Extracting and Visualising Biographical Events from Wikipedia

Irene Russo*,Tommaso Caselli, Monica Monachini***

*ILC-CNR “A. Zampolli” Pisa,** Computational Lexicology & Terminology Lab Vrije Universiteit Amsterdam
irene.russo@ilc.cnr.it, t.caselli@vu.nl, monica.monachini@ilc.cnr.it

Abstract

This work presents a proposal for the development of a natural language processing module for event and temporal analysis of biographies as available in Wikipedia. At the current level of development, we restricted the extraction to temporally anchored events as they represent salient information which can be further used to extract additional events and facilitate their chronological ordering and the representation of a person’s timeline. Visualising data about basic facts concerning groups of people helps with historical reasoning and enables comparisons among them.

Keywords: mining biographies for structured information, visualising biographical data, temporal information

1. Introduction

Historical reasoning concerns facts and stories of the past to describe, compare and explain historical phenomena. Six activities can be listed as part of historical reasoning: a.) historical questions; b.) the use of primary and secondary sources; c.) contextualisation; d.) argumentation; e.) the use of substantive concepts; and, f.) the use of meta-concepts (van Drie and van Boxtel, 2008).

The way information is presented has an impact on historical reasoning: primary and secondary sources are today widely digitized and computational processing of textual and multimodal information (Novak et al., 2014) can meet the needs of users from the humanities, both in research and education. As a matter of fact, in many disciplines, visualisation constitutes a graphical or cognitive aid to thinking because it is based on interrelated textual content, spatio-temporal data and metadata related to images and videos. Data can be mainly visualised for presentation or exploration but in well designed projects there is a continuum between these two modalities (Cairo, 2012). A key aspect in Digital Humanities (DH) is to provide processed results in a way that is usable and easy to browse: data visualisation is today a device which enables the exploration, filtering and searching of data, skipping the interaction with databases.

Narratives can be integrated in visual forms of presentations: the Spatial History Project at Stanford University[1] deals with visualisations that involve the geographic dimensions of Holocaust, portraying, for instance, the mobility in the Budapest ghetto or the arrests of the Italian Jews.

In e-history projects, textual content is processed by Natural Language Processing (NLP) modules which take care of tasks such as Named Entity Recognition and Disambiguation (NERD). These modules identify different types of entities (e.g. Person, Organization, Location etc.) and can link them to external knowledge repositories (e.g. DBpedia) by means of URIs, thus enriching the extracted information. Hypotheses formulation about relations between entities (e.g. people, places and events) is supported once entities and relations between them have been mined from texts. Along this line, data can be explored in novel ways through basic inference rules that find out commonalities between two or more entities.

Our work focuses on comparisons and questions that may arise through visualisations of data automatically extracted from corpora. We designed a simple visualisation tool because visualising data about basic facts concerning groups of people may help with historical reasoning and enables comparisons.

In our framework a fact is a notion where history and data visualisation meet: if for historians a fact is an assertion about people and events that can be located in the past (i.e. a main predicate temporally and spatially grounded as in 1a), for data visualisation, based on data structures like JSON, the same fact, like the date of death, is an instantiated value for a key, as in 1b:

1a. Primo Levi died on 11 April 1987.
1b. “nodes”:[
    {“name”:“Primo Levi”,“deathDate”:“1987-04-11”}
]

As case study, we selected a coherent group of biographies concerning people that have been deported to concentration camps during Nazism, including both those who died because of the deportation and the survivors. Data cleaning and tagging is reported in Section 2.1 while in Section 2.2 we describe the way biographical information has been structured before automatic extraction.

Choosing an operational notion of biographical event as an activity that involve directly the person in question and can be anchored on a timeline whenever a date is available, we set up a basic biographical data array for each subject to encode factual data presented in the visualisation (see Section 3). We end with conclusions and proposals for future work in Section 4.

2. Dataset and Tools

2.1. Holocaust deportees dataset

As a case study 782 Wikipedia pages relative to the biographies of people deported in Nazi concentration camps were

[1] http://web.stanford.edu/group/spatialhistory

downloaded. The set includes 247 short biographies of people that survived. All these people have a key event in common, namely that they have been deported to Nazi concentration camps, and several others along which their lives differ and can be compared.

The biographies are part of a Wikipedia category, namely people who died in Nazi concentration camps[2], and of a list of Holocaust survivors[3].

