A benchmarking for public information by Machine Learning and Regular Language

Fernando Antonio Dantas Gomes Pinto[1], **Jefferson de Barros Santos**[2], **Sérgio Lifschitz**[1], **Edward Hermann Haeusler**[1]

[1]Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)
Departamento de Informática
Rio de Janeiro – RJ – Brazil

[2]Fundação Getúlio Vargas (FGV)
Rio de Janeiro – RJ – Brazil

{fpinto, sergio, hermann}@inf.puc-rio.br, jefferson.santos@fgv.br

Abstract. *Technologies such as Big Data and Transfer Learning have been attracting the interest of industry and academia over the last 15 years. The consequence of this is an almost unanimous preference for technological solutions that use statistical models. This technology is causing a revolution in the information extraction process. In this research, we question whether this technique is the best solution for extracting information from documents. We compare machine learning (ML) and rule-based approaches in the task of recognizing legal entities in the official gazette. We built an annotated dataset with 100 examples of legal documents and submitted this model to an evaluation in IBM Watson Knowledge Studio (WKS). We show that, in a scenario where documents follow a formal structure, rules-based information extraction systems still present themselves as low-cost, more uncomplicated, and more efficient solutions.*

1. Introduction

Information Retrieval (IR) is an area of Computer Science (CS) that deals with the storage of documents and the automatic retrieval of information associated with them. In this process, two techniques are widely used in industry and academia. Mostly the academy has been studying and developing Machine Learning (ML) techniques and the industry has been applying these techniques in real cases, in contrast to the sole use of Rule-based techniques as was the case in the 70s and mid-80s.

Generally, the task of locating and categorizing important nouns and proper nouns in a text written in natural language uses two approaches. The first approach uses supervised machine learning [Mohit 2014], and the second uses regular expressions to locate and categorize legal entities in legal text.

During our research, we observed that the official gazette texts could be classified as semi-structured data. Some sections of the text present a regularity in the disposition of the data, while others appear in a poorly structured way. For example, legal statements that distribute public office are published in well-formatted (structured) sections. The more general information in the official gazette is treated in an unstructured way. In [Constantino et al.], the authors present an ML approach to segmentation in an attempt

to minimize the complexity of identification and extraction in semi-structured documents. In section 4.1 we present our approach with regular expressions.

As the official gazette is a semi-structured document, we question whether the Regular Expression (RE) and Parsing Expression Grammar (PEG) approaches are a viable alternative when compared to an ML approach in the tasks of recognizing entities contained in the official gazette.

The purpose of this article is to present a strand of our research on the use of two techniques for extracting information from legal documents, especially documents that have a structured pattern.

With this, we briefly discuss some technological challenges in the area of Natural Language Processing (NLP) in conducting the research and evaluating both approaches quantitatively. We believe that this discussion can support public managers in a more assertive way when presented with similar and real scenarios.

We will apply both techniques as part of the task of extracting public acts contained in official gazettes. In this research, the entities extracted were tripled to an RDF [1] format and can be further submitted to SPARQL queries in a triple store, as for example the AllegroGraph [Buil-Aranda et al. 2013]. The extraction of the acts from the gazettes to a knowledge base is part of a wider project of KB creation for public documents in the context of e-governance transparency and accountability.

In this paper, we only focus on the task of extracting public acts and the comparison between the approaches. For more details on the whole research and development project, the paper [Pinto et al. 2021] can be consulted.

This research is divided into the following sections: Section 1, this introduction, presents our research motivation. Section 2, discusses the structures of legal documents and the official gazette. Section 3, discusses Related Works. Section 4 presents the research design and methods. In Section 5, we present the Results. In section 6, we make conclusions about research and Future Works, and finally, in Section 7, references.

2. Well Formatted Document

Our objective has always been motivated by the ability of these well-formatted documents, containing formatted legal texts, to reflect a formalism that allows us to apply techniques that map their grammar to rule-based extraction systems.

A clear example of this category of documents is the official gazettes. Generally, these documents do not have a formal organization of sections. Each official gazette, published daily, discloses the most diverse categories of public administration acts. However, even though these documents seem to lack any formalism, some sections, such as those dealing with hiring people for dear audiences, follow some legal rules of publication as highlighted in Figure 1.

Figure 1 shows the kind of legal act available in the official gazettes. Therefore, these characteristics (structure) motivate us to question whether ML is a viable and efficient solution for such extraction tasks.

[1] RDF is a machine-readable data exchange format.

Fundamento: Não sujeito à Lei Federal 8.666/93
Valor: R$ 72.000,00(setenta e dois mil reais)
Autorização: ROSEMARY VIDAL

PROCESSO Nº: 01/900.026/2014- NAD Nº 006/2014
Objeto: Fornecimento de vale transporte
Partes: F-ARTES e FETRANSPOR
Fundamento: Art. 25, Caput da Lei 8.666/93 e suas alterações.
Razão: Inexigibilidade de Licitação
Valor: R$ 33.020,00(trinta e três mil e vinte reais)
Autorização: ROSEMARY VIDAL
Ratificação: EMILIO KALIL

PROCESSO Nº: 01/900.030/2014- NAD Nº 008/2014
Objeto: Prestação de serviços de fornecimento de gás
Partes: F-ARTES e CEG
Fundamento: Art. 25, Caput da Lei 8.666/93 e suas alterações.
Razão: Inexigibilidade de Licitação
Valor: R$ 240.000,00 (duzentos e quarenta mil reais)
Autorização: ROSEMARY VIDAL
Ratificação: EMILIO KALIL

