UNIVERSITÀ DEGLI STUDI DI MILANO

Draft version February 11, 2023
Typeset using LaTeX default style in AASTeX631

"News from the past (P1), Text mining and sentiment analysis" project report, 2022-2023*

ELAHEH ESFANDI [1]

[1] *Master Student of Milan University, data science in economics (DSE), Matriculation number: 963762, elaheh.esfandi@studenti.unimi.it.*

ABSTRACT

Event extraction is intended to extract from the text a characterization of an event, defined by a set of entities associated with a specific role in the event. Historical events are specific types of events that are framed in a historical context, reporting facts, dates, historical figures, and locations The novelty of this detecting historical events task in this paper lies in the following endeavors. firstly, an operative definition of the notion of historical events. Secondly, with the methodology, we try Focusing on specific types of events, specific people events, topic-related events, or short/long-term events. Thirdly, an overview and statistics about the events extracted from a historical corpus of interest. And a case study of the so-called "Mozart" is the main purpose of this project.

Keywords: Named Entity Recognition, Event extraction, Mozart family letters, Text mining.

1. INTRODUCTION

Event detection is a process of analysis of text documents aiming to uncover real events happening in the world. the challenging task assumes that words appearing in similar documents and time windows are likely to concern the same real-world event.

By checking the literature, we could come across the idea of "Event extraction plays an important role in many applications in various fields". For instance, from the most related topic to our study, In the social media field, Ritter et al (2012), Zhou et al, Kunneman (2014) and Van Den Bosch (2016), and Peng et al. (2021) develop novel open-domain event extraction models to extract events from Twitter.

Sakaki et al. (2012) by using social media to provide important events for drivers, developed a system that extracts real-time driving information, such as traffic jams and weather reports. Tanev et al. (2008) performed real-time news event extraction for global crisis monitoring. In the legal field, Erwin Filtz et al (2020) by representing the main legal events, together with relevant temporal information, they did extract events from court decisions that can provide a visual overview of each case.

One of the challenges is that the definition of a historical event itself is problematic. Another challenge is since, when dealing with corpora of historical documents, the language itself may change dealing to shifts in semantics and style.

* "Text mining and sentiment analysis" project report, 2022-2023, Released on November, 12th, 2022

Historical events are specific types of events that are framed in a historical context, reporting facts, dates, historical figures, and locations.

The novelty of this detecting historical events task in this paper lies in the following endeavors. firstly, an operative definition of the notion of historical events. Secondly, with the methodology, we try Focusing on specific types of events, specific people events, topic-related events, or short/long-term events.

Thirdly, an overview and statistics about the events extracted from a historical corpus of interest. And a case study of the so-called "Mozart" which is the main purpose of this project. The rest of the paper is organized as follows. The next section deals with the Research question and analysis. In Section 3, I will go through the Experimental results and the most important key finding. Finally, Section 4 provides Concluding remarks on the user model, general framework, and the Results and findings of the model.

2. RESEARCH QUESTION AND METHODOLOGY

Event extraction is intended to extract from the text a characterization of an event, defined by a set of entities associated with a specific role in the event. As we got from the literature, there are almost three available techniques like data-driven approaches, knowledge-driven approaches, and hybrid approaches.

Data-driven methods require lots of data and a little domain knowledge and expertise, while having a low interpretation ability. This method doesn't deal with meaning explicitly but discovers relations in corpora without considering semantics. Conditional Random Field based (CRF) (Sarawagi Cohen, 2004) is a Data-driven method that applies the classifier to a set of texts to produce a set of annotated texts. The interest and efficiency of CRFs come from considering the dependencies between labels related to each other in the graph.

knowledge-driven approaches are often based on models that express rules representing expert knowledge. It is intrinsically based on linguistic and lexicographical knowledge, as well as on existing human knowledge concerning the content of the text to be treated. This alleviates the problems with the statistical methods concerning the meaning of the text.

