Police Report Linking Algorithm Based on Named Entity Recognition

Mauro Daniel Alvarez[0009-0005-5288-9651][1] , Leandro Antonelli[0000-0003-1388-0337] [2 3]

[1] Facultad de Informática UNLP, La Plata, Bs As, Argentina
[2] LIFIA – Facultad de Informática, UNLP
[3] CAETI – Facultad de Tecnología Informática – Universidad Abierta Interamericana

{mauro.alvarez}@info.unlp.edu.ar
{leandro.antonelli}@lifia.info.unlp.edu.ar

Abstract. Police reports include a complaint made by the victim of a crime. This complaint is a story that describes the facts from the victim's point of view. Most of the time, crimes are perpetrated by unknown authors. This causes cases to be shelved until new evidence arrives. The use of natural language processing allows us to take advantage of non-structured text in victim's complaints by generating links that could lead to the reopening of an archived investigation. Through the use of NER[1], it is possible to extract entities of interest from a report of a complaint that arrives and link it with other reports of existing complaints, allowing the generation of a maps of links in order to detect similarities between cases. This idea will increase the possibility that cases in archived status being opened.

Keywords: NLP, NER, police reports, criminal justice, graphs, similarity

1 Introduction

Victim's complaints are composed of non-structured text. Some of them contain information that can be categorized such as names, places, timetables, etc.

With a knowledge base of complaints, it is possible to compare similarities between a report of an entry complaint and the existing ones.

Initially, it is necessary to reduce the number of reports in the existing complaint database, i.e., to allow only those reports of complaints that have some degree of similarity to the incoming complaint.

The aim of this paper is to propose a NER-based strategy for the identification of similar complaints. This strategy is made up of two phases.

The first phase consists of reducing the number of reports in the existing complaint database, i.e., reducing and preserving only those reports of complaints the have some degree of similarity to the incoming complaint.

[1] Named Entity Recognition

The second phase consists of the recognition of named entities or NERs. From this moment on, the aim is to compare those reports of complaints that have some relationship in categories such as names, people, locations, timetables, etc.

From the result of this comparison, the most similar entities are selected in terms of named entities and an appropriate visualization is generated for the user.

The following section describes the theoretical framework that was used for the development of the algorithm.

The contribution section details the two-phase algorithm and finally gives a brief conclusion about the algorithm.

2 Theoretical framework

Natural language processing (NLP) enables machines to understand and generate written text that can be interpreted by humans.

All natural language processing requires standardization [1], that is, the removal of words and symbols that do not add meaning to the text, such as punctuation symbols, white spaces and also the conversion of lowercase letters from the text, since, for machines, the capitalization of words is indistinct.

The process also encompasses the extraction of lemmas [2]. An interesting case is made up of verbs. With lemma extraction, we avoid having to deal with conjugated forms of verbs. The goal is to unify as much as possible the language in order to apply Jaccard similarity index more efficiently.

The unification of vocabulary is necessary due to the fact that this method does not perform any kind of semantic comparison, only syntactic, so that conjugated verbs and their root base or lemma do not belong to the intersection, they are simply considered two words that have no relationship at all.

Jaccard method is defined as the intersection divided by the union of the sets [3].

For example, with sets A and B, the Jaccard similarity index indicates that they are the numbers of common elements that A and B have, divided by the number of elements that the union of those sets have.

On the other hand, there is the recognition of named entities [4], which allow the extraction, as their name indicates, of entities.

The entities are classified in different categories, in order to allow grouping unstructured texts where these entities participate. For example, the texts could be grouped where certain proper nouns appear, or even use the geographical location to determine how many articles refer to that area. The above idea es reinforced by the crime pattern theory [5].

The aforementioned grouping would make it possible to detect patterns in order to identify people who commit crimes based on data such as type of crimes, locations, schedules, among others.

3 Contribution

Using as starting point the paper An Approach to Cluster Scenarios According to their Similarity using Natural Language Processing [6]. The following increment in the solution is proposed.

In this case, the concept of corpus will be used to describe a set of stories of complaints which are composed of unstructured text.

These complaints were previously subjected to a normalization process.

3.1 First phase of the strategy

When a new complaint arrives, it could be determined how similar it is with respect to the knowledge available in the model, that is, what degree of similarity corresponds to the reports available in the knowledge base. To determine this similarity, the Jaccard similarity index is used in order to yield those complaints that are most similar to the new document.

In order to reduce the number of documents, complaints with higher similarity index are selected and named entities are extracted from the new document and the similar ones, ending phase one of the algorithm. The selection criteria for documents is based on Jaccard similarity higher or equal to forty percent (40%), those below that value are consider negligible.

Be it the following complaints and a recently arrived complaint called a pivot, in Fig. 1. it can be seen the comparison made of the pivot against nine (9) documents. In this case it can be noted that the complaints differ greatly syntactically, with the first two complaints being selected.

