# Reconhecimento de Entidades Nomeadas em Textos Legais

**Autor:** Gabriel Dalforno Silvestre
**Orientador:** André C. P. L. F. de Carvalho

Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

**Abstract**

O avanço das técnicas de inteligência articifical, em especial na área de processamento de linguagem natural, nos últimos anos tem levado organizações a automatizarem grande parte de seus processos. Neste projeto vamos desenvolver um sistema inteligente capaz de reconhecer entidades nomeadas em documentos fornecidos pela Câmara dos Deputados do Brasil, desde a análise dos dados até a avaliação das técnicas especializadas nesta tarefa. Os resultados mostraram que o CRF é o *baseline* mais apto para o trabalho e que explorar a heterogeinidade dos documentos melhora a capacidade preditiva deste modelo principalmente na modelagem das entidades que são mais importantes em cada grupo.

***Palavras-Chave***— Processamento de Linguagem Natural, Reconhecimento de Entidades Nomeadas, Aprendizado de Máquina, Análise de Textos Legais

# 1 Introdução

A grande quantidade de dados que tem sido gerada nas últimas décadas tem posto a prova várias tarefas que até então eram realizadas de forma manual por indivíduos com *expertise* no domínio das informações coletadas. Este cenário se reflete em várias áreas do conhecimento, como na medicina, engenharia, direito, etc. Porém, com o avanço das pesquisas nas áreas de Ciência de Dados e Aprendizado de Máquina, foi possível realizar muitas dessas tarefas de forma automática por um sistema inteligente capaz de aprender padrões a partir de um conjunto de dados rotulado [1]. No âmbito legal a situação não é muito diferente, recentemente vimos o crescente interesse em aplicações de técnicas de Inteligência Artificial, principalmente da subárea de Processamento de Linguagem Natural (PLN), para auxiliar os profissionais da área em trabalhos diversos [2, 3]. Dentre estes podemos destacar o trabalho de sumarização de documentos em termos de palavras-chave que são de suma importância para aceleração do processo de agrupamento de documentos, por exemplo, que no contexto do PLN é chamado de Reconhecimento de Entidades Nomeadas (REN) [4].

Alguns trabalhos dedicaram-se a tarefa de REN em textos legais em lingua portuguesa, alguns focados no desenvolvimento de conjuntos de dados, por exemplo, [5] que apresentou o "LeNER-Br", um conjunto de dados com um total de 70 documentos legais de tribunais judiciais brasileiros e [6] que apresentou o "DOU-Corpus", um conjunto de dados que contém 470 documentos do Diário Oficial da União (DOU). E outros focados na melhoria do desempenho de modelos especializados na tarefa de REN, dentre estes vale ressaltar [7] que desenvolveu uma técnica baseada em aprendizado profundo para melhoria da qualidade das predições de modelos clássicos em dados da Justiça do Trabalho Brasileira e [8] que estudou a aplicação de uma abordagem híbrida para REN em textos da Coleção Dourada do HAREM.

O objetivo deste projeto é desenvolver um modelo inteligente para REN em documentos oriundos da Câmara dos Deputados do Brasil. Na Seção 2, iremos apresentar o conjunto de dados desenvolvido para construção do modelo, na Seção 3, iremos executar uma análise estatística descritiva do Corpora. Já na Seção 4, será apresentada uma introdução teórica à tarefa de REN assim como a apresentação dos modelos e métricas de avaliação que serão utilizadas nos experimentos apresentados na Seção 5.

# 2 UlyssesNER-Br Corpora

O UlyssesNER-Br Corpora é o nome dado ao conjunto de dados que contém os documentos cedidos pela Câmara dos Deputados do Brasil e que foram rotulados pela nossa equipe. Há 3 tipos de documentos: projetos de lei, contidos no conjunto de dados apelidado de PL-corpus; solicitações de trabalho que compõe o ST-corpus e os comentários das enquetes que foi apelidado de C-corpus.

Um conjunto de 17 entidades nomeadas, julgadas como sendo as mais pertinentes para o desenvolvimento do sistema, foram extraídas de todos os documentos do corpora. Estas abrangem desde referências a leis e organizações governamentais até pessoas, localizações, etc. A Tabela 1 mostra todas as entidades nomeadas que podem ser encontradas no UlyssesNER-Br, existem dois níves de agregação, o mais alto das categorias e o mais baixo dos tipos.

| **Categoria** | **Tipo** | **Exemplos** |
|---|---|---|
| **ORGANIZAÇÃO** | ORGgovernamental | Capes, Unicamp, Correios, Ministério da Economia |
| | ORGnaogovernamental | FEDEX, GOL, AZUL, UPS |
| | ORGpartido | PT, PSOL, PSDB |
| **LOCAL** | LOCALconcreto | Niterói-RJ, Brumadinho, Mariana |
| | LOCALvirtual | Google Maps, Whatsapp, Facebook |
| **PESSOA** | PESSOAindividual | Presidente Jorge Sampaio, Iron Tyson, Carlos Guerra |
| | PESSOAcargo | técnica de enfermagem, advogada, deputado |
| | PESSOAgrupocargo | médicos, bombeiros, engenheiros |
| | PESSOAgrupoind | Setúbal, Vilella, Moreira Sales |
| **FUNDAMENTO** | FUNDlei | art. 37, XV da CF/88; art. 41, §3º, e 48 da Lei 8.112/91 |
| | FUNDprojetodelei | PL 805/2020; PEC 187/2016 |
| | FUNDapelido | Lei das Licitações, Código Civil, ECA |
| | FUNDsolicitacaotrabalho | Solicitação de Trabalho nº 3543/2019 |
| **DATA** | ———— | 2019, julho de 2007 |
| **EVENTO** | ———— | eleições de 2018, copa do mundo |
| **PRODUTO DE LEI** | PRODUTOsistema | SUS, SUAS |
| | PRODUTOprograma | Minha Casa Minha Vida, Bolsa Família |
| | PRODUTOoutros | dundo partidário, auxílio emergencial |

Table 1: Entidades nomeadas encontradas no Ulysses-NER-Br Corpora.

