Information Extraction from Historical Texts: a Case Study

Paulo Quaresma[1] and Maria José Bocorny Finatto[2]

[1] Universidade de Évora, Portugal
[2] UFRGS – Universidade Federal do Rio Grande do Sul, RS, Brasil
pq@uevora.pt, mariafinatto@gmail.com

Abstract. In this paper a set of information extraction experiments over historical texts are described. The experiments were done over the Spanish book *Observaciones de Curvo* written by Suarez de Ribera in 1735 based on the 1707 Curvo Semedo's work *Observaciones medicas doutrinaes de cem casos gravissimos* to evaluate which information can be extracted in a fully automatized way.
Using publicly available NLP tools we extracted named entities (persons and places) and identified events. This information was used to populate a specialized ontology, allowing the application of powerful visualization and inference processes.
A preliminary evaluation of the quality of the extracted information showed that, in spite of the use of generic NLP tools, this process is able to automatically identify relevant information and to help human experts in the creation of historical knowledge bases.

1 Introduction

Text information extraction is an increasingly relevant NLP task, aiming to automatically structure unstructured text. On the other hand, historical documents have a huge potential amount of information, which is not easily accessible to researchers or citizens.

In this context, a project aiming to automatically populate a specialized ontology with information extracted from historical texts was created by researchers of the University of Évora, Portugal, and UFRGS – Universidade Federal do Rio Grande do Sul, RS, Brasil.

In this paper we describe the initial experiments done over the Spanish book *Observaciones de Curvo* written by Francisco Suarez de Ribera in 1735 based on the 1707 Curvo Semedo's work *Observaciones medicas doutrinaes de cem casos gravissimos*.

It is important to refer that all the processing was done by applying NLP computational tools without any human intervention: from the OCR until the

ontology population. Our main goal with this option was to analyse and evaluate how a pipeline of computational processes was able to deal with historical texts in a fully automatized way.

In the next section we will briefly describe the used corpus; in section 3 we will present the applied methodology; in section 4 a more detailed discussion of the NLP architecture will be presented; in section 5 the used ontology is described; and in section 6 a preliminary evaluation is presented and discussed.

2 Corpus

As already referred, as base corpus for the research work we have selected the book *Observaciones de Curvo* written by Francisco Suarez de Ribera in 1735 based on the Curvo Semedo's work *Observaciones medicas doutrinaes de cem casos gravissimos*. This book is available for download (in pdf and txt formats) from the Spanish National Library[3] and is composed by 549.267 tokens. It is important to refer that the text version is the output of a OCR process but it was not revised. An example of the existent problems can be seen from the initial sentences:

OBSERVACIÓN PRIMERA.
DE UNA CÓLICA NEPHRITICA, que afligió al Excekntifsimo feñor Principe de Ligne, y Marques de Arronches.
N feis de E n e r o del año de 1 68¡$. afligió al dicho Excelentifsimo feñor la cólica nephritU ca : t e n g o aíTentado, que la verdadera cien-: cia no confine en lo r u i d o f o , ó campanudo de las p a l a b r a s , ni en la apariencia, ó pompa exterior de los v e n i d o s , mas si en las obras ordenadas con acier"¡ t o , y efectuadas con felicidad:

As it can be seen from this example there are many problems with the OCR quality of the document. We had two main options: a) manually or semi-manually revise the texts; b) use the text as is, without any revision. We decided to select option b) because one of our goals was to evaluate which information can be extracted from historical documents in a fully automatized way.

A distinct approach was followed in a distinct project over the original work of Curvo Semedo[3]. In this work the basis was a fully revised version of the digitized images.

3 Methodology

As working methodology we followed a *classical* NLP pipeline of processes:

1. Lexical analysis
2. Syntactical analysis

[3] http://bdh-rd.bne.es/viewer.vm?id=0000081871&page=1

3. Semantical analysis
4. Ontology population

The lexical and syntactical analysis identifies lemmas, perform part-of-speech tagging and dependency parsing. With the semantic analysis we are able to do named entities recognition (persons, organizations, places, time) and semantic role labelling over the results of the previous modules. The last module – ontology population – receives as input the output of the previous module and creates instances of an ontology, allowing the formal representation of the extracted information.

As NLP tools we have used Freeling[5] from Lluís Padró, which supports several languages, including Spanish and Portuguese.

In the scope of this work we have focused on the extraction of entities and events and we have used a specialized ontology developed in OWL in the context of another research project[6, 7].

4 NLP tools

The architecture is based on a pipeline of NLP modules and it is represented in figure 1.

