Event identification in the Monsoon Books (1616-1618)*

Ana Sofia Ribeiro[1 [0000−0002−1822−5908]], Anderson Sacramento[2 [0000−0002−2288−6899]], Marlo Souza[2 [0000−0002−5373−7271]], and Renata Vieira[1 [0000−0003−2449−5477]]

[1] CIDEHUS, Universidade de Évora, Portugal asvribeiro@uevora.ptEvent
[2] Universidade Federal da Bahia, Brasil

Abstract. The *Estado da Índia* constituted the most complex overseas Portuguese set of territories. This paper investigate a digital methodology employing Natural Language Processing to study historical events regarding it during the period 1616-1618. We explore the application of an event extraction tool over an extract of the *The Monsoon Books*. Our preliminary results expose the current problems and help us shape further work for automatic processing of historic corpora.

Keywords: Event identification · 17th century Portuguese · Portuguese India

1 Introduction

The *Estado da Índia* constituted the most complex overseas Portuguese set of territories, encompassing a geography as wide as the Indian Ocean borders [4]. At the same time, literature underlines that this political unit also should be perceived within the geographical scope of Portuguese informal presence in Asia, that is, the places where free-riding Portuguese individuals constituted significant communities (mostly *mestizos*), albeit there were nor Portuguese formal structures of any sort, nor under Portuguese jurisdictional power [5]. This paper intends to investigate a digital methodology employing Natural Language Processing to study this reality in a fully-integrated way during the period in which Portugal was part of the Hispanic Monarchy (1580-1640). Political affairs, socioeconomic realities, cultural and religious matters, relationships with autochthonous neighbouring powers and communities, as well as the relations with competing European powers are key variables to be addressed.

Particularly, we explore the application of an event extraction method aiming to support the process of data identification of historical junctures, which is a huge time-consuming task for historians dealing with massive document corpora.

 Supported by by Portuguese Foundation for Science and Technology (FCT) under the projects CEECIND/01997/2017, UIDB/00057/2020.

A process like this applied to the study of a colonial macro-region as the eastern Portuguese Empire identifies and categorize human actions and episodes which determine not only patterns of historical junctures in time and place, but also disruptive events that underline changing processes.

2 The Monsoon Books

The *Documentos Remetidos da Índia* or *Livros das Monções* (Monsoon books) collect letters exchanged between the monarchs and Portuguese government councils and India viceroys where all types of affairs concerning the so-called Portuguese *Estado da Índia* were discussed. They comprise a geographical scope from Eastern Africa to Japan. The use of this collection is paramount to understand the internal dynamics of the Portuguese Estado da Índia until the 19th century. In fact, they are considered the core documents produced by Portuguese authorities in Asia. The fact of being a type of documental corpora concerning all types of issues makes the Monsoon Books unique and a privileged lab for building a new analytic model and approach to understand internal dynamics of colonial empires macro-regions. The Monsoon Books are composed by the sets of documents located in both in the Portuguese National Archives, in Lisbon, and in the Historical Archives of Goa, in Panjin, India. Since this paper intends to test an automatic event extraction model in order to conceive an interpretative framework of European colonial presence in overseas macro-regions, we employ some of the already transcribed and published books referring to the years of 1616-1618 [3].

3 Event Identification and Classification

The goal of event identification and classification is to detect the event mentions of target event types in plain text. Given an input text, an event detection (ED) system should be able to identify whether the sentences contains events of interest by means of the identification of event trigger terms (event identification) and classify them into specific event types (event classification). For instance, in the following sentence:

> "Meridian National Corp. **said** it **sold** 750,000 shares of its common stock to the McAlpine family interests, for $1 million, or $1.35 a share."

According to the TimeBankPT corpus [2], the words "**said**" and "**sold**" describe event occurrences (triggers) for two distinct event mentions, one of type *Statement* and the other of type *Commerce Selling*, respectively, if we consider the FrameNet [1] lexicon as a source of target event types. Most of the work done on ED in the literature has focused on contemporary variants of European languages [8] and, even for the contemporary variants, few studies have addressed event identification and classification for the Portuguese language. Notably, the work of Sacramento and Souza [6] describes a method and the only,

to our knowledge, publicly available system for event extraction on Portuguese sentences. TEFE encodes ED as a sequence labelling problem and employs bidirectional recurrent neural networks to simultaneously predict event triggers and their types. It was trained on an enriched TimeBankPT corpus with events types from the FrameNet project, using deep neural networks and contextualized word embeddings from a Portuguese BERT model [7].

