RAMBLE ON: Tracing Movements of Popular Historical Figures

Stefano Menini[1-2], **Rachele Sprugnoli**[1-2], **Giovanni Moretti**[1],
Enrico Bignotti[2], **Sara Tonelli**[1] **and Bruno Lepri**[1]

[1]Fondazione Bruno Kessler / Via Sommarive 18, Trento, Italy
[2]University of Trento / Via Sommarive 9, Trento, Italy
{menini,sprugnoli,moretti,satonelli,lepri}@fbk.eu
{enrico.bignotti}@unitn.it

Abstract

We present RAMBLE ON, an application integrating a pipeline for frame-based information extraction and an interface to track and display movement trajectories. The code of the extraction pipeline and a navigator are freely available; moreover we display in a demonstrator the outcome of a case study carried out on trajectories of notable persons of the XX Century.

1 Introduction

At a time when there were no social media, emails and mobile phones, interactions were strongly shaped by movements across cities and countries. In particular, the movements of eminent figures of the past were the engine of important changes in different domains such as politics, science, and the arts. Therefore, tracing these movements means providing important data for the analysis of culture and society, fostering so-called cultural analytics (Piper, 2016).

This paper presents RAMBLE ON, a novel application that embeds Natural Language Processing (NLP) modules to extract movements from unstructured texts and an interface to interactively explore motion trajectories. In our use case, we focus on biographies of famous historical figures from the first half of the XX Century extracted from the English Wikipedia. A web-based navigator[1] related to this use case is meant for scholars without a technical background, supporting them in discovering new cultural migration patterns with respect to different time periods, geographical areas and domains of occupation. We also release the script to generate trajectories and a stand-alone version of the RAMBLE ON navigator[2], where users can upload their own set of movements taken from Wikipedia biographies.

2 Related Work

The analysis of human mobility is an important topic in many research fields such as social sciences and history, where structured data taken from census records, parish registers, mobile phones etc. are employed to quantify travel flows and find recurring patterns of movements (Pooley and Turnbull, 2005; Gonzalez et al., 2008; Cattuto et al., 2010; Jurdak et al., 2015). Other studies on mobility rely on a great amount of manually extracted information (Murray, 2013) or on shallow extraction methods. For example Gergaud et al. (2016) detect movements in Wikipedia biographies assuming that cities linked in biography pages are locations where the subject lived or spent some time.

However we believe that, even if NLP contribution has been quite neglected in cultural analytics studies, language technologies can greatly support this kind of research. For this reason, in RAMBLE ON we combine state-of-the-art Information Extraction and semantic processing tools and display the extracted information through an advanced interactive interface. With respect to previous work, our application allows to extract a wide variety of movements going beyond the birth-to-death migration that is the focus of Schich et al. (2014) or the transfers to the concentration camps of deportees during Nazism as in Russo et al. (2015).

3 Information Extraction

In Figure 1, we show the general NLP workflow behind information extraction in *RAMBLE ON*. The goal is to obtain, starting from an unstructured

[1]Available at http://dhlab.fbk.eu/rambleon/

[2]Both can be downloaded at http://dh.fbk.eu/technologies/rambleon

Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, April 3-7 2017, pages 77–80

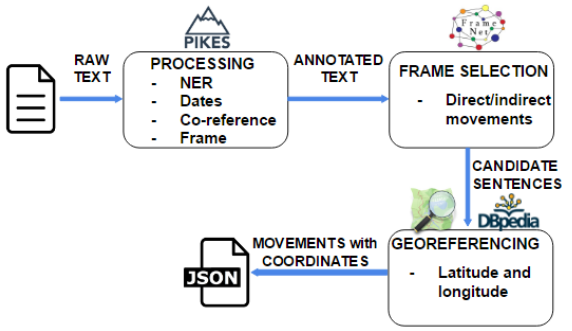

Figure 1: Information extraction workflow.

text, a set of destinations together with their coordinates and a date, each representative of the place where a person moved or lived at the given timepoint.

Input Data In our approach information extraction is performed on Wikipedia biographical pages. In the first step, these pages are cleaned up by removing infoboxes, tables and tags, keeping only the main body as raw text.

Pre-processing Raw text is processed using PIKES (Corcoglioniti et al., 2015), a suite of tools for extracting frame oriented knowledge from English texts. PIKES integrates Semafor (Das et al., 2014), a system for Semantic Role Labeling based on FrameNet (Baker et al., 1998), whose output is used to identify predicates related to movements and their arguments because its high-level organization in semantic frames is an useful way to generalize over predicates. PIKES also includes Stanford CoreNLP (Manning et al., 2014). Its modules for Named Entity Recognition and Classification (NERC), coreference resolution and recognition of time expressions are used to detect for each text: (i) mentions related to the person who is the subject of the biography; (ii) locations and organizations that can be movement destinations; (iii) dates.

