Text Mining for History: first steps on building a large dataset

Suemi Higuchi, Cláudia Freitas, Bruno Cuconato, Alexandre Rademaker
FGV/CPDOC, PUC-Rio, FGV/EMAp, IBM Research and FGV/EMAp
suemi.higuchi@fgv.br, claudiafreitas@puc-rio, bcclaro@gmail.com, alexrad@br.ibm.com

Abstract

This paper presents the initial efforts towards the creation of a new corpus on the history domain. Motivated by the historians' need to interrogate a vast material - almost 9 million words - in a non-linear way, our approach privileges deep linguistic analysis on an encyclopaedic style data. In this context, the work presented here focuses on the preparation of the corpus, which is prior to the mining activity: the morphosyntactic annotation, the definition of semantic types for named entity (NE) and named entities relations relevant to the History domain. Taking advantage of the semantic nature of appositive structures, we manually analysed a sample of 1,049 sentences in order to verify its potential as additional semantic clues to be considered. The results show that we are on the right track.

Keywords: digital humanities, text mining, corpus annotation, appositives

1. Introduction

Language is a rich repository of information about our practices, constituting a raw material for research in the Human and Social Sciences. In close connection with Computational Linguistics, Humanities and Social Sciences, the growing and fertile field of the Digital Humanities has at its disposal tools and resources that offer new ways to explore text's materiality. In this paper we present the initial efforts towards the creation of a golden resource dedicated to text mining in the history domain. The mining strategy is linguistically motivated: we believe that certain semantic relations, relevant to the history domain, have a linguistic realization, and therefore it is essential the inclusion of linguistic data, such as part-of-speech, lemma, and syntactic information, in the corpus. The target of the mining - and the corpus - is the *Dicionário Histórico-Biográfico Brasileiro* (Brazilian Historical and Biographical Dictionary), DHBB for short, that contains almost 8,9 millions of tokens and about 307 thousand of sentences.

DHBB is a reference work, written by historians and published by the Research and Documentation Center of Brazilian Contemporary History (CPDOC) of Getulio Vargas Foundation (FGV). It contains almost eight thousand entries with information ranging from the life and career trajectories of individuals, to the relationships between the characters and events that the country has hosted. The main motivation to mine the DHBB came from the need of researchers to query the material looking for information that requires almost total reading of the whole body of texts, such as kinship (or personal) relationship between politicians; their connection to entities such as institutions, movements, events or places, throughout their public life, or even the desire to answer questions such as "Which politicians were born before the 1960s, had military training, and held a position in the Executive Branch?".

We know there are a vast amount of knowledge spread around the entries in a non-linear way. After all, dictionaries and encyclopaedias are made to be consulted, and not to be read linearly. In this context, the focus of this paper is to report the first efforts related to the preparation of the material – in particular, the morphosyntactic linguistic annotation, the definition of semantic types for named entity (NE) and named entities relations relevant to the History domain, taking the appositives as an important feature to observe when analyzing the syntactic structures within the sentences. For the annotation of the semantic relations in the corpus, we will make use of linguistically motivated rules, inspired by strategies like (Santos and Mota, 2010).

Our main purpose is not only to explore and mine the DHBB, but to create a public corpus to foster Portuguese NLP in general, and NLP in the history domain, in particular. Most of large Portuguese annotated corpora are composed by newspaper texts; DHBB entries, on the other hand, are written in encyclopaedic style, and this "novelty" can be a challenge for automatic parsers.

2. Corpus Preparation

It is important to mention that the first edition of the DHBB dates from 1984 in printed version only, and it was only in 2010 that its content was fully brought to the Internet. Since its begining, CPDOC has developed a internal information system to maintain the data through forms and reports that interact with a relational database. The database structure can be summarized as one main table that contained a text field with the entries encoded in HTML and some few metadata: basically, the name of the entry and its nature (if biographical or thematic). This structure showed to be quite limiting when it concerns maintenance and improvements to the dictionary. These issues are described in details in (Paiva et al., 2014) that also explains how the nature of DHBB's data suggested that its entries could be easily maintained as text files using a lightweight human-readable markup syntax, like YAML (Ben-Kiki and Evans, 2005) and Markdown (Gruber, 2004). A considerable effort was made then to bring up this structure and the final adoption of plain text files was justified by some clear reasons: easiness of maintenance using any text editor (tool independence); conformity to long-term standards by being software and platform independent; easiness to exploit the possibilities of DHBB's files as a resource for NLP; enrichment of the entries with metadata of any kind at any time, even those extracted from NLP.