Caso	Indice de similitud
Caso 120928	0.42857142857142855
Caso 120927	0.4166666666666667
Caso 122633	0.3333333333333333
Caso 122531	0.2222222222222222
Caso 122521	0.20689655172413793
Caso 122693	0.2
Caso 120924	0.19642857142857142
Caso 122709	0.16666666666666666
Caso 122355	0.09375

Fig. 1. Table - Jaccard Similarity Index

3.2 Second phase of the strategy

For the extraction of named entities, the SpaCy [7] library is used, which already has support for this task. This library generates automatic tagging of named entities, for instance, it can differentiate between ORG, LOC, PERSON, TIME, etc.

The idea is as follows, a model will be trained for the NER module of SpaCy. For that matter, the NER Annotator [8] tool will be used.

The NER module is part of the SpaCy pipeline. This module can be extracted to perform adjustments to the model. In this case, the model will be trained in the criminal justice domain, such as crimes, related legal articles and some other more general sections like street names and their belonging to a neighborhood.

This training gives the model the ability to link words like robbery with the penal code article that corresponds to that crime, or to recognize the neighborhood that belongs to a certain street.

Tentative training sets for the NER module are detailed below.

Table 1. Crimes training template[2]

Word	NER	Additional information
Robbery	Article 140 of penal code	Common theft
Burglary	Article 140 of penal code	Common theft
Vandalism	Article 201 of penal code	Damage of property
…	…	…
Word N	Article N of penal code	Description N.

Table 2. Neighborhood training template

Street	Number	NER
San Martín	0	Centro neighborhood
Muzio	10	Padre Juan neighborhood
Urquiza	20	Padre Juan neighborhood
Marconi and. San Martín	S/D	Centro neighborhood
…	…	…
Street N	Number N	Neighborhood N

Table 3. Suspects template

Full Name	Alias	NER
John Doe	Doey	John Doe + Doey SUSPECT

[2] They do not correspond to actual articles numbers, they are for example purposes only. Besides each country has it own penal code.

Ron J	Joey	Ron J + Joey SUSPECT
…	…	…
N	N	N SUSPECT

Once the new named entities have been obtained, the Jaccard method is re-executed with recognized entities, i.e., with the categories, in order to obtain a successive refinement and thus determine which ones are most similar to the story of the new complaint.

This process will yield the most similar cases just like the representation shown in Fig. 1. This information will then be processed by the visualization tool [9] to generate a graphed map composed of the crime, the neighborhood where the event occurred, those involved if they exist, otherwise they are categorized as unknown authors.

Fig. 2. depicts the resulting schema with graphs.

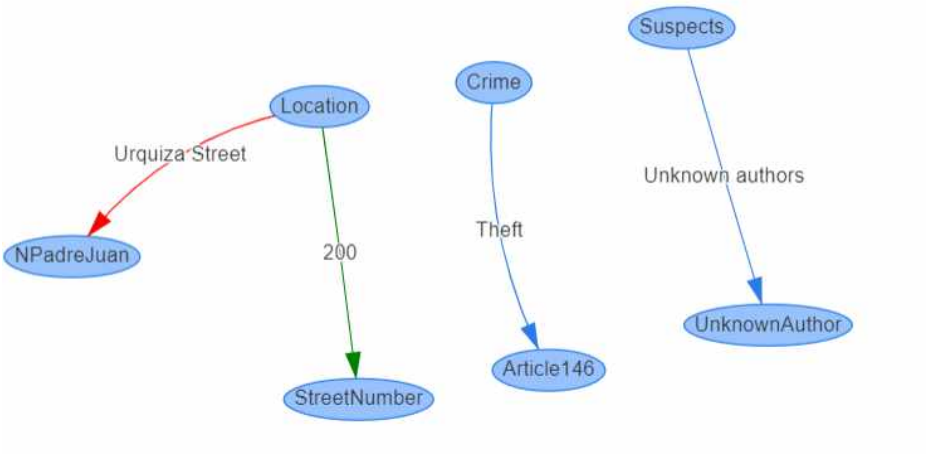

Fig. 2. Representation of a complaint story with the Vis tool

This idea could be extended to the N most similar cases to the pivot, in order to obtain a visualization of a set of N cases that can provide the investigator with clues, such as detection of crime patterns on the cases.

This would make it possible to expand the possibility of reopening cases currently shelved for further investigation.

3.3 Initial implementation

This algorithm is currently under development. We proceeded to with a proof of concept[3]. The idea is to determine feasibility through a minimum implementation and then proceed to a definitive implementation.

The PoC of phase one (1) of the algorithm is available, it was tested with a sample of nine (9) complaints and the pivot. The application of the PoC of phase one (1) reduced the complaints of interest to two (2) as can be seen in Fig. 1.