Frame Selection Starting from the frames related to the *Motion* frames in FrameNet and a manual analysis of a set of biographies, we identified 45 candidate frames related to direct (e.g. *Departing*) or indirect (e.g. *Residence*) movements of people. After a manual evaluation of these 45 frames on a set of biographies annotated with PIKES, we removed 16 of them from the list of candidate frames because of the high number of false positives. These include for example *Escaping*, *Getting_underway* and *Touring*. Combining the information from the CoreNLP modules in PIKES with the remaining 29 frames listed in Table 1, our application extracts a list of candidate sentences, containing a date and a movement of the subject together with a destination. These represent the geographical position of a person at a certain time.

FRAME
Arriving
Attending
Becoming_a_member
Being_employed
Bringing
Cause_motion
Colonization
Come_together
Conquering
Cotheme
Departing
Detaining
Education_teaching
Fleeing
Inhibit_movement

FRAME
Meet_with
Motion
Receiving
Residence
Scrutiny
Self_motion
Sending
Speak_on_topic
State_continue
Temporary_stay
Transfer
Travel
Used_up
Working_on

Table 1: List of frames selected for RAMBLE ON[4]

Georeferencing To georeference all the destinations mentioned in the candidate sentences RAMBLE ON uses Nominatim[5]. Due to errors by the NERC module (e.g., *Artaman League* annotated as geographical entity), some destinations can lack coordinates and thus are discarded. Moreover, for each biography, the places and dates of birth and death of the subject are added as taken from DBpedia.

Output Data Details about the movements extracted from each Wikipedia biography, e.g. date, coordinates and the original snippet of text, are saved in a JSON file as shown in the example below. This output format accepts additional fields with information about the subject of the biography, e.g. gender, occupation domain, that could be extracted from other resources.

```
"name": "Maria_Montessori",
"movements": [{
 "date": 19151100,
 "place": "Italy",
 "latitude": 41.87194,
 "longitude": 12.5673,
 "predicate_id": "t3147",
 "predicate": "returned",
 "resource": "FrameNet",
 "resource_frame": "Arriving",
 "place_frame": "@Goal",
 "snippet": "Montessori's father died in
 November 1915, and she returned to Italy."
}]
```

[5] https://nominatim.openstreetmap.org/

4 Ramble On Navigator

Movements as extracted with the procedure described in Section 3 are graphically presented on an interactive map that visualizes trajectories between places. The interface, called `RAMBLE ON Navigator`, is built using technology based on web standards (HTML5, CSS3, Javascript) and open source libraries for data visualization and geographical representation, i.e. d3.js and Leaflet. Through this interface, see Figure 2, it is possible to filter the movements on the basis of the time span or to search for a specific individual. Moreover, if information about nationality and domain of occupation is provided in the JSON files, the Navigator allows to further filter the search. Hovering the mouse on a trajectory, the snippet of text from which it was automatically extracted appears on the bottom left. Information about all the movements related to a place is displayed when hovering on a spot on the map. The trajectories have an arrow indicating the route destination and are dashed if the movement described by the snippet is started before the selected time span.
The online version of the `Navigator` shows the output of the case study presented in Section 5, while the stand-alone application also allows to upload another set of data.

5 Case Study

We relied on the Pantheon dataset (Yu et al., 2016) to identify a list of notable figures to be used in our case study. We chose Pantheon since it provides a ready-to-use set of people already classified into categories based on their domain occupation (e.g., *Arts*, *Sports*), birth year, nationality and gender. More specifically, we considered 2,407 individuals from Europe and North America living between 1900 and 1955. First we downloaded the corresponding Wikipedia pages, as published in April 2016, collecting a corpus of more than 7,5 million words. Then we used the workflow described in Section 3 and we enriched output data with the categories taken from Pantheon. We manually refined the output by removing the sentences wrongly identified as movements (14.02%), for example those not referring to the subject of the biography (e.g., *When communist North Korea invaded South Korea in 1950, he sent in U.S. troops*). The final dataset resulted in 2,929 sentences from 1,283 biographies, since 1,124 individuals had no associated movements. This may be due to either an actual lack of sentences concerning movements or errors in the automatic processing, e.g., missed identification of places or dates. Table 2 shows the distribution per domain of the individuals with associated movements. Moreover, we corrected the coordinates of places wrongly georeferenced (6.7%). The extracted movements are evoked by predicates associated to 66 different lemmas (lexical units in FrameNet). The most frequent lemmas ($>$ 100 occurrences) are: *return* (567), *move* (556), *visit* (253), *travel* (188), *attend* (182), *go* (153), *live* (111), *arrive* (107).

DOMAIN	# of individuals	# of movements
Arts	647 (348)	788
Science & Technology	591 (318)	631
Humanities	502 (276)	709
Institutions	483 (255)	633
Public Figure	69 (30)	55
Sports	59 (30)	54
Business & Law	38 (13)	22
Exploration	18 (13)	37
TOTAL	2,407 (1,283)	2,929

Table 2: Domain distribution of individuals in our use case. In brackets the number of individuals with movements.