## 2.1 Processo de Anotação

O processo de anotação destas entidades foi feito em 3 fases. A Fase 0 foi a fase de treinamento dos anotadores, onde a anotação de um total de 5 solicitações de trabalho e 4 projetos de lei foi feita em conjunto com o objetivo de familizar-se com a estrutura de documentos do domínio legal.

Uma vez terminado o treinamento dos anotadores, passou-se para a Fase 1, em que duas equipes de duas pessoas ficaram responsáveis pela anotação

de 5 projetos de lei por dia e uma equipe de 2 pessoas anotando 50 solicitações de trabalho por dia durante um período de 6 dias. Além disso, cada equipe teve a disposição um curador que avaliava a similaridade entre as anotações da equipe em cada documento usando a estatística do Cohen's *kappa* [11]. No fim de cada dia, cada equipe recebia o *feedback* de seu respectivo curador com o objetivo de manter a similaridade entre as anotações mais alta possível. Espera-se que o valor de *kappa* fique em torno de 0.72 à 0.85, de acordo com [12].

Finalmente, na Fase 2, o processo é repetido com a diferença de que a curadoria é feita somente no fim da semana ao invés de diariamente como na Fase 1. Agora, as equipes responsáveis pelos projetos de lei anotam 10 por dia, e a equipe das solicitações de trabalho anotam 100 por dia.

No caso dos comentários, a anotação iniciou-se na Fase 2, já que os anotadores já tinham acumulado experiência das Fases 0 e 1 aplicadas nos projetos de lei e solicitações de trabalho. Todo o trabalho de anotação foi carregado utilizando a ferramenta INCEpTION [13].

Ao final, o corpora contou com um total de 250 projetos de lei, 790 solicitações de trabalho e 967 comentários devidamente rotulados e prontos para serem analisados e utilizados na construção do modelo REN.

# 3 Análise Exploratória dos Documentos

Nesta seção, o nível de significância das estimações estatísticas estará fixado em $\alpha = 5\%$ e a estimação dos erros serão baseadas numa abordagem conservadora. Além disso, vamos analisar cada um dos corpus separadamente no que diz respeito a propriedades estatísticas dos documentos e das entidades nomeadas presentes naqueles.

## 3.1 Projetos de Lei

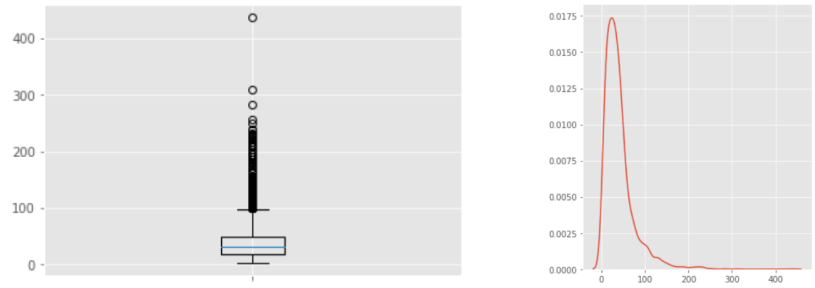
No PL-corpus há um total de 9526 sentenças, porém somente 34% deste total corresponde a sentenças com significado semântico. Devido a estrutura dos arquivos, a maioria das sentenças possuem um único *token*, como uma vírgula ou um ponto. A estimação pontual da média de sentenças por projeto de lei resultou em 63.50 cujo intervalo de confiança é (46.15, 80.86), a grande amplitude do último é explicada pela presença de sentenças de comprimento unitário, como discutido. Os documentos que possuem um grande número de

sentenças deste tipo foram classificados como *outliers* como ilustra a Figura 1.


(a) (b)

Figure 1: Distribuição do número de sentenças por projeto de lei no PL-corpus.

Já na estimação da média do número de *tokens* por sentença obteve-se 40.43 com intervalo de confiança de (39.23, 41.64), neste caso, as sentenças com um único *token* foram removidas devido a sensibilidade da média amostral. A Figura 2 mostra que, apesar da amplitude da amostra dos comprimentos de sentenças ser maior que 400, espera-se que a maioria dos comprimentos estejam limitados por 200, mesmo para *outliers* não tão severos.




(a) (b)

Figure 2: Distribuição do número de *tokens* por sentença no PL-corpus.

Finalmente, há um total de 138741 *tokens* no PL-corpus, dentre estes, 11833 são únicos. Nestas condições, foi calculado o erro ao tentar estimar a probabilidade de aparição de uma entidade nomeada no corpus que foi de 0.26%. A estimação pontual dos valores está presente na Tabela 2. Como

pode-se observar, um pouco mais de 10% dos *tokens* foram anotados como pertencentes à alguma das entidades nomeadas, com destaque para as entidades FUNDlei, seguida por ORGgovernamental e PRODUTOoutros. Por outro lado, outras entidades raramente aparecem neste corpus como ORGpartido e EVENTO, mais ainda, as entidades FUNDsolicitacaotrabalho e PESSOAgrupoind não aparecem nestes documentos. A Figura 3 ilustra o que mostra a tabela em termos do número absoluto de aparições.

| **Entidade Nomeada** | $\hat{p}$ |
|---|---|
| ORGpartido | 0.03% |
| EVENTO | 0.05% |
| FUNDprojetodelei | 0.08% |
| PRODUTOsistema | 0.08% |
| PRODUTOprograma | 0.17% |
| PESSOAgrupocargo | 0.18% |
| PESSOAcargo | 0.28% |
| ORGnaogovernamental | 0.29% |
| LOCALvirtual | 0.35% |
| FUNDapelido | 0.49% |
| PESSOAindividual | 0.63% |
| LOCALconcreto | 0.66% |
| DATA | 0.71% |
| PRODUTOoutros | 0.87% |
| ORGgovernamental | 1.04% |
| FUNDlei | 4.28% |
| O | 89.82% |

Table 2: Estimação pontual da proporção de cada entidade nomeada no PL-corpus.