DHBB corpus will be publicly available according to the Universal Dependencies scheme (Straka and Straková, 2016), which has the advantage of providing a multilingual environment for NLP. However, in this first stage, we used

two different parsers in order to evaluate the corpus processing: PALAVRAS (Bick, 2000), a rule-based multi-level constraint grammar parser, developed specifically for the Portuguese language, and UDPIPE (Straka and Straková, 2016), a machine learning pipeline for tokenization, tagging, lemmatization and dependency parsing. UDPipe is language-independent and can be trained given annotated data in CoNLL-U format.[1]

The motivation for the double processing is twofold: first of all, we aim to compare linguistic analysis of both systems in a genre (encyclopedic) unusual to parsers. Additionally, we believe that comparing the outputs of different systems is a way to optimize the linguistic revision, as suggested by (Truggo, 2016).

Besides morphosyntactic information provided by parsers, the corpus contains information related to named entities and named entities relations, provided by both simple linguistic analysis (Section 3.) and lexical-syntactical patterns, taking advantage of the highly predictable written style of the DHBB.

3. Named Entity and Relations Between Named Entities

Named Entity recognition is a crucial task for text mining, since its main focus is on specific instances of general semantic types like person, location, time and organization. Our definition of named entity closely follows the ACE (Automatic Content Extraction) proposals (Doddington et al., 2004), capturing all kinds of information that can identify something or someone relevant, whether it's a proper name or not. In an entry about *Revolução de 1930* (Revolution of 1930), for example, we want to recover data about this specific event even when it is refered as *revolução* (revolution) as in "Essa carta pode ajudar no esclarecimento de um ponto importante das articulações da *revolução*, pois a bibliografia sobre o período refere-se a dois encontros entre Vargas e Prestes" (This letter can help clarify an important point of the articulations of the *revolution*, since the bibliography on the period refers to two meetings between Vargas and Prestes).

To determine the semantic types relevant for the history domain, we combined knowledge from domain experts and a corpus driven approach, based on a wide reading of entries, aimed to validate and increase the initial proposed classes. As a result, we devised seven classes, presented in Table 1. Taking into account the information we would like to extract from DHBB and inspired by the set of relations proposed by the Second HAREM task (Freitas et al., 2008), we devised our own set of relations to connect the entities. However, during the process of text analysis (in particular, looking at appositives occurrences) a few other relations were identified as relevant to our goals. Table 2 presents the final list of relations and examples..

Classes for DHBB	Examples
[PER] Person	*Getúlio Vargas, Lula, presidente*
[ORG] Organization	*Petrobras, Partido Democrático Social, PDS*
[POL] Political Formulation	*Plano Collor, Programa de Estabilização Monetária, AI-5*
[EVN] Event	*Revolução de 1930, Atentado do Riocentro*
[LOC] Local	*São Paulo, palácio Guanabara*
[DOC] Document	(Titles of books, newspapers, magazines, personal diaries)
[TME] Time	*Janeiro de 2001, 1927 a 1929*

Table 1: Entity Classes for DHBB

Relations	Context
[ident]: is correference of	*...Partido dos Trabalhadores (PT)*
[role]: is role of	*...Alberto Coelho, president*
[loc]: is local of	*...port of Alcantara, in Lisbon*
[part]: is part	*...in Porto Seguro (BA)*
[date]: is a date of	*...promulgation of Nova Carta (18/9/1946)*
[link-inst]: is an institution linked to	*...Vandilson Costa, from Partido Comunista do Brasil*
[link-fam]: is a family relation of	*...Nilo Augusto ..., son of Gercino Coelho and Eunice Coelho.*
[link-pers]: is a personal relation of	*...Orígenes Lessa, friend of his brother Fúlvio*
[attrib]: is an attribute of	*...João Abdalla and Amélia Abdalla, of Arab origin*

Table 2: Relations between entities

4. DHBB: Hands On

In order to compare the performance of UD Pipe and PALAVRAS, parsers with different approaches to linguistic analysis, we conducted a manual analysis of a sample of 39,668 tokens in 1,049 sentences (thirty entries), focusing on proper noun segmentation/named entity identification and appositives.

