Errors in retrieval of entities in video (persons and places).

– Incorrect identification: entities are identified incorrectly in text. For example "Paris" is placed in person category.

The results of errors are reported in Table 1. In the first column we report NER for person (PER) and NER for places (LOC). The column *Tot* reports the total number of occurrences identified by our automated system. *Errs* is the percentage of the number of incorrect identifications, while *Category Errs* is the percentage of the number of correct identification but incorrect classification. Finally, *Tot Errs* is the percentage of the of total errors (i.e. the sum of incorrect identification occurrences and the occurrences of correct identification but incorrect classification).

In the following evaluation we focus on the precision because the evaluation of the recall would require too much effort for the annotators since to annotate 10 minutes randomly extracted from 6600 videos 4 persons were employed for 8 hours a day for 4 months. The results for precision are reported in Table 2 where in the first column, as for the previous table, we distinguish between predictor for person (PER) or places (LOC). The other columns are: the precision for identification process (*Precision*), the precision for classification process when it is correctly identified (*Category Precision*), and the overall precision (*Total Precision*), respectively. The overall precision is a measure of the identification precision and the classification precision. We consider the recognized name entity as statistically independent even in case we have more occurrences of the same word in the same transcript text. For our purpose, it is correct, because identification and classification work on the context analysis and not within a vocabulary. We can see that the precision obtained is the similar for each dataset. It means that the four annotators worked in same way.

The precision of the categorization for each dataset is between about 70% and 91%, which is more than double compared to a random classification process characterized by uniform probability distribution.

	Precision	Category Precision	Total Precision
Dataset 1			
LOC	91.54%	98.36%	89.90%
PER	77.08%	98.58%	75.66%
Dataset 2			
LOC	91.07%	99.75%	90.82%
PER	71.58%	100%	71.58%
Dataset 3			
LOC	88.90%	98.89%	87.78%
PER	70.98%	98.85%	69.83%
Dataset 4			
LOC	91.44%	99.38%	90.82%
PER	73.10%	99.63%	72.73%

Table 2 Precision in retrieval of entities in video

5 Conclusion

Nowadays, advances in technology allow improving traditional teaching, opening up new learning opportunities for students. In this paper we proposed a complete system to extract and localize entities within videos and to enrich them for the teaching purposes.

The experimental results obtained are encouraging when we compare our work with the state of the art. In fact, we demonstrate the feasibility of the proposed system in its possible use within the project defined by the MIUR[13]. In particular, the identification of places is the category with the best recognition precision, which is about 90% for all considered datasets. Then, the proposed system can provide useful information for georeferencing applications, visualization on maps or analysis/sharing of information related to the territory. Furthermore, the extraction and recognition of entities can be used for the automatic extraction of metadata used by indexing and research services in which named entities have particular relevance, such as for example a library, in which the search by authors can be fundamental for comparisons between multiple authors.

In conclusion, we have shown the successful implementation of our system, which can be seen as a creative and innovative educational system with media approach, where the educator is also educated and his own knowledge takes shape in the act of educating. Our system is free and can be used as an aid to learning techniques adopted nowadays by the MIUR. Innovative teaching and learning models are required to replay the actual market request. Just think of this lock down period where only this type of innovation has and will be able to make the step of quality.

[13] The Ministry of Education, University and Research (in Italian: Ministero dell'Istruzione, dell'Università e della Ricerca or MIUR).