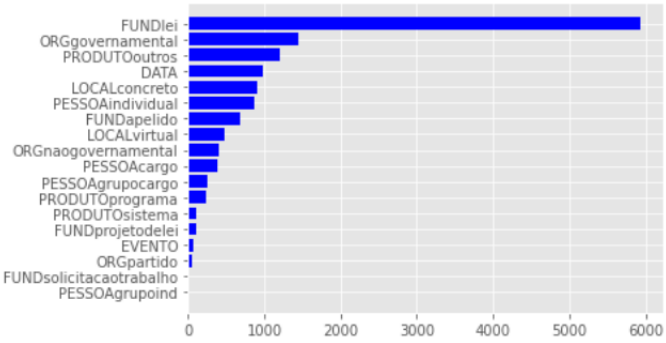



Figure 3: Número absoluto de aparições de cada entidade nomeada no PL-corpus.

## 3.2 Solicitações de Trabalho

Cada solicitação de trabalho no ST-corpus corresponde a uma única sentença, neste caso, temos um total de 790 sentenças no conjunto de dados. A estimação pontual da média do número de *tokens* por sentença resultou em 98.02 com intervalo de confiança de (91.75, 104.30). Na Figura 4 podemos ver que a distribuição do comprimento das sentenças das solicitações de trabalho é similar ao das sentenças dos projetos de lei. Em ambos os casos espera-se que as observações, com excessão de *outliers* severos, não tenham mais do que metade da amplitude total da amostra como limitante superior para o número de *tokens*.


(a) (b)

Figure 4: Distribuição do número de *tokens* por sentença no PL-corpus.

Há um total de 77441 *tokens* no corpus, neste caso, o vocabulário tem 11053 *tokens*. Como este corpus é menor quando comparado ao PL-corpus, tem-se que o erro ao estimar as probabilidades de aparição das entidades nomeadas é de 0.35%. As estimações pontuais destas estão mostradas na Tabela 3 e os valores absolutos na Figura 5. Em ambas, podemos ver que, no que diz respeito as proporções de aparição de cada entidade nomeada, há uma grande similaridade entre as solicitações de trabalho e os projetos de lei. As diferenças mais significativas estão nas entidades DATA e LOCALvirtual, a primeira é muito comum em projetos de lei, porém em solicitações de trabalho é mais escassa, já com a segunda acontece justamente o oposto. Outra diferenças é presença da entidade FUNDsolicitacaotrabalho neste conjunto de dados que não aparecia no PL-corpus.

| **Entidade Nomeada** | $\hat{p}$ |
|---|---|
| ORGpartido | 0.01% |
| EVENTO | 0.10% |
| PRODUTOsistema | 0.11% |
| PRODUTOprograma | 0.20% |
| PESSOAgrupocargo | 0.21% |
| FUNDsolicitacaotrabalho | 0.30% |
| ORGnaogovernamental | 0.33% |
| DATA | 0.37% |
| FUNDprojetodelei | 0.39% |
| PESSOAindividual | 0.40% |
| LOCALconcreto | 0.49% |
| PESSOAcargo | 0.53% |
| FUNDapelido | 0.63% |
| LOCALvirtual | 0.77% |
| PRODUTOoutros | 0.97% |
| ORGgovernamental | 1.13% |
| FUNDlei | 4.33% |
| O | 88.72% |

Table 3: Estimação pontual da proporção de cada entidade nomeada no ST-corpus.

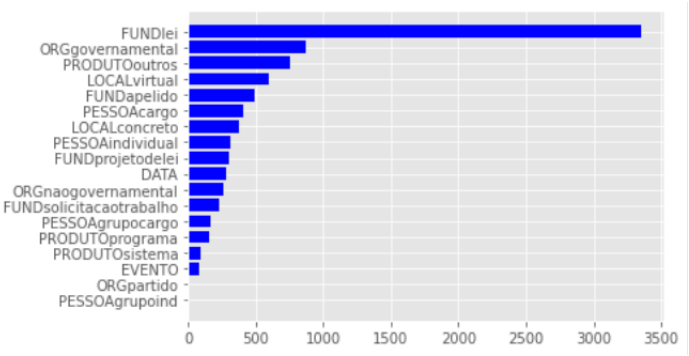



Figure 5: Número absoluto de aparições de cada entidade nomeada no ST-corpus.

## 3.3 Comentários

Cada comentário também é composto por uma única sentença, há um total de 967. Aqui, a estimação pontual da média de *tokens* por comentário foi de 92.02 cujo intervalo de confiança é dado por (91.56, 92.48). A Figura 6 ilustra o comportamento da distribuição da amostra dos comprimentos dos

comentários, vemos que ela segue uma distribuição muita próxima de uma normal, além disso, quando comparado aos outros corpus tem-se que este possue um número muito menor de *outliers*.


(a) (b)

Figure 6: Distribuição do número de *tokens* por sentença no C-corpus.

O C-corpus é composto por 88982 *tokens* em que *11837* são únicos. A Tabela 4 contém as estimações pontuais das proporções cujo erro associado é de 0.33% e Figura 7 mostra as quantidades absolutas.

| **Entidade Nomeada** | $\hat{p}$ |
|---|---|
| PRODUTOprograma | 0.01% |
| PESSOAgrupoind | 0.01% |
| PRODUTOsistema | 0.01% |
| ORGpartido | 0.01% |
| LOCALvirtual | 0.02% |
| FUNDprojetodelei | 0.02% |
| DATA | 0.05% |
| ORGnaogovernamental | 0.06% |
| PESSOAindividual | 0.15% |
| EVENTO | 0.17% |
| LOCALconcreto | 0.18% |
| FUNDapelido | 0.20% |
| FUNDlei | 0.23% |
| PESSOAcargo | 0.31% |
| ORGgovernamental | 0.42% |
| PRODUTOoutros | 0.68% |
| PESSOAgrupocargo | 0.74% |
| O | 96.72% |

Table 4: Estimação pontual da proporção de cada entidade nomeada no C-corpus.