6 Conclusion and Future Work

We presented an automatic approach for the extraction and visualisation of motion trajectories, which is easy to extend to different datasets, and that can provide insights for studies in many fields, e.g., history and sociology.

In the future, we will mainly focus on improving the system coverage. Currently, missing trajectories are mainly due to (i) the presence of predicates not recognized as lexical units in FrameNet, e.g. *exile*; (ii) the lack of information in the English Wikipedia biography, and (iii) the presence of sentences with complex temporal structures, e.g., *Cummings returned to Paris in 1921 and remained there for two years before returning to New York*. These issues can be dealt with by adding missing predicates to FrameNet, extend Pikes to other languages and experimenting with different systems for temporal information processing (Llorens et al., 2010). We also plan to apply the methodology presented in (Aprosio and Tonelli, 2015) to automatically recognize the Wikipedia text passages dealing with biographical information, so to discard sections containing useless information.

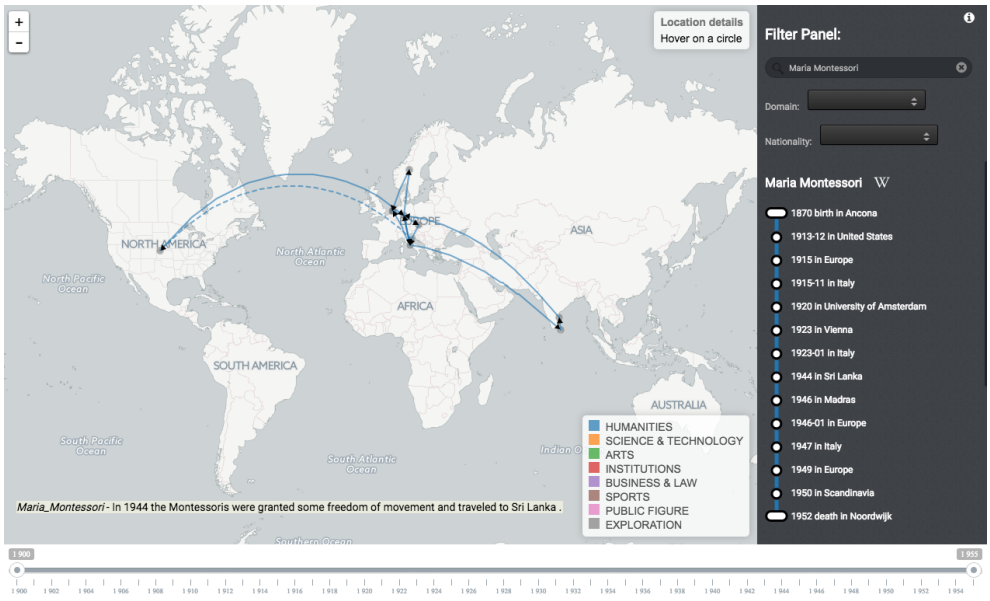

Figure 2: Screenshot of Ramble On Navigator.

References

Alessio Palmero Aprosio and Sara Tonelli. 2015. Recognizing Biographical Sections in Wikipedia. In *Proceedings of EMNLP 2015*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL '98*, pages 86–90.

Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. 2010. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PloS one*, 5(7):e11596.

Francesco Corcoglioniti, Marco Rospocher, and Alessio Palmero Aprosio. 2015. Extracting Knowledge from Text with PIKES. In *Proceedings of the International Semantic Web Conference (ISWC)*.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.

Olivier Gergaud, Morgane Laouénan, Etienne Wasmer, et al. 2016. A brief history of human time: Exploring a database of'notable people'. Technical report, Sciences Po Departement of Economics.

Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. 2015. Understanding human mobility from Twitter. *PloS one*, 10(7):e0131469.

Hector Llorens, Estela Saquete, and Borja Navarro. 2010. Tipsem (English and Spanish): Evaluating CRFs and Semantic Roles in Tempeval-2. In *Proceedings of the 5th SemEval*, pages 284–291.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Sarah Murray. 2013. Spatial Analysis and Humanities Data: A Case Study from the Grand Tour Travelers Project. In *A CESTA Anthology*. Stanford.

Andrew Piper. 2016. There will be numbers. *CA: Journal of Cultural Analytics*, 1(1).

Colin Pooley and Jean Turnbull. 2005. *Migration and mobility in Britain since the eighteenth century*. Routledge.

Irene Russo, Tommaso Caselli, and Monica Monachini. 2015. Extracting and Visualising Biographical Events from Wikipedia. In *Proceedings of the 1st Conference on Biographical Data in a Digital World*.

Maximilian Schich, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky, Mauro Martino, Albert-László Barabási, and Dirk Helbing. 2014. A network framework of cultural history. *Science*, 345(6196):558–562.

Amy Zhao Yu, Shahar Ronen, Kevin Hu, Tiffany Lu, and César A. Hidalgo. 2016. Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific data*, 3.