GLAEE (Elloumi et al., 2013) is a knowledge-driven approach that is based on the generation of annotation patterns that involves a list of keywords and cue words. It purposes to identify events by an alignment between the pattern and the new text. the entities, the keywords, and the cue words require human intervention.

hybrid approaches seem to be somehow between data and knowledge-based approaches, requiring an average amount of data and domain knowledge and having medium interpretation ability. However, it should be noted that the amount of expertise required is high since several techniques are combined. Two-level approach (Elloumi, 2019) is one of the hybrid kinds of method which adapt the recognition of named entities level to the CRF tool based on learning techniques and correspondence between level 1 learning (PERSON, ORG, DATE, NUMEX, PROFIL) as well as learning level 2 (NEW PERSON, COMING PERSON), which brings us back to a double generation of the classifier.

In general, the event extraction task is a dependent domain. It requires human intervention to construct manually the annotation rules or to prepare an annotated corpus as input for the learning phase. To reduce the expert intervention, Sihem Sahnoun, et al, (2020) suggest using anthologies as a knowledge source for describing any event.

they suppose that an event ontology describes the relations between named entities and their possible roles in an event. To check whether a new text informs about a given event, in particular, the roles of some named entities, a matching process is required between the new text and the ontology.

they also suggested the following steps: First, apply the open information extraction (OIE) to extract the most relevant relations within a text, then apply the Named entity recognition (NER) on these relations. And finally, make a matching between the results of the previous step and the ontology for deriving a possible event.

in this paper, we are going through this methodology and we will try to find a relation between people and places in the data set that belong to the Mozart family letters.

3. EXPERIMENTAL STUDY

In this part of the paper, we will go through the overview of the data set used for experiments, the metrics used for evaluating performances, and the experimental methodology. Presents experimental results as plots and tables. The python codes are available in the appendix link of this paper for further information.

3.1. *a case study reconstructing the main events and overview of the data set*

The Mozart family letters are a fundamental source of information concerning daily life at the time and Mozart's own biography. Numerous details of his life - including details of the early tours and the composer's time in Vienna -

	lettera
0	Accademia Filarmonica Palazzo Sanvenanzi via G...
1	alla \r\n\r\n gennaio \r\n\r\nMi dispia...
2	alla \r\n\r\nA Madame Madame Marie Anne ...
3	alla \r\n\r\n dicembre \r\n Venerdì alle ...
4	alla \r\n\r\n Domenica notte probabilmente...
...	...
1555	Consigliere imperiale e vicepresidente del gov...
1556	Consigliere del governo tirolese Bibl Il Moza...
1557	Dal presidente del Tirolean Landesgubernium G...
1558	Nel Settecento quattro conti Attems furono can...
1559	Consigliere segreto dellimperatore

Figure 1. normalize the data set. Source: outcome of this paper

are known only from the letters. By the same token, they also give information concerning his compositional activities, including otherwise unknown works.

Even beyond illuminating the genesis, authenticity, and chronology of his music, however, the letters also give evidence concerning its performance, including questions of ornamentation, scoring, tempo, and the size of the orchestras he played with, in Salzburg and elsewhere.

This data set was gathered in the MySQL structured format with the data set of people, places, and music works in Mozart's family letters. the letters total 114, the places mentioned 173, locales within places 303, persons 452, Mozart's works 42 and other composers' works 50, generating a total of more than 14,500 citations Eisen (Cliff et al, 2011).

Note that, in their letters, the Mozart family sometimes wrote in secret code, presumably to hide material from the censors. These coded passages - realized in the French, English, and Italian translations - are given in square brackets; in the German originals, they are given both in code and realized.

after importing the table that we want into python from SQL, we tried to normalize and clean the texts of the letters. Based on this letter's texts, we had to clean text by removing punctuation, numbers, website, and emoji, at the end the database became the PANDAS data frame with 1560 rows that belong to 1560 letters. the result can be seen in figure(1).

3.2. *the metrics used for evaluating performance and methodology*

The approach proposed as an event extraction by using an OIE system for relation extraction, an automatic NER, and an ontology applied for any event (Sahnoun et al., 2020). The approach has two different phases as shown in Figure(2) that depend on one another.