Quando comparado aos projetos de lei e solicitações de trabalho, tem-se que os comentários exibem um comportamento distinto no que diz respeito

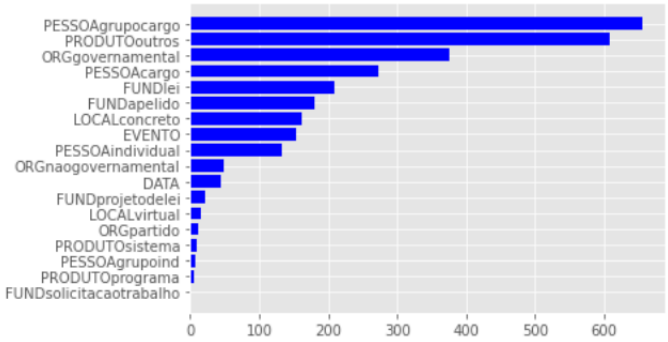



Figure 7: Número absoluto de aparições de cada entidade nomeada no C-corpus.

as estatísticas sobre as entidades nomeadas. Primeiramente, a probabilidade de um *token* carregar uma entidade é de menos de 5%, quase metade das chances nos outros dois corpus. Além disso, vemos que a entidade PESSOA-grupocargo domina os comentários e FUNDlei que era a mais importante nos projetos de lei e solicitações de trabalho aparece em quarto lugar no C-corpus. Como no PL-corpus, a entidade FUNDsolicitacaotrabalho também não aparece pareada com nenhum *token*.

# 4 Reconhecimento de Entidades Nomeadas

O REN não é nada mais do que uma tarefa de Aprendizado de Máquina Supervisionado [9] , onde tem-se definido um espaço de entradas $\mathcal{X}$, um espaço de saídas $\mathcal{Y}$ e uma distribuição de probabilidade $\mu$ definida no produto cartesiano $\mathcal{X} \times \mathcal{Y}$. Nestas condições, temos disponível um conjunto de dados rotulado $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}; i = 1, \ldots, n\}$ e queremos mapear $\mathcal{D}$ numa função $f : \mathcal{X} \to \mathcal{Y}$ tal que $f(x) \approx y, \forall (x, y) \sim \mu$. No caso específico do REN, cada $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n_i)})$ é uma sentença, e $y_i = (y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(n_i)})$ é a sequência de *tags* (ou entidades) associadas à $x_i$ para todo $i = 1, \ldots, n$.

Dentre as abordagens para aprendizado da função $f$ existem duas que são muita utilizadas na prática, a abordagem generativa e a abordagem discriminativa.

## 4.1 Modelos Generativos

Suponha que temos um vocabulário finito $\mathcal{V}$ e um conjunto finito de *tags* $\mathcal{K}$, defina $\mathcal{S}$ como o conjunto de todos os pares sentença/*tag* como em 1.

$$\mathcal{S} = \{(x_1, \ldots, x_n, y_1, \ldots, y_n) : x_i \in \mathcal{V}, y_i \in \mathcal{K}, i = 1, \ldots, n\} \tag{1}$$

Um modelo generativo é uma função $p$ satisfazendo as seguintes propriedades:

- $p(x_1, \ldots, x_n, y_1, \ldots, y_n) > 0, \forall (x_1, \ldots, x_n, y_1, \ldots, y_n) \in \mathcal{S}$
- $\sum_{(x_1,\ldots,x_n,y_1,\ldots,y_n)\in S} p(x_1, \ldots, x_n, y_1, \ldots, y_n) = 1$

Nessas condições, definimos o *tagger* $f$ como em 2.

$$f(x_1, \ldots, x_n) = \arg \max_{y_1,\ldots,y_n} p(x_1, \ldots, x_n, y_1, \ldots, y_n) \tag{2}$$

Um dos exemplos mais comuns deste tipo de modelo são os *Hidden Marvok Models* (HMM). Se considerarmos uma suposição de Markov de ordem 2, tem-se que $p$ se reduz a Equação 3.

$$p(x_1, \ldots, x_n, y_1, \ldots, y_n) = \prod_{i=1}^{n+1} q(y_i|y_{i-1}, y_{i-2}) \prod_{i=1}^{n} e(x_i|y_i) \tag{3}$$

Denotando por $c(u, v, s)$ o número de vezes que a sequência $(u, v, s)$ aparece no conjunto de treinamento, $c(u, v)$ o número de aparições do par $(u, v)$, $c(s \rightsquigarrow x)$ o número de vezes que a *tag* $s$ aparece pareada com o token $x$ e $c(x)$ o número de aparições do token $x$ define-se $q$ e $e$ como em 4.

$$\begin{aligned} q(s|u, v) &= \frac{c(u, v, s)}{c(u, v)} \\ e(x|s) &= \frac{c(s \rightsquigarrow x)}{c(x)} \end{aligned} \tag{4}$$

Uma implementação do HMM está disponível em [14].

## 4.2 Modelos Discriminativos

Os modelos discriminativos são definidos da mesma forma que os generativos com a excessão de que a função $p(x_1, \ldots, x_n, y_1, \ldots, y_n)$, i.e., a probabilidade conjunta, é substituída pela probabilidade condicional $p(y_1, \ldots, y_n | x_1, \ldots, x_n)$. Dentre aqueles, um dos mais utilizados é o *Conditional Random Fields* (CRF).

