Automatic information extraction:
a distant reading of the Brazilian Historical-Biographical Dictionary

PROPOR 2022
March 21st-23rd, 2022

Suemi Higuchi, Claudia Freitas, Diana Santos
CPDOC/FGV, PUC Rio, Linguateca & UiO

Introduction

We present some results of applying natural language processing (NLP) techniques in the domain of History, having as object of investigation the Brazilian Historical-Biographical Dictionary (DHBB).

After improving or adding annotation of specific fields, information extraction techniques based on manually derived patterns were applied to three relevant problems: the age of entrance in Brazilian politics, the academic background of Brazilian politicians, and family ties among the political elites.

- Dicionário Histórico-Biográfico Brasileiro (DHBB) from FGV;
- Encyclopedic work;
- More than 7,500 biographical and thematic entries;
- Domain: contemporary history of Brazil (from 1930 to nowadays)

The research

Aim: to create, from the DHBB, an annotated corpus for automatic information extraction's purpose.

Motivation: Researches demand: "Is it possible to extract certain information without having to consult individual entries?"

Examples:

- How is the educational background of political cadres characterized over time?
- How old were Supreme Court ministers when they were nominated?
- Who are the politicians who have family ties to other politicians? Which ties?

Metadata is not enough
Close reading is a hard task

Solution: NLP methods and techniques
Distant reading of the DHBB

Methodology

- Creation of the corpus
  - Available at https://github.com/cpdoc/dhbb
- Annotation of the corpus
- Using PALAVRAS parser
- Incorporation into Linguateca (AC/DC format)
  - Available at https://www.linguateca.pt/acesso/corpus.php?corpus=DHBB
- More annotation
  - Processed by AC/DC project
- Addition of semantic layer
- Improvement NER with lists
- Definition of the themes for extraction
- Identification of the set of patterns
- Extraction using the patterns

PARSER

Corpus creation and annotation

each entry = a document containing metadata and text

- *Tokenization*
- *Morphosyntactic and semantic analysis*
- *Conversion to AC/DC format*
- *More morphosyntactic and semantic analysis*

Annotated corpus

Morphosyntactic tags
- lemmas
- part of speech (pos tag)
- verb tenses
- singular and plural nouns
- syntactic function

Information provided by metadata:
- full name of the entry holder (*fonte*)
- original identification (*entidade*)
- gender (*sexo*)
- positions held (*cargos*)

Semantic annotations:
- proper names (people, organizations, places, etc.)
- family relationships (lexicon of terms that denote kinship relationship)

Other semantic information can be aggregated
= endless enrichment possibilities

Use of rules to improve NER

Correspondence between names referring to the same person:

Lemmas (nicknames, parts of the name, typos...)		DHBB entry name (entidade)
Bolsonaro	=	Jair Messias Bolsonaro
Getulio Vargas	=	Getúlio Dornelles Vargas
Jango	=	João Belchior Marques Goulart
Lula	=	Luis Inácio da Silva

To fix segmentation errors:

Simpatizante	de	Vargas
PoS = noun	PoS = preposition	PoS = proper name

Extraction

Extraction approach

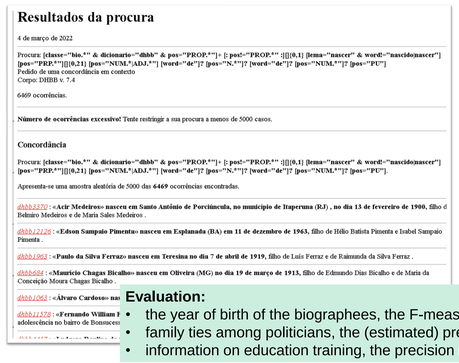
The methodology is based on the use of textual patterns, where extraction rules are manually created and rely mainly on lexical-syntactic aspects and semantic annotation of the corpus

For each theme:

Compilation of the set of patterns
- observe in a sample of entries how the sentences that bring the desired information are constructed;
- translate these constructions into lexical-syntactic patterns with regular expressions; concatenate all expressions.