PROCESSO Nº: 01/900.027/2014- NAD Nº 009/2014
Objeto: Prestação de serviços de água/esgoto
Partes: F-ARTES e CEDAE
Fundamento: Art. 25, Caput da Lei 8.666/93 e suas alterações.
Razão: Inexigibilidade de Licitação
Valor: R$ 524.800,00 (quinhentos e vinte e quatro mil e oitocentos reais)
Autorização: ROSEMARY VIDAL
Ratificação: EMILIO KALIL

PROCESSO Nº: 01/900.028/2014- NAD Nº 011/2014
Objeto: Contribuição patronal PASEP
Partes: F-ARTES e BANCO DO BRASIL
Fundamento: Não sujeito à Lei Federal 8.666/93
Valor: R$ 27.720,00 (vinte e sete mil, setecentos e vinte reais)
Autorização: ROSEMARY VIDAL

PROCESSO Nº: 01/900.031/2014- NAD Nº 010/2014
Objeto: Prestação de serviços de luz e força motriz
Partes: F-ARTES e LIGHT
Fundamento: Art. 25, Inciso XII, da Lei 8.666/93 e suas alterações.
Razão: Dispensa de Licitação
Valor: R$ 2.340.000,00 (dois milhões, trezentos e quarenta mil reais)
Autorização: ROSEMARY VIDAL
Ratificação: EMILIO KALIL
*Omitidos no D.O.Rio nº 197, de 03/01/2014.

SECRETARIA DA CASA CIVIL

Secretário: **Pedro Paulo Carvalho Teixeira**
Rua Afonso Cavalcanti, 455 - 13ºandar - Tel.: 2976-3187

RESOLUÇÃO "P" Nº 169 DE 13 DE FEVEREIRO DE 2014
O SECRETÁRIO CHEFE DA SECRETARIA MUNICIPAL DA CASA CIVIL, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE
Exonerar, a pedido, **ANAMARIA CARVALHO SCHNEIDER**, matrícula 57/253.542-5, com validade a partir de 3 de fevereiro de 2014, do Cargo em Comissão de Coordenador I, símbolo DAS-09, código 039447, da Coordenadoria de Demandas Institucionais, da Subsecretaria de Gestão, da Secretaria Municipal de Saúde.

RESOLUÇÃO "P" Nº 170 DE 13 DE FEVEREIRO DE 2014
O SECRETÁRIO CHEFE DA SECRETARIA MUNICIPAL DA CASA CIVIL, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE
Exonerar **STAEL CHRISTIAN RIANI FREIRE**, matrícula 60/293.276-2, do Cargo em Comissão de Gerente II, símbolo DAS-07, código 039394, da Gerência de Atendimento a Demandas, da Coordenadoria de Administração de Contratos de Gestão com Organizações Sociais, da Subsecretaria de Gestão, da Secretaria Municipal de Saúde.

RESOLUÇÃO "P" Nº 171 DE 13 DE FEVEREIRO DE 2014
O SECRETÁRIO CHEFE DA SECRETARIA MUNICIPAL DA CASA CIVIL, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE
Nomear **STAEL CHRISTIAN RIANI FREIRE**, matrícula 60/293.276-2, para exercer o Cargo em Comissão de Coordenador I, símbolo DAS-09, código 039447, da Coordenadoria de Demandas Institucionais, da Subsecretaria de Gestão, da Secretaria Municipal de Saúde.

RESOLUÇÃO "P" Nº 172 DE 13 DE FEVEREIRO DE 2014
O SECRETÁRIO CHEFE DA SECRETARIA MUNICIPAL DA CASA CIVIL, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLUÇÃO "P" Nº 169 DE 13 DE FEVEREIRO DE 2014
O SECRETÁRIO CHEFE DA SECRETARIA MUNICIPAL DA CASA CIVIL, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE
Exonerar, a pedido, **ANAMARIA CARVALHO SCHNEIDER**, matrícula 57/253.542-5, com validade a partir de 3 de fevereiro de 2014, do Cargo em Comissão de Coordenador I, símbolo DAS-09, código 039447, da Coordenadoria de Demandas Institucionais, da Subsecretaria de Gestão, da Secretaria Municipal de Saúde.

RESOLUÇÃO "P" Nº 170 DE 13 DE FEVEREIRO DE 2014
O SECRETÁRIO CHEFE DA SECRETARIA MUNICIPAL DA CASA CIVIL, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE
Exonerar, a pedido, **ANAMARIA CARVALHO SCHNEIDER**, matrícula 57/253.542-5, com validade a partir de 3 de fevereiro de 2014, do Cargo em Comissão de Coordenador I, símbolo DAS-09, código 039447, da Coordenadoria de Demandas Institucionais, da Subsecretaria de Gestão, da Secretaria Municipal de Saúde.

RESOLUÇÃO "P" Nº 170 DE 13 DE FEVEREIRO DE 2014
O SECRETÁRIO CHEFE DA SECRETARIA MUNICIPAL DA CASA CIVIL, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE
Exonerar **STAEL CHRISTIAN RIANI FREIRE**, matrícula 60/293.276-2, do Cargo em Comissão de Gerente II, símbolo DAS-07, código 039394, da Gerência de Atendimento a Demandas, da Coordenadoria de Administração de Contratos de Gestão com Organizações Sociais, da Subsecretaria de Gestão, da Secretaria Municipal de Saúde.