Pelo bem da simplicidade, denotemos $\underline{x} = (x_1, \ldots, x_n)$ e $\underline{y} = (y_1, \ldots, y_n)$. A ideia do CRF está em transformar cada par $(\underline{x}, \underline{y}) \in \mathcal{S}$ num vetor de atributos $\Phi(\underline{x}, \underline{y}) \in \mathbb{R}^d$ e assumir que a função $p(\underline{y}|\underline{x})$ é parametrizada como em 5.

$$p(\underline{y}|\underline{x}; w) = \frac{\exp(\langle w, \Phi(\underline{x}, \underline{y})\rangle)}{\sum_{\underline{y'} \in \mathcal{S}} \exp(\langle w, \Phi(\underline{x}, \underline{y'})\rangle)} \tag{5}$$

Logo, dado um conjunto de dados rotulado $\mathcal{D} = \{(\underline{x}^i, \underline{y}^i), i = 1, \ldots, m\}$ basta encontrar $w^* \in \mathbb{R}^d$ tal que a Equação 6 é satisfeita.

$$w^* = \arg\max_{w \in \mathbb{R}^d} \sum_{i=1}^{m} \log p(\underline{y}^i | \underline{x}^i; w) - R(w) \tag{6}$$

Em que o primeiro termo refere-se à máxima verossimilhança e $R : \mathbb{R}^d \to \mathbb{R}$ são os termos da regularização sobre o parâmetro $w$. Este pode ser estimado usando algum algoritmo de otimização não-linear como o *Limited-memory BFGS* [10]. Uma implementação do modelo CRF pode ser encontrada em [15].

## 4.3 Métricas de Avaliação

Como o REN cai na categoria de aprendizado supervisionado, dado um *tagger* $f$, define-se uma função $\mathcal{E} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ que nos traz informações à respeito da qualidade das predições do *tagger* $f$ em termos de um número real. A aplicação $\mathcal{E}$ é chamada de métrica de avaliação. Esta é muita usada quando deseja-se comparar a capacidade preditiva de vários *taggers*.

No caso específico do REN, as métricas mais utilizadas na literatura são: Precisão, *Recall* e *F1-score*. Denote por $(y_1, \ldots, y_n) \in \mathcal{Y}$ as *tags* de uma sentença $(x_1, \ldots, x_n) \in \mathcal{X}$, $(\hat{y_1}, \ldots, \hat{y_n}) \in \mathcal{Y}$ as *tags* que foram preditas por um *tagger* $f$, e $\mathbb{I}(A)$ como sendo 1 se $A$ for verdadeiro e 0, caso contrário.

Fixada uma *tag* $s \in \mathcal{K}$, tem-se que a Precisão do *tagger* $f$ em prever a *tag* $s$, é dada pela Equação 7. Esta métrica mede as chances do modelo acertar a

predição da *tag* $s$ considerando todas as vezes que recomendou esta *tag* para um *token* no conjunto de testagem.

$$\frac{\sum_{i=1}^{n} \mathbb{I}(y_i = s \cap \hat{y}_i = s)}{\sum_{i=1}^{n} \mathbb{I}(y_i = s \cap \hat{y}_i = s) + \sum_{i=1}^{n} \mathbb{I}(y_i \neq s \cap \hat{y}_i = s)} \tag{7}$$

Já o *Recall* com respeito a *tag* $s$ está ilustrado na Equação 8, note que a fórmula é similar a Precisão, porém neste caso consideramos a porcentagem do modelo acertar a predição de $s$ a partir de todas as vezes que esta *tag* aparece nas sentenças escolhidas para avaliação do modelo.

$$\frac{\sum_{i=1}^{n} \mathbb{I}(y_i = s \cap \hat{y}_i = s)}{\sum_{i=1}^{n} \mathbb{I}(y_i = s \cap \hat{y}_i = s) + \sum_{i=1}^{n} \mathbb{I}(y_i = s \cap \hat{y}_i \neq s)} \tag{8}$$

Finalmente, tem-se o *F1-score* da *tag* $s$ que é dado pela Equação 9. Este não é nada além da média harmônica entre a Precisão e o *Recall*. Todas as 3 métricas estão limitadas entre 0 e 1, onde valores próximos ao último indicam uma performance mais satisfatória.

$$\frac{2}{\frac{1}{Precisao} + \frac{1}{Recall}} \tag{9}$$

# 5 Experimentos

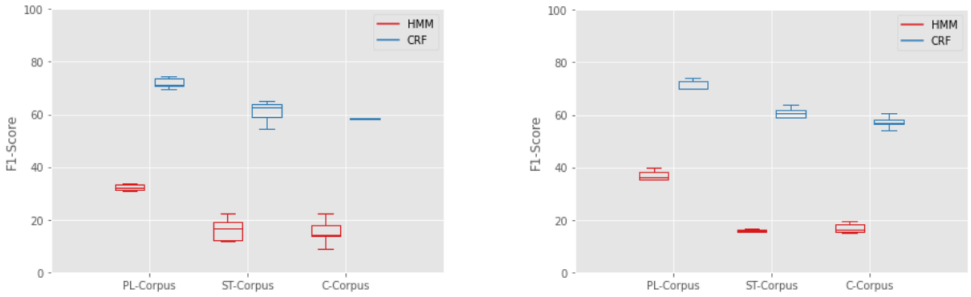
Os experimentos foram divididos em duas etapas. Na primeira, estamos preocupados em descobrir qual das duas abordagens NER apresentadas na seção anterior tem mais chance de sucesso na modelagem dos documentos do corpora. A partir desta informação, iremos testar estratégias para tornar o modelo escolhido mais robusto e confiável para aplicações reais.

Com respeito a primeira etapa, vamos aplicar a seguinte sequência de passos para os 3 corpus considerando tanto o nível de agregação das categorias quanto o dos tipos.

1. Divisão do conjunto de dados em conjunto de treinamento, contemplando 75% do total de documentos, e conjunto de teste, os 25% restantes de forma aleatória;

2. Aplicação da técnica de validação cruzada com $k = 5$ no conjunto de treinamento, este último é um agregado das sentenças de todos os seus documentos;

3. A abordagem que mostrar um *F1-score* maior será escolhida como a mais confiável.




(a) Categorias (b) Tipos

Figure 8: Resultados dos modelos CRF e HMM na validação cruzada.