A learning phase and a recognition phase. The learning phase consists of modeling an event by an ontology (classes, sub classes, and instances in relations), and constructing a set of rules manually.

The recognition phase operates through four steps: The input of the system is a text in natural language so the first stage is open information extraction. this means generating by an OIE system which allows to extraction of textual relationship triplets existing in each sentence. The aim of this step is to restrict the content of the text into relationships that are well-defined and have a specific meaning for each sentence in the text.

the second stage is named entity recognition. The input of the NER tool which is a triplet found in the previous step will have an automatic recognition of named entities after an organization step. now, the system is capable detect persons, organizations, and locations, in any part of the triplet. Among Python libraries, we chose spaCy for NER and NLTK for lemmatization and tokenization.

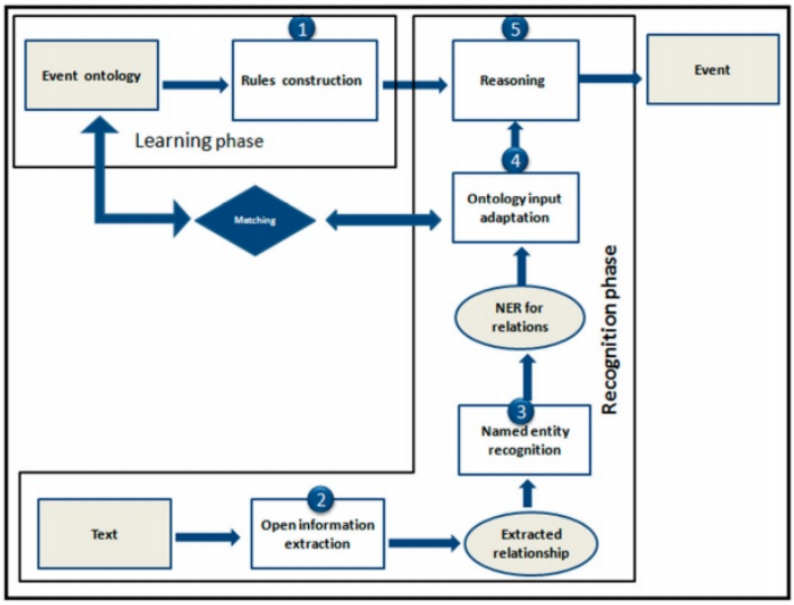

Figure 2. all phase of NER.

After recognizing the NEs, now is the ontology input adaptation stage. the verbs will be passed through a lemmatization layer that will convert conjugated verbs to their infinitive form. Every token recognized by a NE will be added as an instance in the ontology.

in this stage, Tokens are added as instances to the ontology and will be linked by relations whenever the number of named entities is greater than or equal to 2 to have a possible relationship among them, The lemmatized verb and the other relations between delimiters';' should be included in the relation list of the ontology and NEs can be linked with these relations.(Sahnoun et al., 2020).

the reasoning stage is the ending part of the algorithm, which contain the logical consequences from a set of rules to affect each instance of its event.

3.3. *experimental results*

by following the introduced stages in the previous part about the algorithm technique that we are using, the first stage after interring and normalizing the Mozart letter is to do the information extraction from all the available letters in the database. so the beginning is to tokenization all the sentences by the OIE system.

after we had done the tokenized as you could see the outcome in figure(3). we dropped "NA" values, and join them a gin to Extract the locations, time, and name from the text by using the "spacy" library.

In many languages, words appear in several inflected forms. "Lemmatization" is a text normalization technique in natural language processing. "Lemmatization" uses vocabulary and morphological analysis to remove affixes of words. our text is all in Italian sentences so for doing the sentiment analysis we need to "Lemmatizing" in Italian languages. so in the second stage, we used the package of"nltk" for doing the "Lemmatizing Italian sentences". .

in the third stage, for Extracting locations, time, and names from the text we used "Named Entity Recognition". "Named Entity Recognition" is the process of NLP that deals with identifying and classifying named entities. NER systems are developed with various linguistic approaches, as well as statistical and machine learning methods. NER has many applications for project or business purposes.