DESPACHOS DO DIRETOR
EXPEDIENTE DE 14/02/2014

01/800.185/2014 - Autorizo a despesa na forma abaixo:
1. Objeto: Publicação em jornal de grande circulação
2. Partes: Companhia de Desenvolvimento Urbano da Região do Porto do Rio de Janeiro S/A – CDURP e Editora O Dia S/A
3. Razão: dispensa;
4. Fundamento: Art. 24, inciso II, da Lei 8.666/93;
5. Valor total da despesa: R$ 834,60 (oitocentos e trinta e quatro reais e sessenta centavos)
6. Autoridade: Sérgio Lopes Cabral;
7. Ratificador: Alberto Gomes Silva

DIRETORIA DE ADMINISTRAÇÃO E FINANÇAS
DESPACHOS DO DIRETOR
ERRATA

Publicação do dia 10/02/2014 Pag. 6
01/800.126/2012
Onde se lê:
6. Valor da despesa: R$ 330.000,00 (trezentos e trinta reais)
Leia-se
6. Valor da despesa: R$ 330.000,00 (trezentos e trinta mil reais)

IPLANRIO

Empresa Municipal de Informática S/A
Av Presidente Vargas, 3.131 - 12ºandar - Tel.:3971-1818/ Fax: 3971-1589
E-mail:iplanrio@pcrj.rj.gov.br

DESPACHOS DO PRESIDENTE
EXPEDIENTE DE 13/02/2014

PROCESSO Nº: 01/300.048/2014
Objeto: Ressarcimento de Despesa de Pessoal Requisitado ao SERPRO.
Partes: IPLANRIO e SERVIÇO FEDERAL DE PROCESSAMENTO DE DADOS - SERPRO
Fundamento: Não sujeito a Lei 8666/93
Razão: Não sujeito
Valor: R$ 1.049.000,00 (Hum milhão e quarenta e nove mil reais),
Autorização: VICTOR ZAJDHAFT, matr. Nº 45/620.889-7

EXPEDIENTE DE 13.02.2014

Com base na manifestação do Órgão Gerenciador do Sistema de Registro de Preços da IplanRio, autorizo os órgãos abaixo a fazerem uso dos preços registrados na Ata de Registro de Preços Nº 0004/2013, conforme disposto no Decreto Municipal nº 36.567, de 04 de dezembro de 2012.

FSS 006/2013 e FSS 022/2013 - Secretaria Municipal de Desenvolvimento Social - SMDS
Item 03 – 5.908 unidades Item 04 – 07 unidades Item 14 – 200 unidades Item 18 – 108 unidades Item 31 – 1.028 unidades Item 37 – 101 unidades

EXPEDIENTE DE 13/02/2014
Ressarcimento Imobiliário:

05/502.743/2003	Tânia Josué Ferreira

Defiro no processo piloto n.º 05/501.196/2014

DIRETORIA DE PREVIDÊNCIA E ASSISTÊNCIA
DESPACHOS DO DIRETOR
EXPEDIENTE DE 13/02/2014

Pensão
05/505.364/1994 – Marcelo Volpini Jansen Pereira
Indefiro o pedido de Liberação de Reservas à fl. 158.
05/500.471/1997 – Othoniel de Azevedo Bittencourt
Defiro o pedido de Extinção de Pensão à fl. 17.
05/504.745/2004 – José Bastos
Defiro o pedido de Extinção de Pensão à fl. 82.
05/507.576/2013 – Adelia Fernandes
Defiro o pedido de Reconsideração do Pagamento de Pensão às fls. 58 e 63.
05/500.136/2014 – Itiel Natalino Moreira
Defiro o pedido de Reconsideração do Pagamento de Pensão à fl. 26.
05/507.820/2013 – Oton José Medeiros Brito
Defiro o pedido de Pagamento de Pensão para Julia Lefebvre Brito Ferrari, a titulo precário à fl.19.
05/507.301/2010 – Hamilcar Pacheco Silveira
Indefiro o pedido de Extinção do Pagamento de Pensão à fl. 59.
Defiro o pedido de Reversão do Pagamento de Pensão de Selma Ferreira de Andrade Silveira para Agatha Raissa de Andrade Silveira à fl. 60.

Pecúlio Post Mortem
05/506.928/2013 – Robson Nascimento de Souza
Defiro o pedido de Reconsideração do Pagamento de Pecúlio Post Mortem à fl. 19.
05/500.424/2014 – Gilza Maria Silva Maia
Defiro o pedido de Pagamento de Pecúlio à fl. 14.
05/501.105/2014 – José Carlos Fajardo
05/501.118/2014 – Lucina Nunes Ramade
05/501.130/2014 – Francisco de Assis Araujo
05/501.139/2014 – Maria Rosa da Silva Moura
05/501.147/2014 – Diva Ferreira de Almeida
Defiro o pedido de Pagamento de Pecúlio à fl. 02.

Auxílio Funeral de Segurado
05/507.644/2013 – José Carlos de Souza
Defiro o pedido de pagamento de Auxilio Funeral de Segurado à fl. 11.
Indefiro o pedido de pagamento de Auxilio Funeral de Segurado à fl. 02.
05/501.104/2014 – José Carlos Farjado
05/501.107/2014 – Lucia Ferreira
Defiro o pedido de pagamento de Auxilio Funeral de Segurado à fl. 02.
05/500.519/2014 – Luiz de Oliveira
Defiro o pedido de pagamento de Auxilio Funeral de Segurado à fl. 12

Figure 1. Example of a public act published in the Official Gazettes.