Como podemos ver na Figura 8, o modelo CRF foi claramente superior ao HMM, obtendo uma performance aproximadamente duas vezes melhor em todos os casos considerados. Assim, conclui-se que neste corpora, a parametrização da distribuição de probabilidade em termos de uma mapa de atributos que traduzem as relações entre os *tokens* é menos suscetível a erros do que a estimação direta da distribuição de probabilidade conjunta sob a suposição de Markov. Também vale a pena ressaltar que aplicando o teste *t-student* para comparação de duas médias amostrais com variâncias distintas, concluiu-se que não há diferença estatística entre a performance dos modelos considerando o nível de agregação tipos ou categorias em absolutamente todos os casos.

Desde já vemos que dentre os 3 conjuntos de dados considerados, o PL-corpus é o mais simples de se modelar, seguido pelo ST-corpus e logo atrás o C-corpus. Isto decorre principalmente, do fato de que a quantidade de sentenças com significado semântico disponíveis para treinamento dos modelos no conjunto de dados dos projetos de lei é essencialmente 3 vezes maior que no conjunto dos comentários e 5 vezes maior que nas solicitações de trabalho. Além disso, a linguagem empregada nos projetos de lei e nas solicitações de trabalho é mais formal que a dos comentários, tornando o aprendizado das relações semânticas entre *tokens* e *tags* no último ligeiramente mais complicadas mesmo gozando de um número relativamente grande de sentenças.

De acordo com essas informações, decidiu-se testar a hipótese de que treinando o modelo CRF com sentenças de mais de um corpus de uma só

vez trás resultados mais favoráveis do que modelar cada corpus de maneira independente. Para isso, treinou-se um modelo CRF em todas as possíveis combinações dos 3 corpus e avaliou-se suas performances em cada um dos conjuntos de dados. O treinamento foi feito considerando a mesma partição feita na primeira etapa, e.g., se um modelo CRF terá seus parâmetros estimados usando sentenças do PL-Corpus e do ST-corpus, então em seu conjunto de treinamento haverá 75% das sentenças dos projetos de lei e 75% das sentenças das solicitações de trabalho. Porém, os conjuntos de teste são os mesmos para todos os modelos, desta forma garantimos uma comparação justa e isolamos os efeitos do treinamento em múltiplos conjuntos de dados em tempo de avaliação.

| Nível de Agregação | Conjunto de Treinamento | Conjunto de Teste | Precisão | Recall | F1-score |
|---|---|---|---|---|---|
| **Categorias** | PL-Corpus | **PL-Corpus** | 74.57 | 58.84 | 65.78 |
| | ST-Corpus | | 38.69 | 34.18 | 36.30 |
| | C-Corpus | | 7.51 | 9.27 | 8.30 |
| | **PL-Corpus + ST-Corpus** | | **76.03** | **61.22** | **67.83** |
| | PL-Corpus + C-Corpus | | 74.81 | 58.08 | 65.39 |
| | ST-Corpus + C-Corpus | | 42.77 | 36.48 | 39.38 |
| | UlyssesNER-BR Corpora | | 73.89 | 59.44 | 65.05 |
| | PL-Corpus | **ST-Corpus** | 51.28 | 33.39 | 40.44 |
| | ST-Corpus | | 73.89 | 58.10 | 65.05 |
| | C-Corpus | | 13.76 | 8.68 | 10.64 |
| | **PL-Corpus + ST-Corpus** | | **74.80** | **61.94** | **67.76** |
| | PL-Corpus + C-Corpus | | 53.15 | 40.90 | 46.23 |
| | ST-Corpus + C-Corpus | | 73.23 | 60.27 | 66.12 |
| | UlyssesNER-BR Corpora | | 71.88 | 61.44 | 66.25 |
| | PL-Corpus | **C-Corpus** | 38.76 | 15.00 | 21.63 |
| | ST-Corpus | | 53.68 | 15.87 | 24.50 |
| | C-Corpus | | 70.00 | 38.04 | 49.30 |
| | PL-Corpus + ST-Corpus | | 58.43 | 22.61 | 32.60 |
| | PL-Corpus + C-Corpus | | 76.18 | 52.83 | 62.39 |
| | ST-Corpus + C-Corpus | | **79.50** | 54.78 | 64.86 |
| | **UlyssesNER-BR Corpora** | | 77.48 | **56.09** | **65.07** |
| **Tipos** | PL-Corpus | **PL-Corpus** | 80.67 | 66.60 | 72.96 |
| | ST-Corpus | | 48.51 | 38.62 | 43.01 |
| | C-Corpus | | 14.97 | 12.02 | 13.33 |
| | **PL-Corpus + ST-Corpus** | | **82.00** | **68.67** | **74.75** |
| | PL-Corpus + C-Corpus | | 78.49 | 64.73 | 70.95 |
| | ST-Corpus + C-Corpus | | 45.95 | 36.85 | 40.90 |
| | UlyssesNER-BR Corpora | | 79.63 | 67.39 | 73.00 |
| | PL-Corpus | **ST-Corpus** | 51.92 | 32.70 | 40.13 |
| | ST-Corpus | | 72.25 | 52.25 | 60.64 |
| | C-Corpus | | 16.67 | 10.38 | 12.79 |
| | PL-Corpus + ST-Corpus | | 74.36 | 55.71 | 63.70 |
| | PL-Corpus + C-Corpus | | 56.34 | 39.97 | 46.76 |
| | ST-Corpus + C-Corpus | | 72.85 | 54.33 | 62.24 |
| | **UlyssesNER-BR Corpora** | | **75.78** | **59.00** | **66.34** |
| | PL-Corpus | **C-Corpus** | 43.67 | 15.00 | 22.33 |
| | ST-Corpus | | 52.07 | 13.70 | 21.69 |
| | C-Corpus | | 76.66 | 47.83 | 58.90 |
| | PL-Corpus + ST-Corpus | | 60.12 | 21.96 | 32.17 |
| | PL-Corpus + C-Corpus | | 78.76 | 52.39 | 62.92 |
| | **ST-Corpus + C-Corpus** | | **79.74** | 53.91 | **64.33** |
| | UlyssesNER-BR Corpora | | 78.37 | **54.35** | 64.18 |

Table 5: Resultados do experimento de agregação dos conjuntos de dados para treinamento do modelo CRF.