The Wikipedia pages have been downloaded and saved in plain text files by removing all HTLM tags. The downloaded data have been processed by two different NLP systems: the NewsReader pipeline (Agerri et al., 2014) and the Stanford CoreNLP (Manning et al., 2014).

The NewsReader pipeline is a set of 15 NLP modules that generates a structure in the NLP Annotation Format (NAF) (Fokkens et al., 2014). The pipeline has been developed as part of the EU project NewsReader[4]. Apart from the basic processing modules, such as tokenization, part-of-speech tagging and lemmatization, the additional modules which are relevant for our project include:

- coreference resolution (COREF layer);
- named entity recognition and disambiguation (NERC and NERD layers);
- semantic role labelling (SRL layer).

Stanford CoreNLP is a Java natural language analysis library that includes a part-of-speech (POS) tagger, a named entity recognizer (NER), a dependency parser and a coreference resolution system.

2.2. Structuring biographical events

Events extracted from DBpedia are a subpart of biographical events found in Wikipedia entries and other resources. The aim of this section is to show how the integration of information can be achieved by means of structured information derived from the output of existing NLP modules.

From the NewsReader pipeline output we extracted the disambiguated DBpedia URIs from the NERD layer. The DBpedia URI is used to query the corresponding DBpedia HTML page to collect the basic information which are summarized in Table 1.

Grounding each page on DBpedia URIs, we decided to combine the two sources, Wikipedia and DBpedia respectively, to structure the array that will be the starting point for the visualisations, choosing the keys (see Table 1) that are basic for representing the lives of Holocaust deportees. In Table 2 the data extracted for Marceli Handelsman, a Polish historian that died in the Mittelbau-Dora concentration camp, are reported. Most of them come from DBpedia (Lehmann et al., to appear). The data are obtained by processing the linguistic information contained in the first lines of Wikipedia entries, missing important information that is provided in the remaining text as, in this case, the birth place, the death place and the exact date of death, that we are able to include because we extracted them from the processed text. Moreover, DBpedia sets the date of death to the first day of the year when it finds just the year in the Wikipedia pages infobox; we overwrite this value when the NLP tools extract the exact date for the event (i.e. `deathDate` in this case is 1945-03-20 and not 1945-01-01). `deportationFromCity` is an example of a potentially uncertain value: if the place where the arrest/deportation took place is not mentioned we presume that it coincides with the one where the person was living at the moment.

name	name of the deportee
nationality	nationality of the deportee (current nationality for survivors)
birthDate	date of the birth of the deportee
birthPlace	place of birth of the deportee
deathDate	date of death of the deportee
deathPlace	place of death of the deportee
deathInConcentrationCamp	YES/NO it reports if the deportee died in a concentration camp
deportationDate	date of deportation
deportationCamp1	first camp of deportation
deportationCamp2	second camp of deportation, if available
locationCamp1	location of deportationCamp1
locationCamp2	location of deportationCamp2
deportationFromCity	the city where the deportee was living at deportationDate
deportationAge	age of the deportee at deportationDate
deathAge	age of the deportee at deathDate
wikiOccupation	Wikipedia category
wikiNationality	Wikipedia category
wikiCamp	Wikipedia category
gender	M/F

Table 1: Basic biographical events as keys.

`wikiOccupation`, `wikiNationality` and `wikiCamp` values are imported, when they are available, from the Wikipedia taxonomy. These values allow to group deportees according to three different modalities (their nationality, their job and the concentration camp where they have been deported). This information could look redundant with `deportationCamp` and nationality

[2] `http://en.wikipedia.org/wiki/Category:People_who_died_in_Nazi_concentration_camps_by_occupation`

[3] `http://en.wikipedia.org/wiki/Lis_of_Holocaust_survivors`

[4] `http://www.newsreaderproject.eu`

DATE	PREP	EVENT	NSUBJ	DOBJ	LOCATION
1912	in	write	Gokkes	composition	
1921	in	found	he	choir	Amsterdam
1923	in	marry	Gokkes	Winnik	

Table 4: Output of the post-processing rules over the Stanford dependency parser.