2.1. Syntactic structure of the law

In the context of the syntactic structure of Laws, legal norms are formally expressed in the form of propositions and may appear in documents of the legal system in the form of statements from which facts are verified. For example, according to [Kelsen 2009] in page 81, the legal norm to which theft should be punished is often formulated by the legislator in the following proposition: *"Theft is punished with imprisonment;"*. On the other hand, a norm that grants the Head of State competence to conclude a treaty takes the form: *"The Head of State concludes an international treaty."*.

Basically, a Brazilian normative text follows a well-formed struct publication, a template. This struct, which can be seen, looks like a syntactic structure. An attempt to standardize legal instruments is provided in [Brasil 1998] and [Brasil 2001].

Thus, this normative gives rise to the structure:

1. **Preliminary Part**
   (a) *Epigraph* - The title of the legal norm.
   (b) *Ementa* - The purpose of the legal norm (object).
   (c) *Preamble* - Institution for the practice of the act and its legal basis.
2. **Normative Part**
   (a) *Substantive provisions.* Area intended for the provisions pertaining to the measures necessary for the implementation of the rules (substantive content).
3. **Final Part**
   (a) *Implementation* - Area intended for the provisions (dispositions) pertaining to the actions necessary for the implementation of substantive content legal norms.
   (b) *Transitory part* - Area intended for the transitory provisions (when exist).
   (c) *Validity* - The laws have a period of validity that can be determined or undetermined. In its syntactic form, the laws that determine a period ("vacancy") should use the wording [Brasil 2001]: *"[...] Esta lei entra em vigor após decorridos [the number of] dias de sua publicação oficial"*.
   (d) *Revocations* - when a law cites a "revocation". It must expressly list the laws or legal provisions revoked (when exist) [Brasil 2001].

For this research, what interests us is to syntactically analyze the Valid Legal Statements [2] (VLS) that are less complex than legal structures.

3. Related Works

We found four articles discussing information retrieval in a similar context. The three two deal with the processing of Official Gazettes, and the last one shows a survey between industry and academia examining the choice for each technique. We discuss them below.

[Rodríguez, M., Dantas Bezerra, B 2019] uses NLP techniques, based on [Friedman et al. 2013] to recognize Named Entities in appointment ordinance on Official Gazette. They use the resources available on the Natural Language Toolkit

[2] In the context of this work, a legal statement is any document that describes facts, provided by an individual who declares this information to be a candidate to validity. Example: an invoice, a paycheck, a normative act in the official gazette [Haeusler and Rademaker].

(NLTK) [Loper and Bird 2002] platform for steps of the tokenization process until the entity recognition. A limitation of this work is that the authors present a tool that recognizes only the names of public agents (public employees) in appointment ordinances. In this experiment, it was possible to observe an accuracy of 92% in the extraction of names. Moreover, it is not reported in detail which ML algorithms are used in this article.

[Junior et al. 2018] uses data mining techniques for information retrieval in the official gazette of the Government of Pernambuco, Brazil. This work reports the application of the Random Tree algorithm with a hit rate of 80%. The authors agree that "if the department wants an algorithm with better results, it is necessary to carry out a minimum standardization of the Official Gazette so that the extraction is more efficient". This highlights the complexity in the treatment of data contained in the official gazette. In this case, a study of other information retrieval strategies is necessary.

In [Pinto et al. 2022], uses regular expressions to retrieve information that has been triplicated to an RDF format and can be further queried *SPARQL* against a triple storage database such as AllegroGraph [Buil-Aranda et al. 2013]. Extracting acts from gazettes to a knowledge base is part of a broader project to create a KB for public documents in the context of electronic governance auditing and compliance.

Finally, [Chiticariu et al. 2013] makes a case for the importance of the use of rules-based extraction systems for industry. It presents a research plan with the potential to bridge the gap between academic research and industry practice.

4. Research Design and Methods

This section describes the data collection and analysis procedures used in this research report.

We defined that both techniques would have the same scenario for the experiment. Both would extract the information contained in the official gazettes. The choice of this document is justified in Section 2.

In the official gazette, we decided to treat a subset of human resources information, such as public jobs that do not require a contest for admission. Based on this scope, pattern extractors were developed using regular expressions applied to the grammar of the acts targeted in this research.

The first challenge was to capture official gazettes for a time period. For this activity, and others of this project, the language Python, version 3, was used as a *backend* of the production tool and generation of the RDF triples of public acts published in the official gazettes as already mentioned in Section 1.

In order to highlight the contributions, we will restrict ourselves here to the City of Rio de Janeiro `https://doweb.rio.rj.gov.br/`, Maceió `https://www.diariomunicipal.com.br/maceio/`, Palmas `http://diariooficial.palmas.to.gov.br/`, Recife `https://dome.recife.pe.gov.br/dome/` and Florianópolis `https://www.pmf.sc.gov.br/governo/index.php`. The scope was restricted to publications involving appointments and dismissals of public employees. These are similar to the one shown in Figure 1.