Como podemos ver na Tabela 5, no caso do PL-corpus, a adição das

solicitações de trabalho no treinamento do modelo CRF melhoraram a qualidade das predições tanto na modelagem das categorias quanto na modelagem dos tipos. A semelhança entre as propriedades estatísticas estruturais dos projetos de lei e solicitações de trabalho, discutidas na Seção 3, pode ter contribuído para a alavancagem do *tagger*.

No caso das solicitações de trabalho, vemos que em ambos níveis de agregação, a adição dos projetos de lei também melhorou a resposta do modelo, reforçando ainda mais a ideia de semelhança semântica entre os corpus. Porém, é interessante notar que no caso dos tipos, a inclusão de comentários também foi benéfica. É provável que, como a distribuição das *tags* dos comentários é diferente das solicitações de trabalho, o acréscimo de sentenças do último aumenta a quantidade de exemplos de entidades mais raras no ST-corpus. Além disso, já foi constatado que um dos maiores inimigos do modelo CRF é a falta de dados, equivalentemente, a presença de *tags* raras.

Finalmente, considerar sentenças de todos os corpus no treinamento do modelo nas categorias e das solicitações de trabalho no caso dos tipos também foi suficiente para uma melhora considerável nas predições do modelo em comentários. A abundância de entidades nomeadas nos outros dois corpus compensa a falta no último a nível de modelagem.

Com isso, vemos que treinar o modelo de aprendizado de máquina com sentenças de múltiplos documentos é primordial para uma boa qualidade das predições. Vale ressaltar também que, apesar do modelo treinado no Ulysses-NER-BR corpora completo não ter tido o melhor resultado em alguns casos, ele não ficou muito atrás do melhor modelo em termos absolutos, além disso, foi o que mostrou mais robustez em todos os casos. Estas são características desejáveis em termos de aplicações. Considerando agora, o melhor modelo em cada caso, analisou-se a performance de cada CRF em cada uma das entidades nomeadas.

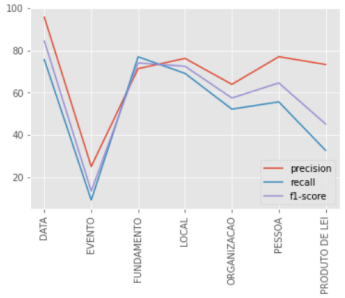
Na Figura 9, vemos que no caso do PL-corpus e ST-corpus, as categorias DATA, FUNDAMENTO, LOCAL e PESSOA foram as que o modelo teve mais facilidade de apontar. A ausência de exemplos da entidade EVENTO levou o modelo a ignorar completamente esta *tag*. Com respeito a ORGANIZAÇÃO e PRODUTODELEI vemos uma performance mediana. Já no caso do C-corpus, vemos uma uniformidade bem maior nas métricas de todas as entidades, variando em torno de 70% em todos os casos.

Por último na Figura 10, foi notado que tratando-se dos tipos, no PL-corpus as *tags* DATA, FUNDapelido, FUNDlei, LOCALconcreto, PESSOAindividual, PRODUTOprograma e PRODUTOsistema obtiveram as melhores

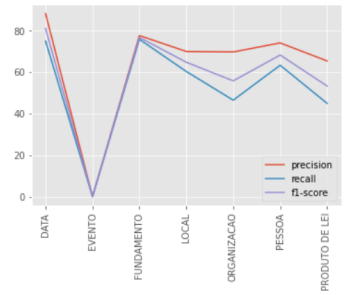
respostas do CRF, com resultados acima de 80%. Já EVENTO, FUNDsolicitaçãotrabalho e PESSOAgrupoind foram ignorados devido a escassez ou, simplesmente, total ausência dessas entidades neste tipo de documento.

No ST-corpus, vemos um comportamento muito parecido com o PL-corpus, com excessão de que neste caso, há a presença da entidade FUNDsolicitaçãotrabalho que não ocorre em projetos de lei e o *tagger* obteve um *score* alto nesta. Também vale ressalta a escassez de EVENTO e ORGpartido que levou o CRF a resultados ruins e a total ausência da entidade PESSOAgrupoind.

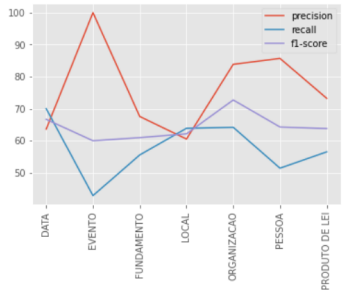
Já no C-corpus, as entidades FUNDprojetodelei, ORGpartido e PRODUTOoutros merecem destaque como as que obtiveram os melhores resultados, no que diz respeito as demais, vemos uma performance mediana reforçando o que foi constantado de que os comentários são ligeiramente mais complicados de se modelar devido à sua estrutura linguística. No caso das entidades escassas ou ausentes no corpus, tem-se FUNDsolicitaçaotrabalho, LOCALvirtual, PESSOAgrupoind, PRODUTOprograma e PRODUTOsistema.