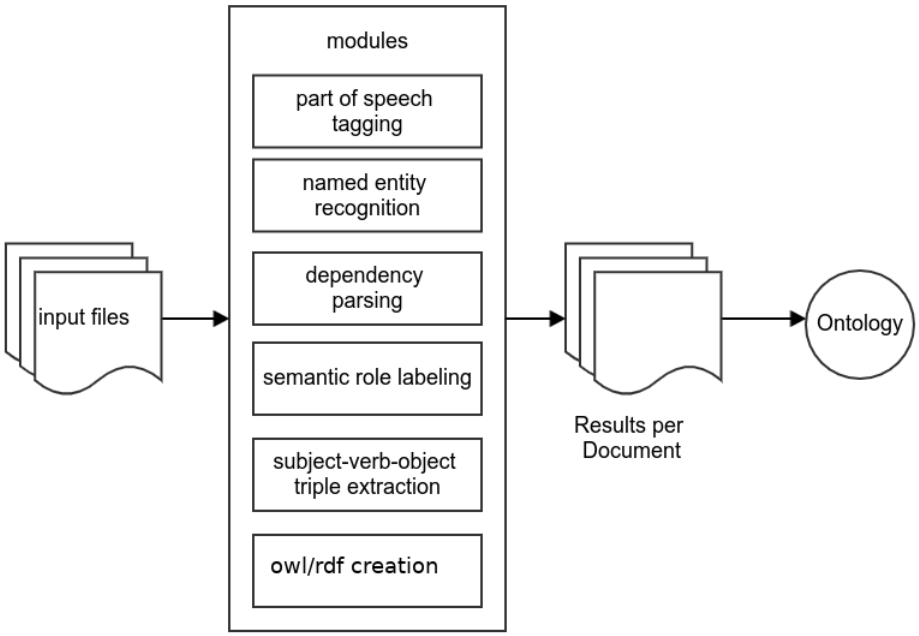
Each sentence is processed by a series of modules - part of speech tagging, named entity recognition, dependency parsing, semantic role labelling, subject-verb-object identification, and the creation of ontology instances in OWL.

The main goal is to identify events in the text, which are used to populate a predefined ontology.

As already referred we used the Freeling framework [5] for this pipeline of event extraction:

- POS tagging
  It annotates each work with an associated part-of-speech tag, using a Hidden Markov Model created for the Spanish language.
- NER
  This module is used to identify which words in the sentence are named entities (persons, organizations and locations). We did not take into account date time and currency (which can be identified by Freeling).
- Dependency parsing
  The default parser for the Spanish language was used.
- Semantic Role Labelling
  In this module expressions are annotated as A0, A1, A2, AM-LOC, or AM-ADV, which indicates if the expression is a subject, a direct or indirect object, a location or an adverbial, respectively.
- SVO triple extraction
  From the output of the previous module, it is possible to identify as subject-verb-object (SVO) triples, which constitute the basis of events.
- OWL creation
  Finally, using the previous information, RDF triples and OWL instances can generated and inserted in a specific event ontology.

Fig. 1. Architecture overview.



Below is an example of the obtained output for a simple sentence of the corpus:

```
... y como por caufa de ella murieron en una cafa cinco hijos ...
```

The NER output is:

```
y y CC 0.999989
como como CS 0.967153
por por SP 1
caufa caufa NCFS000 0.919869
de de SP 0.999961
ella él PP3FS00 1
murieron morir VMIS3P0 1
en en SP 1
una uno DI0FS0 0.951973
cafa cafa NCFS000 1
cinco 5 Z 0.999454
hijos hijo NCMP000 1
```

And the final Freeling output is:

```
Pred100.4: die.01|expire.01|perish.01 t100.20 (murieron) [
             AM-LOC t100.21 (en una cafa) [t100.21 .. t100.23]
             A1 t100.25 (cinco hijos) [t100.24 .. t100.25]
           ]
```

5 Ontology

According to [4] an ontology is a formal specification of a conceptualization; it allows the representation of entities and events, along with their properties and relations, according to a system of categories.

In the context of this work we used as basis the Simple Event Model (SEM) as a baseline model. A graphical representation of this ontology is given in Fig. 2 and its detailed design is presented in [8].

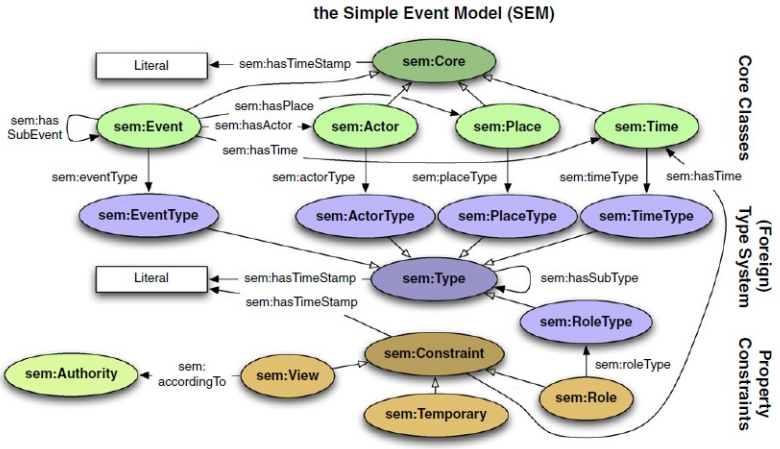


Fig. 2. the Simple Event Model

As the main concept in this ontology, we can identify *Event*, which has relations with *Actor*, *Place*, and *Time*. All of these other concepts have specific types and play roles in the event representation.