References

1. N. Raju, H. Anita, International Journal of Applied Engineering Research **12**(24), 14750 (2017)
2. S.K. Gaikwad, B.W. Gawali, P. Yannawar, International Journal of Computer Applications **10**(3), 16 (2010)
3. J. Nothman, N. Ringland, W. Radford, T. Murphy, J.R. Curran, Artificial Intelligence **194**, 151 (2013)
4. E.F. Tjong Kim Sang, F. De Meulder, in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* (2003), pp. 142–147. URL `https://www.aclweb.org/anthology/W03-0419`
5. D. Nadeau, S. Sekine, Linguisticae Investigationes **30**(1), 3 (2007). URL `http://www.ingentaconnect.com/content/jbp/li/2007/00000030/00000001/art00002`. Publisher: John Benjamins Publishing Company
6. A. Ritter, S. Clark, Mausam, O. Etzioni, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Edinburgh, Scotland, UK., 2011), pp. 1524–1534. URL `https://www.aclweb.org/anthology/D11-1141`
7. X. Liu, S. Zhang, F. Wei, M. Zhou, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, Portland, Oregon, USA, 2011), pp. 359–367. URL `https://www.aclweb.org/anthology/P11-1037`
8. J.J. Jung, Expert Systems with Applications **39**(9), 8066 (2012). DOI https://doi.org/10.1016/j.eswa.2012.01.136. URL `http://www.sciencedirect.com/science/article/pii/S0957417412001546`
9. B.M. Sundheim, in *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995* (1995). URL `https://www.aclweb.org/anthology/M95-1002`
10. S. Ullo, G.R. Sinha, Sensors **20**, 3113 (2020). DOI 10.3390/s20113113
11. S.L. Ullo, S.K. Khare, V. Bajaj, G.R. Sinha, IEEE Access **8**, 124055 (2020)
12. M. Ceccarelli, M. di Bisceglie, C. Galdi, G. Giangregorio, S.L. Ullo, in *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, vol. 5 (2008), vol. 5, pp. 220–223
13. R. Giuliano, G. Cardarilli, C. Cesarini, L.D. Nunzio, F. Fallucchi, R. Fazzolari, F. Mazzenga, M. Re, A. Vizzarri, Electronics. p. 1055 (2020)
14. K.J. Han, A. Chandrashekaran, J. Kim, I. Lane, arXiv preprint arXiv:1801.00059 (2017)
15. C.C. Chiu, T.N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R.J. Weiss, K. Rao, E. Gonina, et al., in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2018), pp. 4774–4778
16. B. Kotelly, *Art and Business of Speech Recognition: Creating the Noble Voice* (Addison-Wesley Longman Publishing Co., Inc., USA, 2003)
17. T. Wilhelm-Stein, R. Herms, M. Ritter, M. Eibl, in *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, ed. by E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, E. Toms (Springer International Publishing, Cham, 2014), pp. 110–115
18. C.G. Snoek, M. Worring, Foundations and trends in information retrieval **2**(4), 215 (2008)
19. W. Hu, N. Xie, L. Li, X. Zeng, S. Maybank, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **41**(6), 797 (2011)
20. B. Patel, B. Meshram, arXiv preprint arXiv:1205.1641 (2012)
21. R.C. Veltkamp, H. Burkhardt, H.P. Kriegel, *State-of-the-art in content-based image and video retrieval*, vol. 22 (Springer Science & Business Media, 2013)
22. R. Xu, C. Xiong, W. Chen, J.J. Corso, in *Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015)
23. A. Torabi, N. Tandon, L. Sigal, arXiv preprint arXiv:1609.08124 (2016)
24. Y. Liu, S. Albanie, A. Nagrani, A. Zisserman, arXiv preprint arXiv:1907.13487 (2019)
25. N.C. Mithun, J. Li, F. Metze, A.K. Roy-Chowdhury, in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval* (2018), pp. 19–27

26. A. Miech, D. Zhukov, J.B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic, in *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 2630–2640
27. A.L. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, B. Russell, in *Proceedings of the IEEE international conference on computer vision* (2017), pp. 5803–5812
28. N.C. Mithun, S. Paul, A.K. Roy-Chowdhury, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 11,592–11,601
29. A. Ritter, S. Clark, O. Etzioni, et al., in *Proceedings of the conference on empirical methods in natural language processing* (Association for Computational Linguistics, 2011), pp. 1524–1534
30. M. Speranza, in *EVALITA 2009* (2009)
31. P. Basile, A. Caputo, A.L. Gentile, G. Rizzo, in *of the Final Workshop 7 December 2016, Naples* (2016), p. 40
32. R.R. Larson, Geographic information systems and libraries: patrons, maps, and spatial information [papers presented at the 1995 Clinic on Library Applications of Data Processing, April 10-12, 1995] (1996)
33. R.S. Purves, P. Clough, C.B. Jones, M.H. Hall, V. Murdock, Foundations and Trends in Information Retrieval **12**(2-3), 164 (2018)
34. Y.T. Zheng, Z.J. Zha, T.S. Chua, Multimedia Tools and Applications **51**(1), 77 (2011)
35. N. Golubovic, C. Krintz, R. Wolski, S. Lafia, T. Hervey, W. Kuhn, in *Proceedings of the 10th Workshop on Geographic Information Retrieval* (2016), pp. 1–2
36. E. Purificato, A.M. Rinaldi, Multimedia Tools and Applications **77**(20), 27447 (2018)
37. A. Messina, R. Borgotallo, G. Dimino, D.A. Gnota, L. Boch, in *2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services* (2008), pp. 219–222