(a) PL-corpus

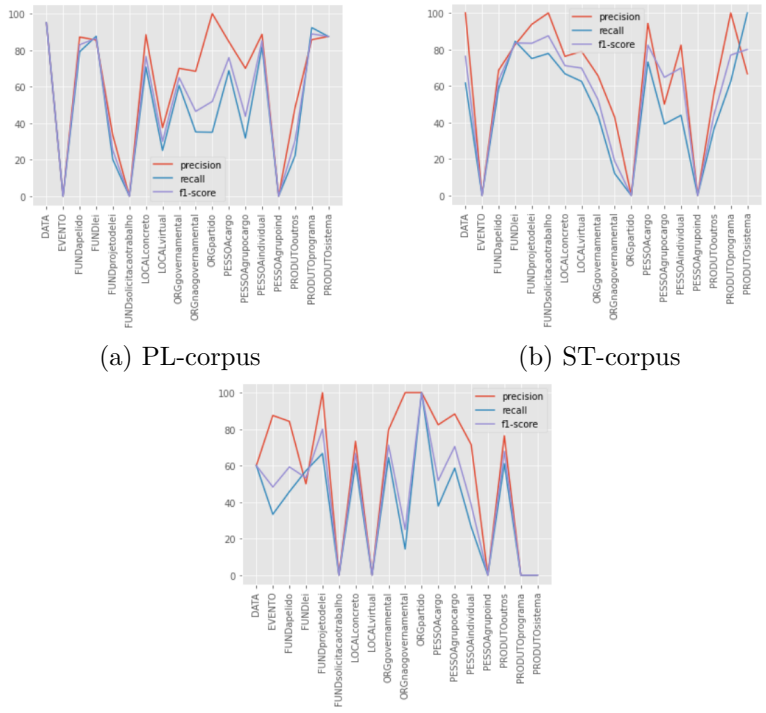



(b) ST-corpus




(c) C-corpus

Figure 9: Resultados do modelo CRF com mais chance de sucesso por categoria.




(a) PL-corpus

(b) ST-corpus

(c) C-corpus

Figure 10: Resultados do modelo CRF com mais chance de sucesso por tipo.

# 6 Conclusão e Trabalhos Futuros

Neste trabalho foi feita uma análise estatística descritiva e modelagem comparativa de modelos NER em 3 corpus compostos por documentos legais oriundos da Câmara dos Deputados do Brasil. Concluímos que o CRF é, dentre os modelos NER clássicos, o mais apto para tratamento dos dados e que o treinamento deste modelo utilizando sentenças de múltiplos documentos de diferentes espécies aumenta consideravelmente a qualidade das predições em comparação com a modelagem de cada corpus de maneira independente. Além disso, seu desempenho em cada entidade é proporcional à importância daquela no documento tratado o que o torna um forte *baseline* para comparação com futuras melhorias no *framework*.

Assim, sugere-se a testagem de arquiteturas de aprendizado profundo especializadas na tarefa de REN e possivelmente, tirar vantagem de técnicas

de aprendizado não-supervisionado para treinamento da rede neural, já que um dos principais obstáculos do trabalho é a rotulação manual dos dados. Outro caminho, é a exploração de técnicas de *active learning* para otimização do processo de anotação dos documentos a partir dos modelos treinados oriundos desta pesquisa.

# 7 Agradecimentos

Gostaríamos de agradecer a Câmara dos Deputados do Brasil pelo fornecimento dos conjuntos de dados, validação do processo de anotação e financeamento do projeto, e aos professores Ellen Souza, Nádia Félix, Rosimeire Costa e Hidelberg Oliveira pela imensa ajuda com o desenvolvimento do *framework*, desde o processo de tratamento dos dados até a avaliação dos modelos REN.

# References

[1] Lorena, Ana & Faceli, Katti & Almeida, Tiago & de Carvalho, Andre & Gama, João. (2021). Inteligência Artificial: uma abordagem de Aprendizado de Máquina (2a edição).

[2] Melo, Jairo. (2020). Inteligência artificial: uma realidade no Poder Judiciário. https://www.tjdft.jus.br/institucional/imprensa/campanhas-e-produtos/artigos-discursos-e-entrevistas/artigos/2020/inteligencia-artificial. Acessado em 23-11-21.

[3] Alencar, Ana. (2020). A Inteligência Artificial no Poder Judiciário Brasileiro: entendendo a nova"Justiça Digital". https://turivius.com/portal/inteligencia-artificial-no-poder-judiciario/. Acessado em 23-11-21

[4] Nadeau, D.; Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes, v. 30, n. 1, p. 3–26 (2007). Avaliable from https://www.jbe-platform.com/content/journals/10.1075/li.30.1.03nad.

[5] Luz de Araujo, P.H., de Campos, T.E., de Oliveira, R.R.R., Stauffer, M., Couto, S., Bermejo, P.: LeNER-Br: a dataset for named entity recognition

in Brazilian legal text. In: Villavicencio, A., et al. (eds.) PROPOR 2018. LNCS (LNAI), vol. 11122, pp. 313–323. Springer, Cham (2018).

[6] Alles, V.J.: Construção de um corpus para extrair entidades nomeadas do Diário Oficial da União utilizando aprendizado supervisionado. Master's thesis, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF (2018)

[7] Castro, P.V.Q.: Aprendizagem profunda para reconhecimento de entidades nomeadas em domínio jurídico. Masters thesis, Programa de Pós-graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás (2019).

[8] Pirovani, J. P. C.: CRF+LG: uma abordagem híbrida para o reconhecimento de entidades nomeadas em português. PhD thesis, Universidade Federal do Espírito Santo (2019).

[9] Collins, Michael. Natural Language Processing, Lecture Notes. http://www.cs.columbia.edu/ mcollins/. Acessado em 24-11-21

[10] Mykel J. Kochenderfer and Tim A. Wheeler. 2019. Algorithms for Optimization. The MIT Press.

[11] McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica, 22(3):276–282.

[12] WIEBE, J.; WILSON, T.; CARDIE, C. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, v. 39, n. 2, p. 165–210, 2005

[13] Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R. and Gurevych, I. (2018): The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, New Mexico, USA

[14] Loper, E., Bird, S.: NLTK: The Natural Language Toolkit. CoRR cs.CL/0205028 (2002).

[15] Wijffels J, Okazaki N (2007-2018). "crfsuite: Conditional Random Fields for Labelling Sequential Data in Natural Language Processing based on

CRFsuite: a fast implementation of Conditional Random Fields (CRFs)."
R package version 0.1, https://github.com/bnosac/crfsuite.