The raw and structured text is taken and named entities are classified into persons, organizations, places, money, time, etc. Basically, named entities are identified and segmented into various predefined classes. mainly shown by clearing the text differently. but for having clear and well design outcomes, we transferred the text coloring to the table. as you can see in a table (1), for the purpose of this study the location and a person had been chosen to show.

we also used the "displacy.render" to display a visualize named entities. so we passed the doc object with style as 'ent'. This code will highlight various types of entities in different colors. as you could see in figure(4), which is only a small part of the database, the location called in orange with the name "LOC" is close to the highlighted words. the

```
0       [Accademia, Filarmonica, Palazzo, Sanvenanzi, ...
1       [alla, gennaio, Mi, dispiace, molto, di, non, ...
2       [alla, A, Madame, Madame, Marie, Anne, Mozart,...
3       [alla, dicembre, Venerdì, alle, di, Siamo, qui...
4       [alla, Domenica, notte, probabilmente, dic, ho...
                              ...
1555    [Consigliere, imperiale, e, vicepresidente, de...
1556    [Consigliere, del, governo, tirolese, Bibl, II...
1557    [Dal, presidente, del, Tirolean, Landesguberni...
1558    [Nel, Settecento, quattro, conti, Attems, furo...
1559               [Consigliere, segreto, dellimperatore]
Name: tokenized sents, Length: 1560, dtype: object
```

Figure 3. Lemmatizing Italian sentences. Source: outcome of this paper

Table 1. selecting rows based on condition

number	text	label
0	Filarmonica Palazzo Sanvenanzi	LOC
1	Guerrazzi Fondata	LOC
2	Vincenzo Maria Carrati	PER
3	papa Benedetto XIV	PER
5	Santa Cecilia	PER
...	...	...
14130	Attems	LOC
14132	Attems	LOC
14133	Salisburgo	LOC
14134	Salisburgo	LOC
14135	Bibl V pConsigliere	LOC

7596 rows × 2 column, Source: the outcome of this paper

Figure 4. display the text in colors

light purple was chosen to show the person. the light blue was chosen to show the organization(ORG) like "Accademia Romana" or"Accademia Filarmonica".

.

4. CONCLUDING REMARKS

Information Extraction (IE) is the most popular area in artificial intelligence and had been born and developed between the late 1980s till 1990s. The core task of IE is named entity recognition (NER) which is an emerging field of research in NLP and information retrieval.

NER is based on the extraction of categorized textual objects in classes such as names of persons, location names, organization names, etc. Named entity recognition acts as a vital pre-processor tool in several NLP applications, namely machine translation systems, question-answering systems (KorkontzelosI. et al, 2015), text summarizing systems(SahaS.K. et al,2012), etc.

There are several NER systems that exist such as Systems based on symbolic approaches, systems based on learning approaches, and hybrid systems. The symbolic approach is based on the use of formal grammar built by hand and exploits syntactic labeling associated with words, such as the grammatical category of the word, which is based on dictionaries of proper names.

learning-based methods use annotated data that correspond to documents in which the entities, with their types, are indicated. Subsequently, a learning algorithm will automatically develop acknowledge base using several numerical models such as Conditional Random Fields(CRF), Support Vector Machines (SVM), and Hidden Markov Model (HMM). The combination of these two approaches is a hybrid approach that uses rules written manually and builds its rules based on syntactic information and information extracted from learning data.

in this paper, we used a learning-based method, verbs passed through a lemmatization layer to convert verbs to their infinitive form. An instance is every token recognized by a NE, then it was added to the ontology and linked by relations whenever The number of named entities was greater than or equal to 2 to have a possible relationship among them.

by Relation extraction (RE) we mean the identifying relationships between named entities in each sentence of a given document. A relation usually indicates a well-defined relation between two or more NEs.

we used theses methodology on The Mozart family letters data set. the clear and the brief outcome were shown in the previous part of this paper and also they are available in the codes. but as an example, we could say "Filarmonica Palazzo Sanvenanzi" and "Guerrazzi Fondata" is the whole address of the places that are related to the "Vincenzo Maria Carrati". we have to know that the two words "via" and palazzo" are the keywords in addressing the place in the Italian language. this can be seen in figure(4) and table(1).