4.1. Proper nouns identification and segmentation

Prior to semantic classify the named entities (NE) we need to correct identify them. By the highly idiosyncratic nature of proper nouns, errors resulting from a wrong segmentation are usual. Both PALAVRAS and UDPipe, for example, considered the name *Ministério das Minas e Energia* (Ministry of Mines and Energy) as two separated names: *Min-*

[1]The conversion process from PALAVRAS to *Universal Dependencies* follows the procedures described at (Rademaker et al., 2017).

istério das Minas and *Energia*. Regarding names of people, UDPipe tagged the last name of the person *José Afonso de Melo*, as a noun modifier of the first, and not as part of the whole name. In this case, PALAVRAS did it right, joining the tokens in a single token.

We tried to reduce the segmentation errors of names creating domain lexicons from external resources like category pages of Wikipedia and DHBB metadata (names of biographees and event entries, as well as positions held by biographed person). We also used simple pattern recognition in the texts. We looked for simple patterns like *presidir o [A-Z]* (to chair the [capital letter]) or *estudar em [A-Z]* (to study at [capital letter]). This process lead to more than 15 thousand entities, classified as *events*, *person*, *documents* and *organizations*. The lexicons are available at the project repository.[2]

Some case studies demonstrated positive results when adopting similar approach of using lexicons. In (Florian et al., 2003), for example, the authors investigated the combination of a set of diverse statistical named entity classifiers applied to an English corpus: when no gazetteer (lexicons) or other additional training resources are used, the combined system attains a performance of 91.6 F1 on the English development data; but after integrating gazetteers containing some 50 thousand names of cities, 80 thousand proper names and 3,5 thousand organizations, the F-measure error was reduced by a factor of 15 to 21%.

In order to evaluate the impact of the incorporation of the lexicons, we compared the outputs of UDPipe and PALAVRAS against our golden sample, manually revised. We know that the use of lexicons has limitations such as the cost of its maintenance and its possible limited coverage, specially if we consider that the same person can be mentioned by many different names (ranging from the formal complete name to the most common nickname). On the other hand, we believe that the incorporation of lexical entries, associated with semantic classes, are a simple and effective method to bootstrap the creation of lexical-syntactic patterns, crucial for semantic annotation between NEs.

We evaluated the impact of lexicon incorporation at the automatic processing. Our lexicon of names of people has 19,466 entries. In the sample, we found 358 mentions of names from the lexicon (34% of the sentences). The most frequent name occurs 20 times. We estimate that 107 errors in the segmentation of names with more than three tokens could be fixed when lexicons are used to post-processing the parser analysis. It is worth mentioned that the idea of external lexicons is to bootstrap the linguistic analysis and linguistic revision, providing not only proper nouns with correct segmentation, but also named entities with semantic labels.

The lexicon of names of organizations has 3,642 entries. In the sample, we found 229 occurrences of names of organizations (22% of the sentences). The most frequent names has 84 occurrences. We estimated 53 errors in the segmentation of names of organizations with more than three tokens could be fixed in the post-processing of the parser analysis.

[2] `github.com/cpdoc/dhbb/`

4.2. Appositives

Appositives are syntactic relations especially productive for text mining. They provide descriptive information about the head noun, thus enriching its characterization. When a given noun is tagged as an appositive, a relationship with another term is derived, as we could see in the examples in Table 2, which can generate tuples as:

- ident(Partido dos Trabalhadores, PT)
- role(president, Alberto Coelho)
- link-fam(son, Nilo Augusto)

Considering our golden sample, we conducted an evaluation of appositives in the corpus. Table **??** presents the overall amount of appositives identified by each parser and by the human revision.

	PALAVRAS	UDPipe	Golden
Total	735	880	801

Table 3: Evaluation of Appositives in the Corpus

As can be seen, PALAVRAS recognized 735 cases of appositives and UDPipe 880 cases.[3] Performing the task manually, we recognized 801 occurrences of appositives. Although this result seems to provide us a measure of how close the parsers have been to the human analysis, it does not mean that the mistakes and correctness were given in the same place of the text, and therefore, that one is better than the other. This is still an analysis to be made in the future.