Padrão	Dispensar Servidor
Ato	*Dispensar[,|\s]*
Nome	(?P<nome>[A-ZÉÁÍÓÚÇÃÊÔÕÀÜ\s]+)
Matrícula	(?P<matricula>[0-9\./-]+)
Cargo Efetivo	(?P<cargoEfetivo>[A-ZÉÁÍÓÚÇÃÊÔÕÀÜa-záêéóíçãâôú\-\s]+)
Dia	(?P<dia>[0-9]+)
Mês	(?P<mes>[J | j]aneiro | [F | f]evereiro | [...] | [D | d]ezembro)
Ano	(?P<ano>[0-9]+)
Cargo Comissionado	(?P<cargo>[A-ZÉÁÍÓÚÇÃÊÔÕÀÜa-záêéóíçãâôú\-\s]+)
Símbolo	(?P<simbolo>[A-Z\-0-9\/\s]+)

Table 1. Regular expression pattern for the "Dispensar" act.

Some PDF files used in this research can be found in this repository: `https://github.com/fernandoantoniodantas/LREC2022/tree/main/PDFs/`.

4.1. Rule-based Approach

With these PDF files, the second step was to build the extractor of information contained in each official gazette. In this process, we used the Python RE library `https://docs.python.org/3/library/re.html` and a PEG grammar presented in the Figure 2.

As presented in Section 1, Regular Expression (RE) is a notation for specifying lexical patterns. Its syntactic construction is composed of atomic symbols (characters), union, concatenation, and Kleene closure of other regular expressions. Readers unfamiliar with the concept or terminology can refer to the book [Aho et al. 2006].

Thus, our objectives follow the formalization of text-based information. In this case, as we saw in Figure 1, the form of information presentation follows a model that we can define in terms of formal grammar, where the sections of the official gazette are well-defined. Consequently, enabling the extraction of legal entities by rules.

Parsing Expression Grammar (PEG) is a formalism that describes language recognizers, and it is a simpler alternative to presenting the syntactic formation rule (grammar) of certain languages. In the Figure 2, we present an example of formal grammar (PEG) that maps a public act of the Official Gazette.

Once the PEG is defined, we can map its formalism to a set of rules in regular expressions responsible for extracting information from the Official Gazette. We can see a small example of these patterns in Table 1.

4.2. Machine Learning-Based

For this task, was used Named-Entity Recognition (NER). It is a Natural Language Processing (NLP) widely used in information extraction, consisting of identifying names of specific entities in textual data according to predefined labels. We chose to use NER tasks because they have excellent tool support that speeds up preprocessing and model training.

This activity started with creating a training dataset of our named entity recognition model. Thus, this set is formed by small examples of legal declarations (extracted from the official gazette) annotated, indicating the words and their classifications

```
⟨publicAct⟩  ::=  ⟨top⟩⟨segment⟩
      ⟨top⟩  ::=  DECRETO "P" No.⟨port⟩ DE ⟨per⟩
      ⟨per⟩  ::=  ⟨day⟩ DE ⟨month⟩ DE ⟨year⟩
  ⟨segment⟩  ::=  ⟨segment1⟩⟨segment2⟩, símbolo⟨symbol⟩
 ⟨segment1⟩  ::=  RESOLVE ⟨act⟩⟨name⟩, matrícula⟨mat⟩,
 ⟨segment2⟩  ::=  ⟨compl⟩Cargo em Comissão de⟨publicfunc⟩
      ⟨act⟩  ::=  Nomear|Exonerar
     ⟨name⟩  ::=  [A-Z]+
     ⟨port⟩  ::=  [0-9]+
      ⟨day⟩  ::=  [0-9]+
    ⟨month⟩  ::=  [A-Z]+
     ⟨year⟩  ::=  [0-9]+
      ⟨mat⟩  ::=  [0-9/.-]+
    ⟨compl⟩  ::=  [A-Za-z0-9,-]+
⟨publicfunc⟩  ::=  [A-Z0-9]+
   ⟨symbol⟩  ::=  [A-Z0-9.-]+
```

Figure 2. Official Gazette grammar PEG.

within our legal domain. Generally, these annotations are created by human annotators or through automatic annotation tools and can be refined and reviewed by NLP experts to ensure the quality of the training data.

In previous work [Pinto et al. 2021], we annotated some examples of legal statements and submitted them to a Spacy [3] NER pipeline (like NER BERT [4]) for training an NER model [5]. A small sample of this annotated dataset is shown in Figure 4. It is important to note that this annotation process takes a lot of work, as it is necessary to indicate the indexes of occurrence of each entity in the text.

The Figure 3 depicted an example of the model with the classification of entities in the appointment for public office.

RESOLUÇÃO**RESOLUCAO** "P"
Nº 326 **NUMRESOLUCAO** DE 7 **DIARESOLUCAO** DE JANEIRO **MESRESOLUCAO** DE 2013 **ANORESOLUCAO**. O SECRETÁRIO CHEFE DA SECRETARIA MUNICIPAL DA CASA CIVIL, no uso das atribuições que lhe são conferidas pela legislação em vigor, RESOLVE Nomear **ACAO** ANA PAULA DE ALMEIDA NIGRO **SERVIDOR**, matrícula 60/199.218-9 **MATRICULA**, para exercer o Cargo em Comissão de Assessor III **CARGO**, símbolo DAS-07, código 029846, da XVI Administração Regional, da Coordenadoria Especial da Área de Planejamento 4, da Subsecretaria de Integração das Áreas de Planejamento, da Secretaria Municipal de Governo.

Figure 3. Entities classified with our trained model [Pinto et al. 2021].