The Protege [2] tool was used for the ontology creation and GraphDB [1] for populating and querying the data. GraphDB is a Semantic Graph Database, compliant with W3C Standards, which provides the core infrastructure for ontology solutions.

6 Results

After applying the pipeline of NLP processes we were able to obtain the following output:

- NER (named entities)
  - 410 places (282 distinct)
  - 2011 persons (1294 distinct)
- Events
  - 14005 events
  - 2896 with subject – A0
  - 8271 with direct objects – A1
  - 1685 with indirect objects – A2
  - 901 with place information – AM-LOC
  - 2747 with adverbials – AM-ADV

We have done a preliminary evaluation of the quality of the extraction processes and we were able to calculate the precision of the extraction, i.e. the percentage of extracted concepts that are correct.

For each kind of information we present bellow this value and some examples:

- Places
  - Precision: 21%
    - Sevilla
    - Lisboa
    - Salamanca
    - Rúa de la Paz
- Persons
  - Precision: 22%
    - Curvo
    - Cardenal
    - Rey
    - Hypocrates
    - Galeno
- Events
  - Precision: 5% (precision was calculated over a sample of 10% of the extracted events)
    - Verbo: murieron; A1: cinco hijos; AM-LOC: en una cafa
    - Verbo: caerá; A1: la enferma; AM-ADV: en una hydropéfia
    - Verbo: dar; A1: la Unción; A2: a el paciente
    - Verbo: tomava; A1: una taza com caldo, ó agua; AM-LOC: en las manos; AM-ADV: hirviendo

As it can be seen from these examples, it was possible to extract relevant information and, in some cases, with an associated high level of complexity (see, for instance, the last example of the *events*). Nevertheless, the precision of extraction is still quite low.

We have made an analysis of the main error situations and the main sources of errors can be characterized in the following way:

– Places: Most of the errors are related with incorrect OCR and with misclassified entities (e.g. persons classified as places).
– Persons: The main source of errors for this class of entities is clearly the bad quality of the initial OCR, which created many incorrect words that, being unknown to the NLP tools, tend to be classified as proper nouns.
– Events: As expected, the precision for this kind of information is very low. This can be explained, again, by the poor quality of the initial OCR system and the nonexistence of a lexical and syntactical module adapted to the Spanish language of the XVIII century.

7 Conclusions and Future Work

As main conclusion we believe we were able to show that it is possible to use standard NLP computational tools to automatically extract information from historical texts.

It is important to emphasize that we presented an initial evaluation with a not revised corpus directly generated from a OCR system. Thus, the obtained results are far from perfect and they show the relevance of having good quality texts as input to the processing pipeline. Nevertheless, the proposed approach can be used as a baseline and a basis for additional research work.

As future work, a deeper evaluation of the results should be done and NLP tools adapted to the lexicon and syntax of the historical corpora need to be developed.

We will also foresee the creation of a web based application for the visualization and access to the created ontology.

References

1. Graphdb, http://graphdb.ontotext.com/, [Available online: accessed on 24/02/2020]
2. Protege, https://protege.stanford.edu/, [Available online: accessed on 24/02/2020]
3. Finatto, M.J., Quaresma, P., Gonçalves, M.F.: Portuguese corpora of the 18th century: old medicine texts for teaching and research. In: Fiser, D., Pancur, A. (eds.) Proceedings of the Conference on Language Technologies and Digital Humanities, Ljubljana, Slovenia, September 20-21 (2018), http://dspace.uevora.pt/rdpc/handle/10174/23606
4. Guarino, N., Giaretta, P.: Ontologies and knowledge bases: Towards a terminological clarification. In: Towards very Large Knowledge bases: Knowledge Building and Knowledge sharing. pp. 25–32. IOS Press (1995)
5. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012). ELRA, Istanbul, Turkey (May 2012)
6. Quaresma, P., Nogueira, V.B., Raiyani, K., Bayot, R., Gonçalves, T.: From textual information sources to linked data in the agatha project. In: INAP – Proceedings of the 22nd International Conference on Applications of Declarative Programming and Knowledge Management, Cottbus, Germany, September 9-13, 2019. pp. 1–11 (2019)

7. Quaresma, P., Nogueira, V.B., Raiyani, K., Bayot, R.: Event extraction and representation: A case study for the portuguese language. Information 10(6) (2019), https://www.mdpi.com/2078-2489/10/6/205
8. Van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the simple event model (sem). Web Semantics: Science, Services and Agents on the World Wide Web 9(2), 128–136 (2011)