name	Marceli Handelsman
nationality	Polish
birthDate	1882-01-01
birthPlace	Warsaw
deathDate	1945-01-01 no 1945-03-20
deathPlace	Mittelbau-Dora
deathInConcentrationCamp	yes
deportationDate	1944
deportationCamp1	Gross-Rosen
deportationCamp2	Mittelbau-Dora
locationCamp1	Rogonica, Poland
locationCamp2	Nordhausen, Germany
deportationFromCity	Warsaw (induced)
lengthJourney1	282 km
lengthJourney2	681 km
deportationAge	62
deathAge	63
wikiOccupation	historian
wikiNationality	Polish
wikiCamp	Mittelbau-Dora

Table 2: Example of basic biographical events about Marceli Handelsman.

name	Helen Berman
nationality	Dutch-Israeli
birthDate	1936-04-06
birthPlace	Amsterdam, Netherlands
deathDate	living
deathPlace	living
deathInConcentrationCamp	no
deportationDate	
deportationCamp	
deportationFromCity	
lengthJourney	
deportationAge	
deathAge	living
wikiOccupation	Painter
wikiNationality	
wikiCamp	

Table 3: Example of basic biographical events about Helen Berman.

extracted from data processed by NLP modules but, since it was part of Wikipedia pages infobox compiled by users, has a higher degree of certainty and, in case of mismatch, should be the considered the right value. In Table 3 data relative to Helen Berman, one of the survivors, make evident a different type of information that is missing. She was a child when she was deported and her short biography on Wikipedia mainly refers to her artistic career as a painter, briefly mentioning the event of deportation without details about it. Some of the keys in the JSON structure will be without a value because information is missing both in DBpedia and Wikipedia and this could be a problem for a set of persons in the dataset; other sources of information should be found to integrate them.[5] Some basic biographical events, such as birthDate and deportationDate among others, are anchored because of their nature to time, i.e. they make sense only when a temporal expression fills their values. We consider event anchoring as the first step to discover commonalities between biographies of different people as it provides a way of comparing events with respect to a timeline. Time anchored events have been extracted through a rule-based module on top of the output of the Stanford CoreNLP. The rules take in input all dependency relations between a temporal expression, marked as DATE in the NER analysis of the Stanford CoreNLP, and a verb in the same sentence. We assume that the dependency relation between the temporal expression and the verb expresses a temporal relation of inclusion, i.e. anchors the event on a timeline. More complex temporal relations, such as before, after, begins, ends, overlap or simultaneous, can be expressed. In particular, in case a temporal expression is introduced by a preposition, a further set of rules carved to express the temporal meaning (or relation) associated to the preposition have been developed. This set of rules is based on the manually annotated data from the English TimeBank[6] (Pustejovsky et al., 2003). After the time anchoring of the verb has been established we extracted also dependency relations concerning the subject, direct object and, if available, locations. Subjects and objects are then mapped to the NewsReader output to solve entity disambiguation (NERD) and pronominal coreference. Table 4 reports the output structure of the post-processing rules for event anchoring extraction. The data obtained from the Stanford parser and their integration with the NWR output facilitate the comparison of biographies. Examples 2a and 2b show how by looking for the same event lemma (e.g. "immigrate") in the extracted data, we

[5] One possible source could be the Central Database of Shoah Victims Names (http://www.yadvashem.org/), an international endeavor initiated and led by Yad Vashem; an estimated 4.3 million murdered Jews have been commemorated and a database of Shoah Survivors will be released soon.

[6] http://www.timeml.org/site/timebank/timebank.html

can easily aggregate people by the type of event and still be able to tfind out differences (in this case, the fact that there are two emigration instances, one in 1978 in Israel and one in 1947 in the Unites States).

2a. Helen Berman: *In 1978, she emigrated to Israel.*
1978, in, emigrate, she, Israel.
2b. Ruth Klüger: *In 1947 she emigrated to the United States.*
1947, in, emigrate, she, States.

Anchoring in time is often paired with anchoring in space because the information about where the event took place is necessary for completeness, for this reason we integrated the data structure of event lemma with information about the closest syntactic constituent that is a LOCATION according to the NER module of the Stanford parser (Israel and United States, in the cases above).

3. Visualisation of data

The implementation of the visualisation is based on D3.js, a JavaScript library designed to display digital data in a dynamic graphical form. We propose two interrelated visualisation modalities:

- A force-directed graph (Figure 1) where each person is a node connected to other nodes when they share the same value. It allows the visualisation of clusters of persons based on the different values in Table 1, highlighting data that have been extracted from Wikipedia, DBpedia and biographies parsed with the Stanford CoreNLP and the NewsReader pipeline. In this way, the names of the persons can be visible moving a pointer over a node and the source, i.e. the corresponding Wikipedia page in this work, will open in a different window, when clicking on the node.
  This type of visualisation will allow the user to directly explore the source of information from which the data have been extracted and thus verify if the proposed clusters are relevant or if they are due to errors in data extraction.