So we discussed using the Watson Knowledge Studio (WKS) [6] annotation tool.

[3] A natural language processing library.

[4] A pre-training model of natural language processing.

[5] For practice: https://drive.google.com/drive/folders/16p8ejnFlu8WajJfWqbn3EG3fU8ahYHhV?usp=share_link

[6] https://www.ibm.com/cloud/watson-knowledge-studio

By having a text markup process, the annotation process has become more efficient.

”RESOLUÇÃO “P” Nº 3668 DE 5 DE AGOSTO DE 2020 O SECRETÁRIO CHEFE DA SECRETARIA MUNICIPAL DA CASA CIVIL, no uso das atribuições que lhe são conferidas pela legislação em vigor, RESOLVE Exonerar MELINA BRITO RODRIGUES DE SOUZA, matrícula 11/218.309-3, Agente de Administração, com validade a partir de 13 de maio de 2020, do Cargo em Comissão de Diretor IV, símbolo DAS-06, código 027526, da Divisão de Infraestrutura e Logística, da Coordenadoria de Gestão Administrativa, da Coordenadoria Geral de Atenção Primária da AP-2.1, da Subsecretaria de Promoção, Atenção Primária e Vigilância em Saúde, da Secretaria Municipal de Saúde. RETIFICAÇÃO D.O. RIO Nº 099 DE 29 DE JULHO DE 2020 ”,”entities”:[(0,13, ”RESOLUCAO”), (17,21, ”NUMRESOLUCAO”),(24,27, ”DIARESOLUCAO”),(30,36, ”MESRESOLUCAO”),(40,44, ”ANORESOLUCAO”), (183,191, ”ACAO”),(192,223, ”SERVIDOR”), (235,247, ”MATRICULA”), (343,353, ”CARGO”)],

Figure 4. An annotated instance of the training set.

For this task and with the defined scope shown in Section 4, 100 examples of publications involving appointments and dismissals of public employees were selected and submitted to a manual annotation process with IBM Watson Knowledge Studio (WKS). This low number of samples for training is justified because the WKS uses a pre-trained NER model.

Therefore, a set of 7 entities was defined: dia (day), mês (month), ano (year), nome (name), portaria (act id), cargo (public job name), and ato (act). As can be seen in the Figure 5, an example of an annotated publication on the WKS.

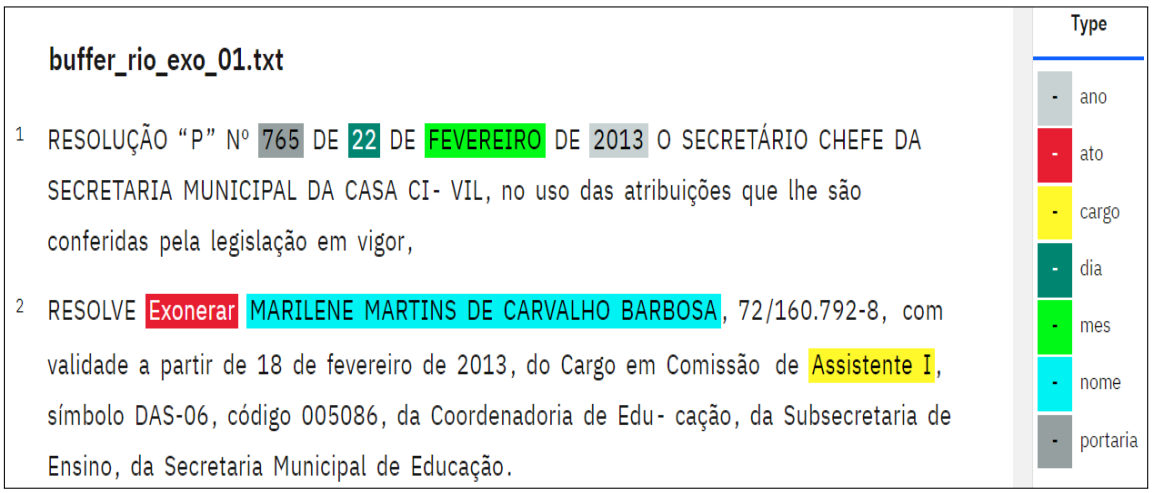

Figure 5. Example of annotated entities in the IBM Watson Knowledge Studio tool.

For readers outside of Brazil, the paragraph 1 and 2 of Figure 5 have translated into Figure 6:

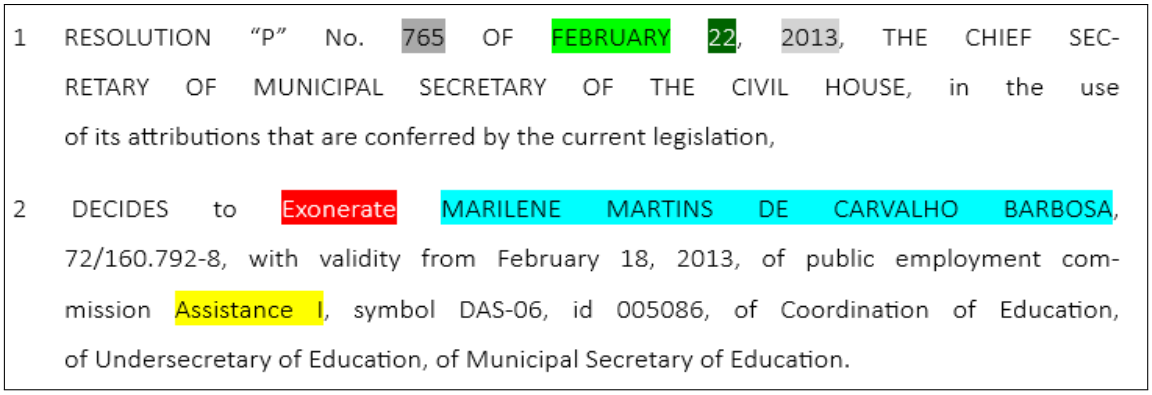

Figure 6. Example of annotated entities in the IBM Watson Knowledge Studio tool translated to English.