It is worth noting that appositives are tricky linguistic structures to parse automatically, since its main formal clue, the punctuation, can be easily confused with coordination and vice-versa. For example, in the sentence "Entre 1959 e 1960, coordenou o setor financeiro da campanha eleitoral do marechal Henrique Teixeira Lott, candidato à presidência da República apoiado pelo PSD e o PTB." (Between 1959 and 1960, he coordinated the financial sector of the election campaign of Marshal Henrique Teixeira Lott, candidate for the presidency of the Republic supported by the PSD and PTB), UDPipe erroneously considered "candidato" as in coordination with "setor". Also, in the sentence "... votou a favor da emenda constitucional que previa a reeleição de presidente da República, governadores e prefeitos, ..." (... voted in favor of the constitutional amendment that foresees the reelection of the president of the Republic, governors and mayors, ...), both PALAVRAS and UDPipe were mistaken in identifying an appositive structure between "governadores" and "reeleição" when it is clear that this is a case of coordination.

Besides, the explicit semantic nature of appositives led us to a semantic strategy for revision the parser analysis.

[3] PALAVRAS uses two tags to indicate the general idea of appositives (we take both into account) and UD uses only one.

That is, we extracted from the parsed sentences the triples formed by the `appos` relation – the linearization of the noun phrases that have theirs heads connected by a `appos` relation. The extracted triples can be trivially analysed by an human and abnormal noun phrases indicate a possible parser mistake.

From the 801 manually verified appositives, in Table 4 we present the distribution of semantic relations that we associated for each appositive relation.

Freq	semantic relation
2	link-pers
11	loc
11	link-inst
21	govern
59	link-fam
61	date
62	part
78	attrib
193	ident
289	role

Table 4: Appositives relations and their types

5. Concluding Remarks

In this paper, we present the first efforts towards the creation of a corpus for the history domain, an ongoing research for a research group on digital humanities from the CPDOC/FGV.

Motivated by the historians' need to interrogate a vast text material in a non-linear way, our approach privileges deep linguistic analysis, as opposed to shallow techniques, such as topic modeling, which we believe to be complementary. In this context, a crucial step is to prepare the material that will be mined. In our case, the preparation includes the following levels of annotation: morphosyntactic, named entities and relationships between entities. Although the morphosyntactic annotation is already being successfully performed automatically, it is not be possible to rely entirely on the results. As we have seen, almost 10% of the revised cases of appositives were wrong analyses. As to semantic annotation, at least for the Portuguese language, it is still unreliable. In addition, there are some domain particularities, like entities such as "Policy Formulation", that would hardly be included in general-purpose NE systems. As a way to overcome these difficulties, we invested in the derivation of semantic relations from certain syntactic relations, such as appositive structures, and in the creation of linguistically motivated rules. There is a long way to go.

6. Bibliographical References

Ben-Kiki, O. and Evans, C. (2005). Yaml ain't markup language (yaml^{TM}) version 1.1.

Bick, E. (2000). *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, pages 837–840.

Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named entity recognition through classifier combination. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 168–171, Stroudsburg, PA, USA. Association for Computational Linguistics.

Freitas, C., Santos, D., Gonçalo Oliveira, H., Carvalho, P., and Mota, C. (2008). Relações semânticas do rerelem: além das entidades no segundo harem. *Linguateca*, pages 77–96, Jan.

Gruber, J. (2004). Markdown language.

Paiva, V. D., Oliveira, D., Higuchi, S., Rademaker, A., and Melo, G. D. (2014). Exploratory information extraction from a historical dictionary. In *IEEE 10th International Conference on e-Science (e-Science)*, volume 2, pages 11–18. IEEE, October.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva Universal Dependencies for Portuguese, V. (2017). Universal dependencies for portuguese. In *Proceedings of the International Conference on Dependency Linguistics*, Pisa, Italy, September.

Santos, D. and Mota, C. (2010). Experiments in human-computer cooperation for the semantic annotation of portuguese corpora. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Straka, M. and Straková, J. (2016). UDPipe. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Truggo, L. F. (2016). Classes de palavras - da grécia antiga ao google: um estudo motivado pela conversão de tagsets. Master's thesis, Programa de Pós-Graduação em Estudos da Linguagem, Departamento de Letras PUC-Rio.