5. REFERENCES

1. Alan Ritter, Oren Etzioni, and Sam Clark. Open domain event extraction from twitter. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1104–1112, 2012.

2. Deyu Zhou, Liang-Yu Chen, and Yulan He. A simple bayesian modeling approach to event extraction from Twitter. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 700–705, 2014.

3. Florian Kunneman and Antal Van Den Bosch. Open-domain extraction of future events from Twitter. Natural Language Engineering, 22(5):655–686, 2016.

4. Hao Peng, Jianxin Li, Yangqiu Song, Renyu Yang, Rajiv Ranjan, Philip S. Yu, and Lifang He. Streaming social event detection and evolution discovery in heterogeneous information networks. ACM Transactions on Knowledge Discovery from Data (TKDD), 15(5):1–33, 2021.

5. Erwin Filtz, María Navas-Loro, Cristiana Santos, Axel Polleres, and Sabrina Kirrane. Events matter: Extraction of events from court decisions. In Legal Knowledge and Information Systems, Legal Knowledge and Information Systems pages 33–42. IOS Press, 2020.

6. Takeshi Sakaki, Yutaka Matsuo, Tadashi Yanagihara, Naiwala P. Chandrasiri, and Kazunari Nawa. Real-time event extraction for driving information from social sensors. In 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), pages 221–226. IEEE, 2012.

7. Hristo Tanev, Jakub Piskorski, and Martin Atkinson. Real-time news event extraction for global crisis monitoring. In International Conference on Application of Natural Language to Information Systems, pages 207–218. Springer, 2008.

8. Sarawagi, S., Cohen, W. (2004). Semi-Markov conditional random fields for information extraction. NIPS'04: Proceedings of the 17th International Conference on Neural Information Processing Systems, Vancouver, USA (pp. 1185–1192).

9. Elloumi, S. (2019). An adaptive model for sequential labeling systems. Multimedia Tools and Applications, 78, 22183–22197. https://doi.org/10.1007/s11042-019-7558-8, ISSN:1573-7721

10. Elloumi, S., Jaoua, A., Ferjani, F., Semmar, N., Besancon, R., Al-Jaam, J., Hammami, H. (2013). General learning approach for event extraction: Case of management change event. Journal of Information Science, 39(2), 211–224. https://doi.org/10.1177/0165551512464140

11. Eisen, Cliff et al. In Mozart's Words, 'Introduction' ¡http://letters.mozartways.com¿. Version 1.0, published by HRI Online, 2011. ISBN 9780955787676.

12. KaurA. et al. Evaluation of named entity features for Punjabi language Procedia Comput. Sci. (2015)

13. KorkontzelosI. et al. Boosting drug named entity recognition using an aggregate classifier Artif. Intell. Med. (2015)

14. SahaS.K. et al. A comparative study on feature reduction approaches in Hindi and Bengali named entity recognition Knowledge Based systems.

15. Sihem Sahnoun, Samir Elloumi, and Sadok Ben Yahia. 2020. Event detection based on open information extraction and ontology. Journal of Information and Telecommunication, pages 383–403.

16. Eisen, Cliff et al. In Mozart's Words, 'Introduction' ¡http://letters.mozartways.com¿. Version 1.0, published by HRI Online, 2011. ISBN 9780955787676.

APPENDIX

A. DECLARATION BY AUTHOR

"I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study."

B. ALL THE AVAILABLE LINKS FOR CODES BY AUTHOR

GitHub:https: https://github.com/elahehesfandi/News-from-the-past-
or this is the repository name: Github:elahehesfandi/News-from-the-past
Google-colab :.....