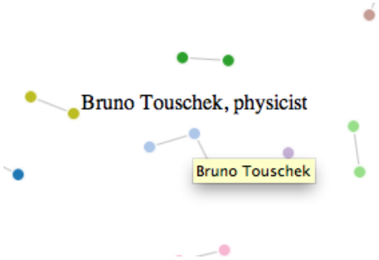

Figure 1: Force-directed graph visualisation.

- An horizontal timeline as illustrated in Figure 2. This visualisation allows to represent also relevant dates concerning larger events, e.g. World War II, which have crossed the lives of the people in our dataset. The timeline visualisation reports, for each deportee, a set of dates and sentences extracted from Wikipedia and describing a biographical event.

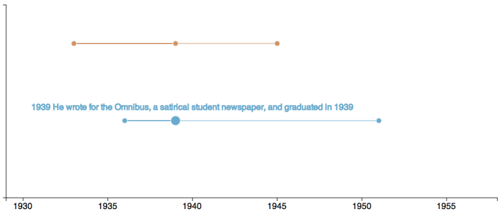

Figure 2: Horizontal timeline visualisation.

4. Conclusions and Future Work

We propose a method and a preliminary development of an NLP module for extracting biographical events from biographical notes such as the ones that are available in Wikipedia. Focusing on temporally anchored events will allow to extract salient events which can be further used to identify other biographical events and facilitate their chronological ordering and the representation of a person storyline.

One of the limitations of our visualisation as a tool for data presentation is the lack of potential interactions with historians: spatial and temporal data extracted from texts can be ambiguous or uncertain and some events can be wrongly extracted. Users should be able to validate the information discovered, labelling the results as true, worth further investigation or useless because noisy. We plan to make the visualisations interactive, with the possibility to delete or annotate each piece of information.

As future work we aim at labelling the biographical events as positive or negative, integrating a Sentiment Analysis component in our module by means of a psychologically grounded dataset (Lewinsohn and Amenson, 1978). So far this task was conducted manually by associating the predicate and semantic role information of the extracted events to the entries in the psychological dataset. The manual labelling aims at developing a reliable training set data for the development of a learning algorithm.

5. Acknowledgements

One of the author wants to thanks the NWO Spinoza Prize project Understanding Language by Machines (sub-track 3) for partially supporting this work.

6. References

Agerri, R. I. Aldabe, Z. Beloki, E. Laparra, M. De Lacalle, G. Rigau, A. Soroa, M. van Erp, P. Vossen, C. Girardi and S. Tonelli (2014), Event Detection, version 2 D4.2.2. NewsReader Project Deliverable.

Cairo, A. (2012), The Functional Art: An introduction to information graphics and visualization.

van Drie, J. and C. van Boxtel (2008), Historical Reasoning: Towards a Framework for Analyzing Students' Reasoning about the Past. Educational Psychology Review, 20 (2), 87-110.

Fokkens, A., A. Soroa, Z. Beloki, N. Ockeloen, G. Rigau, W. R. van Hage and P. Vossen (2014), NAF and GAF: linking linguistic annotations. In: Proceedings 10th joint iso acl sigsem workshop on interoperable semantic annotation, Reykjavik, Iceland, 2014, p. 9.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer and C. Bizer, (to appear), DBpedia A Large scale, Multilingual Knowledge Base Extracted from Wikipedia. To appear in the Semantic Web Journal.

Lewinsohn, J. and C.S. Amenson (1978), Some Relations between Pleasant and Unpleasant Events and Depression. In: Journal of Abnormal Psychology, 87(6): 644 654.

Manning, C., D., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard and D. McClosky (2014), The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

Novak, J., I. Micheel, L. Wieneke, M. Dring, M. Melenhorst, J. Garcia Moron, C. Pasini, M. Tagliasacchi and P. Fraternali (2014), HistoGraph, A visualization Tool for Collaborative Analysis of Networks from Historical Social Multimedia Collections. 2014 18th International Conference on Information Visualisation (IV).

Pustejovsky, J., P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro and M. Lazo (2003), The Timebank Corpus, Proceedings of Corpus Linguistics 2003.