In this work, we employ TEFE to historical data, on the previously transcribed texts from the Monsoon Books, and evaluate its usefulness to the identification of historical junctures in historical corpora.

4 Applying Event Identification to the Monsoon Letters

Table 1. Events identified by TEFE in Examples 1 - 5

Trigger	Event Type	Source
"saber"	Awareness	Example 1
"entendido"	Awareness	Example 1
"partiram"	Departing	Example 1
"comaçaram"	Activity Start	Example 1
"fazer"	Intentionally act	Example 1
"causa"	Causation	Example 1
"receber"	Receiving	Example 1
"saber"	Awareness	Example 2
"tire"	Removing	Example 2
"sabendo"	Awareness	Example 2
"ver"	Perception experience	Example 3
"cumprirá"	Activity ongoing	Example 4
"fez"	Intentionally act	Example 4
"fez"	Causation	Example 4

Each volume of Monsoon books encompasses more than 300 printed pages of narrative text. In this sense, applying a computational tool of event extraction enhances historical research to rapidly extract text information for a large collection of data. One way of studying this source is by organizing the events it describes. Language technology may help the reader with hints, extraction and quantification of these events. The system described in the last section was developed with the purpose of finding and classifying mentions to events, as well as identifying the participants of the events. The system receives an input sentence such as "*Eu el-rey faço saber aos que este alvará virem que tenho entendido que pelo mau comcerto que tiveram as naus que, o anno passado de seiscentos e quinze, partiram do porto de Goa para este reino*[...]" and identify that "partiram" (departed) is an instance of a Departure event (described by the Departing Frame) and that "*as naus*" (the ships), "*porto de Goa*" (Goa's harbor) and "*este reino*" (this kingdom) are entities participating in such event, as depicted below:

$$\underbrace{\textit{as naus}}_{Theme}\ \text{que[...]}\quad \underbrace{\textbf{partiram}}_{trigger{:}Departing}\quad \text{do}\ \underbrace{\textit{porto de Goa}}_{Source}\ \text{para}\ \underbrace{\text{este reino}}_{Goal}$$

In this work we present an initial assessment of the application of this system in the source under study, to understand the needs for adapting the tool for language variants (from a different time span, in this case). Next we present some passages of the studied source, isolated in five examples, and what the system has produced as output, along with a discussion of the challenges to face.

> **Example 1**: *Eu el-rey faço saber aos que este alvará virem que tenho entendido que pelo mau comcerto que tiveram as naus que, o anno passado de seiscentos e quinze, partiram do porto de Goa para este reino, e por virem sobrecarregadas, em saindo dáquella barra começaram logo a fazer agoa, o que foi causa de se perderem as naus Capitania e Sam Boaventura, e a Sam Philippe vir mui arriscada, e por esse respeito receber minha fazenda e meus vassalos notável perda;*

The passage mentions a communication made about bad mantainance of ships and the consequent loss of value related to overweighted ships coming from Goa to Portugal which sunk. In this passage, as presented in Table 1, 7 events were identified, related to the communication brought about bad maintenance of the ships, the leave of these ships from Goa, the start of the problems caused by the bad maintenance, which caused the loss of values due to the problems encountered with 3 of the ships.

> **Example 2**: *e porque convem muito a meu serviço saber-se com a consideração necessaria de como se procedeo no concerto e carga das ditas naus, e se ouve culpa de alguem de partirem tarde, hei por bem e mando ao Desembargador Gonçalo Pinto da Fonseca tire devassa na conformidade d´este alvara, sabendo mui particularmente a causa que houve pera as ditas naus virem sobrecarregadas e tam mal concertadas, e partirem tarde, e como se procedeo no concerto e carga d´ellas como se refere;*

In continuation from the previous passage, here there is a demand for the causes of the poor maintenance of the ships and how this was done, asking also for the identification or the responsible related to the delay of their departure, the causes for the overweight. The results of processing this passage on TEFE is depicted on Table 1. Two events identified were related to the requirement of information, and the third (removing) was in fact used as asking rather than removing, an error which might be explained due to verb ambiguity.