The next step is the development of the phase two (2) PoC. It consists of the training of a NER model. In this case, a simplified model of named entities related to justice system language will be defined.

Taking the results from the PoC of phase one (1), it will be fed into the PoC of phase two (2) to perform the extraction of named entities from both the complaints and the pivot complaint.

The output of this process consists of a JSON-formatted object. It will contain the links and details of the cases as well as the named entities. This information in this state does not represent an adequate form, and it is necessary to send this object with links to the Visjs library, in order to obtain a graphical representation of the object and its relations.

4 Algorithm validation

The algorithm contains two phases and the validation of both is independent, because phase one (1) consists of the implementation of the Jaccard similarity index. It will be composed of unit tests with the unitest library [10] of Python.

As for the NER model, being a machine learning model, validation is carried out by making changes to the hyperparameters and controlling the loss rate.

It is important to define which complaints will be in the training set, which ones in the test set and finally which will be in the validation set. Omitting this separation of sets will result in overfitting the model, and it would not perform properly to information the model has not previously seen.

The above validations are of a technical nature, and it is also necessary to validate manually. In this case, it is proposed to assemble a corpus with complaints of cases in solved state. The idea is to determine the output generated graphically in order to verify if it is consistent with what happened in the case history and if the related cases are linked appropriately.

To determine the degree of accuracy, it is necessary to involve domain stakeholders such as prosecutors, who can verify whether the maps generated would be useful in an investigation.

[3] Proof of Concept or PoC.

5 Related work

There are different articles that aim to provide investigators, police officers and other actors with tools to solve cases that are currently waiting for new clues. The article A distance Measure for Determining Similarity between Criminal Investigations [11] takes advantage of the volume of data to determine what crimes may have been committed by a set of individuals. It is a system composed of several modules, such as a text miner, transformers, etc., and after going through this pipeline, a visual report is generated for police officers.
Another interesting work that is related with phase two (2) of the algorithm presented in this paper, is the article entitled Extracting Meaningful Entities from Police Narrative Reports [12]. This paper describes an extractor of named entities to deal with textual police reports.

6 Future work

The PoC presented as such is not suitable for use in real environment, therefore, it is proposed to evolve the PoC into a final product.

In this case, a large part of work will consist of collecting data about cases, and their classification for the subsequent training of the NER model. In the author's work environment, there are large amounts of information about cases and bibliography on legal concepts and also the advice of professionals in the area such as prosecutors.

7 Conclusion

The algorithm described constitutes a starting point that provides support for the investigation of criminal offences. The inclusion of a graphical interface will complement this algorithm, making it intuitive for operators in the organization to use.

Considering the possibilities offered by natural language processing, the failure to automate the available information, in this case the reports of complaints, represent a waste of data that could open new paths in stagnant investigations.

As a note of interest, the implementation of this tool in the future will represent a milestone in the organization in which the main author of this paper currently works, since there is no tool based on artificial intelligence.

References

1. Aliero, A. & Bashir, S. & Aliyu, H. & Tafida, A. & Kangiwa, B. & Dankolo, N. (2023). Systematic Review on Text Normalization Techniques and its Approach to Non-Standard Words. International Journal of Computer Applications. 185. 975-8887.
2. Khyani, D. & Siddhartha B. S. (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing. Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology. 22. 350-357.

3. vor der Brück, T., Pouly, M.: Text similarity estimation based on word embeddings and matrix norms for targeted marketing. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies vol 1 pp. 1827-1836. (2019).
4. Salman, N. & Muhammad, G., & Sohaib, & Khalid Alvi, Sohaib & Kiran, Anam & Rehman, Shafique Ur & Murtaza, Ghulam & Campus, Jehlum & Jehlum, Pakistan. (2022). Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance.
5. Brantingham, P. & Brantingham, P. (2013). Crime pattern theory. Environmental Criminology and Crime Analysis. 78-93. 10.4324/9780203118214.
6. Delle Ville J., Torres D., Fernández A., Antonelli L. An Approach to Cluster Scenarios According to their Similarity using Natural Language Processing. Lifia, Fac. de Informática, UNLP, La Plata, Bs As, Argentina.
7. SpaCy. https://spacy.io/. Visited April 10, 2024.
8. NER Annotator. https://tecoholic.github.io/ner-annotator/. Visited April 8, 2024.
9. Vis JS. https://visjs.org/. Visited April 7, 2024.
10. Unittest. https://docs.python.org/3/library/unittest.html Visited April 18, 2024.
11. Cocx, T. & Kosters, W. (2006). A Distance Measure for Determining Similarity Between Criminal Investigations. 511-525. 10.1007/11790853_40.
12. Chau, M. & Xu, J. & Chen, H. (2002). Extracting Meaningful Entities from Police Narrative Reports.
13. Lane H., Dyshel M., Natural Language Processing in Action, Second edition, Version 12.
14. Python. https://python.org/. Visited April 10, 2024.