Extraction
- query the corpus;
- postprocess the results using R;
- specific information is extracted and crossed with metadata.

YEAR OF BIRTH	
Examples of sentences	«Moroni Bing Torgan» nasceu em Porto Alegre, no dia 10 de junho de 1956. «Álvaro Francisco de Sousa» nasceu no dia 28 de fevereiro de 1903.
Pattern	[classe="bio.*" & dicionario="dhbb" & pos="PROP.*"]+ [: pos!="PROP.*" :][]{0,1} [lema="nascer" & word!="nascido|nascer"] [pos="PRP.*"][]{0,21} [pos="NUM.*|ADJ.*"] [word="de"]? [pos="N.*"]? [word="de"]? [pos="NUM.*"]? [pos="PU"]
Occurrences	6,469



Evaluation:
- the year of birth of the biographees, the F-measure was **.99**
- family ties among politicians, the (estimated) precision was **.59**
- information on education training, the precision reached **.991**

Some distant readings from DHBB

The age of starting public careers, per generation

- Those aspiring to political careers are getting in the public service steadily younger

The education background of the politicians

Areas of training most frequent among the biographees, throughout the generations

Area	ALL	undated entries	b. 1900 (geração 1)	1901-1920 (geração 2)	1921-1940 (geração 3)	1941-1960 (geração 4)	1961-1980 (geração 5)	1981-2000 (geração 6)
Direito	2.914	59	618	711	829	559	133	5
Militar	854	18	316	326	162	27	4	1
Engenharia	733	16	157	135	189	206	27	3
Medicina	672	19	157	163	122	195	15	1
Administração	516	11	7	35	132	256	71	4
Economia	517	11	10	73	132	258	32	1
Filosofia	267	7	26	63	97	64	10	-
Contabilidade	186	7	4	42	77	52	4	-
Letras	179	2	30	35	46	50	16	-
C. Sociais	137	2	9	17	40	55	13	1

- among the 48 found areas, in the overall reckoning law school appears preponderant in all generations, followed by military training, with just one third of the first;
- decline in military training from the second to the third generation;
- civilian training was replacing the military career as the most suitable way to reach important political positions.

Family ties in politics

Distribution of family ties for Presidents and Senators

Biographies with family ties/ total of biographies	Born:	Identification of at least one family tie with another biographee / total of biographees			
		Men	Women	ALL	%
Presidents 13 /26 50 %	NO DATE	0 /2	-	0 /2	0 %
	Before 1900	6 /12	-	6 /12	50 %
	1901 – 1920	5 /8	-	5 /8	62,5 %
	1921 – 1940	2 /2	-	2 /2	100 %
	1941 – 1960	0 /2	-	0 /2	0 %
	1961 – 1980	-	-	-	0 %
	1981 – 2000	-	-	-	0 %
Senators 260 /742 35 %	NO DATE	5 /17	0 /1	5 /18	27,7 %
	Before 1900	89 /169	-	89 /169	52,6 %
	1901 – 1920	72 /193	-	72 /193	37,3 %
	1921 – 1940	56 /183	0 /5	56 /188	29,8 %
	1941 – 1960	31 /139	4 /12	35 /151	23,2 %
	1961 – 1980	3 /17	0 /6	3 /23	13,0 %
	1981 - 2000	-	-	-	-

- Presidents and senators are the politicians who most appear with family ties, 50% and 35% respectively, this being most perceived in the first generations (in the present version of DHBB). Ministers and deputies follow with 18%, keeping a stable average over generations.

Final considerations

Some challenges and issues:

- How to find a balance between a sufficient number of patterns and a good enough coverage?
- The manual work required to create the expressions;
- Languages natural expressiveness.

Main contributios:

- Creation of an annotated encyclopedic corpus made available for language and humanities studies;
- Presentation of a methodology based on a philosophy of cyclical enrichment: the more information is obtained, the more it is added to the corpus itself;
- Compilation of a set of patterns that can be adapted to other corpora containing a similar type of annotations.