> **Example 3**: *e depois de tirada a dita devassa a mande serrada por vias nas primeiras naus que pera este reino partirem dáquellas partes, dirigida ao Conselho de minha fazenda, pera se ver n´elle e prover no caso como mais convier a meu serviço;*

At this point, instructions are given on how the answer to the enquiry should be sent. The event "ver" (to see), depicted in Table 1, is identified but in this case the present stage of development of this tool is not flexible enough and comprises misunderstood meanings of the text, common in a Portuguese text with 400 years old.

> **Example 4**: *o que cumprirá sem duvida alguma, por este que valerá como carta e não passará pela chancellaria, o qual vai por tres vias. Francisco de abreu o fez em Lisboa a xbij (desasete) de fevereiro de seiscentos e desaseis. Diogo Soares o fez escrever.*

In this particular example, the tool enables us to automatically understand the material authorship of the particular letter, identifying an agent responsible for the act of writing the document in a certain space and time, as can be seen in Table 1. In this analysis, it is not relevant to make an analysis of word statistics, since the documental corpora was not entirely digitally intervened yet. By the results shown in Table 1 we immediately perceive how the diverse documents present in the analysed volume of Monsoon Books were used to report to the metropolis or to India about the remote events relevant to the Portuguese administration of *Estado da India* and the Cape Route in the beginning of the seventeenth century. The most common event type is awareness underlining the role of communication that these documents performed.

5 Final remarks

This paper analyses the application of event identification, extraction and classification as a step for the study of historical sources. This initial study was required to know the current problems and help us shape further work. As such, we are now planning to annotate events in a portion of this historical source. Annotation will serve for adapting the tool for this specific temporal language variant and the required historic study needs, and also for more rigorous evaluation of the extraction provided by the tool. Regarding the language differences, we plan to study whether normalization could improve the extraction of events. The adapted version of the tool will be used for further studies about this historical source. These tools may serve as reader´s guidance for the observation of historical sources contents. The tool enhances a more efficient process of information extraction of huge series of historical texts. From the examples shown, and regarding the nature of this source, we may consider uses such as helping researchers to find all events related to ships, actions of the Crown, conflicts, etc. If associated with other NLP tools it would be possible to create a list of the names of the ships, perhaps finding information about the travels made by them. It also serves the purpose to observe the geographical location of a person in a certain time and space and the type of actions they perform as historical characters. Even a statistical analysis of the type of events associated with time and places can trace new insights in the historical interpretation of a certain reality. We can detect patterns of human action by the most frequent event types

on a certain time frame, as well as interpret the exceptional events and evaluate their historical meaning.If we are able to cross examine the interaction between the most frequent event types and the type of arguments they were most linked to in the future, we can easily perceive historical trends in the type of events which mostly concern the Portuguese administration in the Eastern part of the Portuguese empire.

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1. pp. 86–90. Association for Computational Linguistics, Montreal, Quebec, Canada (Aug 1998). https://doi.org/10.3115/980845.980860, https://aclanthology.org/P98-1013
2. Costa, F., Branco, A.: TimeBankPT: A TimeML annotated corpus of Portuguese. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 3727–3734. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012)
3. Patto, R.A.B.: Documentos Remetidos da India ou Livros das Monções. Tomo IV. Academia Real das Ciências de Lisboa, Lisboa (1893)
4. Pearson, M.N.: The Portuguese in India. Cambridge University Press, Cambridge (1987)
5. Polónia, A.: Indivíduos e redes auto-organizadas na construção do império ultramarino português. In: Garrido, Á., Freira Costa, L., Duarte, L. (eds.) Economia, Instituições e Império. Estudos em Homenagem a Joaquim Romero de Magalhães. pp. 349–372 (2012)
6. Sacramento, A.d.S.B., Souza, M.: Joint event extraction with contextualized word embeddings for the portuguese language. In: Brazilian Conference on Intelligent Systems. pp. 496–510. Springer (2021)
7. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems, BRACIS. pp. 403–417. Springer, Rio Grande do Sul, Brazil (2020)
8. Sprugnoli, R., Tonelli, S.: Novel event detection and classification for historical texts. Computational Linguistics **45**(2), 229–265 (2019)