5. Results

In this section, we will present the results of the information extraction process based on rules and machine learning applied to the scenario described in Section 4.

An important aspect is the definition of the scope of this research. Traditionally, the industry makes its benchmarking of techniques based on these five criteria: Data Capacity, Training Speed, Model Precision, and Inference Speed. In this research, we adopted the **Model Precision** and a new task, the **Development Time** for both approaches.

The result of implementing the regular expressions rules was the extraction of information contained in the Official Gazette. To facilitate the process of analyzing this data, we created a tool that, upon recognizing the patterns, generates an audit file (Figure 7) with the recovered information.

In fact, our technique using regular expressions is adding one more type than the specified in the project, but which was not used as an analysis criterion by both techniques in this comparison.

All the examples used in the training set of our NLP machine were recognized by our rule-based system. The machine learning model obtained an accuracy of 0.99. The Table 2 summarizes these results.

Approach	Examples	Accuracy
ML-based	100	0.99
Rule-based	100	1.00

Table 2. Result between the two techniques.

The time of preparation of the experiment was also analyzed. We only compare the phase that precedes the execution of the extraction algorithms. Therefore, we did

```
(PUC-RIO/TECMF)   ::PROCESSAMENTO DO DIÁRIO:: ANO: 27 No.: 000025 TIPO: NORMAL      * RIO DE JANEIRO * ARQUIVO: 2055.PDF SEQ.: 0002

(PADRAO 1.23) RESOLUÇÃO   1054 02/04/2013 DESIGNAR  XXXXXXXXXX  LEDA MARIA DA FONSECA                          XX/XX/XXXX SECRETÁRIO II
(PADRAO 2.1)  RESOLUÇÕES  0825 19/04/2013 NOMEAR    10/137135-0  DIANA MARIA MENDES DOS SANTOS                 15/04/2013 ASSESSOR ADJUNTO
(PADRAO 2.3)  RESOLUÇÕES  0826 19/04/2013 DESIGNAR  10/192770-6  ELISABETE MORAES DOS SANTOS                   15/04/2013 ASSISTENTE II
(PADRAO 2.4)  RESOLUÇÕES  0827 19/04/2013 DESIGNAR  10/215889-7  RENATA GUIMARÃES BEZERRA                      XX/XX/XXXX COORDENADOR PEDAGÓGICO
(PADRAO 2.4)  RESOLUÇÕES  0828 19/04/2013 DESIGNAR  10/234939-7  CLÁUDIA DE OLIVEIRA ABREU                     XX/XX/XXXX COORDENADOR PEDAGÓGICO
(PADRAO 2.5)  RESOLUÇÕES  0822 19/04/2013 DISPENSAR 12/127995-9  DOROTÉA FROTA SANTANA                         23/04/2013 DIRETOR-ADJUNTO
(PADRAO 2.6)  RESOLUÇÕES  0819 19/04/2013 DISPENSAR 12/223456-5  SIMONE PEREIRA DE CASTRO VIEIRA               XX/XX/XXXX DIRETOR-ADJUNTO
(PADRAO 2.6)  RESOLUÇÕES  0823 19/04/2013 DISPENSAR 12/199981-2  ANA CLAUDIA POLYCARPO RIBEIRO                 XX/XX/XXXX DIRETOR-ADJUNTO
(PADRAO 2.8)  RESOLUÇÕES  0820 19/04/2013 EXONERAR  11/094145-0  MARIA MARTA DE BARROS PATRÍCIO                23/04/2013 DIRETOR IV
```

Figure 7. Audit file with information extracted from official gazettes.

not evaluate the machine processing time of both approaches. Four essential steps were observed: time to select examples, time to annotate the examples, develop rules, and review.

We can see in the Table 3 the Machine Learning-based approach for this task is more costly than the rule-based one. We emphasize that the preprocessing of the experiments was carried out by one person.

Approach	Task	Time
ML-Based	Selection of examples for annotation	360 min.
ML-Based	Annotation process	240 min.
ML-Based	Review	120 min.
Rules-Based	Rules development	240 min.
Rules-Based	Review	120 min.

Table 3. Preprocessing experiments time.

6. Conclusion and future works

In this research, we defined our scope in public acts available in the Official Gazettes. Two information extraction approaches were used, one based on Machine Learning and the other based on Rules.

Due to the characteristics of the Official Gazette, the use of regular expressions was presented as a simple and efficient solution. We observed that the ML preparation activities had a higher cost (time) than the preparation of rules in a regular expression.

Does this effort pay off? In our experiment, the process of extracting examples for annotation was a time-consuming activity. Even for a team carrying out this activity, the problem would be another issue: reliability and cost with workers.

In addition, the black-box aspect of the process inherent to statistical approaches makes it necessary, many times, to have a new treatment of the training set or adjustments to the model's hyperparameters. We are leading to more costs during extraction preparation. In this project, 12 hours were consumed to prepare the training set. It took 6 hours to prepare the extraction rules. The Table 3 presents the details of time measurement.

Based on the presented, we have to conclude that rule-based extraction techniques can easily replace this ML practice. In scenarios where documents have a well-defined grammar (some structure), rules-based information extraction systems still present themselves as a low-cost, simpler, and more efficient solutions.

One of the contributions of this project was the annotation of these 100 examples (made by specialist). This dataset is available for use in other experimental research. Our idea is to add new examples, types, and relationships to this base.

As a proposal for future work, we present two branches of research that we are working on. These researches involve the formalization of public acts in official gazettes.

The first is related to this paper, where given a well-formatted document, a translator could interpret its grammar in the Portuguese language and derive its extraction rules, automating the process of writing the patterns in regular expressions. The idea is to treat the grammar as a computational model and propose the induction of regular grammar on VLS in the Official Gazette.

The second is related to the legality or illegality of the acts of a public manager. We will try to capture the propositions present in the text of the law and its correct application in accordance with a legal act. According to [Kelsen 2009], the legal norm functions as an interpretation scheme, and an act of human conduct, the result of a normative interpretation, constitutes a **legal act that can be valid or invalid**.

In the legal context, our interest is in the search for legal acts that do not correspond to the laws that govern them. In short, we are not analyzing the normative interpretation (hermeneutics [7]) but the application of the norm (validity or invalidity of acts). Otherwise, in a formal semantic context, our interest in this research is the (semi)automatic search for real examples that prove the legal invalidity. In other words, we are looking for public acts (appointments of public agents, undue receipts, etc.) that are non-compliance with legislation.

The main artifacts produced in this project can be found in the following repository: `https://github.com/fernandoantoniodantas/LREC2022`

7. Acknowledgement

In particular, we would like to thank Alexandre Rademaker (FGV/IBM) for their valuable suggestions. Without their support, it would not have been possible for us to complete this project.

References

Aho, A. V., Lam, M. S., Sethi, R., and Ullman, J. D. (2006). *Compilers: Principles, Techniques, and Tools (2nd Edition)*. Addison-Wesley Longman Publishing Co., Inc., USA.

Brasil (1998). Lei complementar nº 95, de 26 de fevereiro de 1998. *Diário Oficial [da] República Federativa do Brasil*.

[7] In the context of law, hermeneutics is the philosophical science of law focused on the interpretation of its objects through models and interpretive structures.

Brasil (2001). Lei complementar nº 107, de 26 de abril de 2001. *Diário Oficial [da] República Federativa do Brasil*.

Buil-Aranda, C., Hogan, A., Umbrich, J., and Vandenbussche, P.-Y. (2013). Sparql web-querying infrastructure: Ready for action? In Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J. X., Aroyo, L., Noy, N., Welty, C., and Janowicz, K., editors, *The Semantic Web – ISWC 2013*, pages 277–293, Berlin, Heidelberg. Springer Berlin Heidelberg.

Chiticariu, L., Li, Y., and Reiss, F. R. (2013). Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 827–832, Seattle, Washington, USA. Association for Computational Linguistics.

Constantino, K., Cruz, V. A. L., Zucheratto, O. M. M., França, C., Carvalho, M., Silva, T. H. P., Laender, A. H. F., and Gonçalves, M. A. Segmentação e classificação semântica de trechos de diários oficiais usando aprendizado ativo. In *Anais do XXXVII Simpósio Brasileiro de Banco de Dados (SBBD 2022)*, pages 304–316. Sociedade Brasileira de Computação - SBC.

Friedman, C., Rindflesch, T. C., and Corn, M. (2013). Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *Journal of Biomedical Informatics*, 46(5):765–773.

Haeusler, E. H. and Rademaker, A. *On how kelsenian jurisprudence and intuitionistic logic help to avoid contrary-to-duty paradoxes in legal ontologies*, pages 44–59. Lógica no Avião.

Junior, R., Melo, W., Fagundes, R., and Maciel, A. (2018). Extração de informação e mineração de dados no diário oficial de pernambuco. *Revista de Engenharia e Pesquisa Aplicada*, 3.

Kelsen, H. (2009). *Teoria pura do direito*. WMF Martins Fontes, São Paulo, 8. ed. edition. ISBN: 83-336-0836-5.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Mohit, B. (2014). Named entity recognition. In Zitouni, I., editor, *Natural Language Processing of Semitic Languages*, pages 221–245. Springer Berlin Heidelberg.

Pinto, F. A., Haeusler, E., and Lifschitz, S. (2021). Transparência pública automatizada a partir da gramática do diário oficial. In *Anais do IX Workshop de Computação Aplicada em Governo Eletrônico*, pages 59–70, Porto Alegre, RS, Brasil. SBC.

Pinto, F. A. D. G., Lifschitz, S., and Haeusler, E. H. (2022). A graph knowledge-base for auditing human resources public management. In *Anais do X Workshop de Computação Aplicada em Governo Eletrônico*, Porto Alegre, RS, Brasil. SBC.

Rodríguez, M., Dantas Bezerra, B (2019). Processamento de linguagem natural para reconhecimento de entidades nomeadas em textos jurídicos de atos administrativos (portarias). *Revista de Engenharia e Pesquisa Aplicada*, 5(1):67–77.