# A Comparative Analysis of Machine Learning Named Entity Recognition Tools for the Brazilian and European Portuguese Language Variants

**Breno David Lopes Pinheiro**[1], **Ellen Polliana Ramos Souza**[1], **Douglas Vitório**[1], **Hidelberg Oliveira Albuquerque**[1]

[1]Universidade Federal Rural de Pernambuco - Unidade Acadêmica de Serra Talhada
Caixa postal 063 – 56909-535 – Serra Talhada – PE – Brazil

brenodvlp@gmail.com,

{ellen.ramos, douglas.alisson, hidelberg.albuquerque}@ufrpe.br

***Abstract.*** *Textual information, despite being digital, are not computationally structured, requiring the use of techniques to structure and extract information. This work aims to evaluate machine learning NER (Named entity recognition) tools for the Brazilian and European Portuguese language variants. The Apache OpenNLP, Stanford CoreNLP and spaCy tools were selected for the evaluation. The HAREM corpus was used to train and evaluate the generated machine learning model; a tool was developed to preprocess the HAREM corpus. Two comparisons were performed: a general comparison and between the Portuguese variants. It was found that variants can confuse the training and evaluation process of NER models.*

***Resumo.*** *Informações textuais, apesar de digitais, não são computacionalmente estruturadas, necessitando do uso de técnicas para estruturá-las e extrair informações. Este trabalho tem o objetivo de avaliar ferramentas de REN utilizando machine learning para as variantes brasileira e europeia da língua portuguesa. As ferramentas Apache OpenNLP, Stanford CoreNLP e spaCy foram selecionadas; o corpus HAREM foi usado para treinar e avaliar os modelos; uma ferramenta foi desenvolvida para pré-processar o corpus HAREM. Dois tipos de comparações foram realizadas: uma geral e outra entre variantes do português. Foi possível identificar que as variantes podem afetar no treinamento e avaliação de modelos de REN (Reconhecimento de entidades nomeadas).*

## 1. Introdução

Documentos e textos, historicamente disponíveis somente em formato físico, passaram a ser distribuídos em formato digital. Entretanto, diferente da informação distribuída através de meios impressos, as informações digitais podem ser transmitidas mais rapidamente e em uma escala global graças à *Internet* [Gruhl et al. 2004].

Com esse grande número de informações textuais não estruturadas, para recuperar informações e descobrir padrões textuais de forma eficiente, faz-se necessário o uso de técnicas de mineração de texto. Extração de Informação (EI) é uma sub-área da mineração de texto que busca extrair automaticamente informação útil de dados não estruturados. Para isso, são utilizadas técnicas que buscam transformar documentos textuais não estruturados que possam ser processados por máquina [Weiss et al. 2004].

Por outro lado, pesquisas na área de extração de informação utilizando variantes da Língua Portuguesa são escassas, onde há um maior esforço no desenvolvimento de abordagens e ferramentas de alta qualidade para extração de informação nas línguas inglesa, alemã e francesa [do Amaral and Vieira 2014]. Contudo, para a atividade de extração de informação, é importante saber qual o melhor método para cada domínio ou língua, dado que muitas das pesquisas têm foco em abordagens mono linguísticas [Akbik et al. 2016]. Porém, com aprendizado de máquina (AM), é possível construir métodos de extração de informação independentes de idioma.

Dentre os métodos de extração de informação, há o de reconhecimento de entidades nomeadas, o qual busca reconhecer unidades de informação como nomes (pessoas, organizações, lugares), expressões numéricas (tempo, data, moeda, porcentagens) e qualquer outro tipo de entidade que seja nomeável. É possível realizar essa tarefa utilizando técnicas de aprendizado de máquina de forma supervisionada e não supervisionada [Nadeau and Sekine 2007].

Neste cenário, o objetivo deste trabalho é avaliar métodos de reconhecimento de entidades nomeadas para as variantes europeia e brasileira da Língua Portuguesa. Assim, as contribuição do trabalho são:

- Mapeamento, seleção e avaliação de ferramentas de reconhecimento de entidades nomeadas para a Língua Portuguesa que fazem uso de aprendizado de máquina;
- Desenvolvimento de ferramenta para tratamento e pré-processamentà do texto, considerando diferentes variantes da Língua Portuguesa;
- Avaliação das ferramentas selecionadas paras as variantes portuguesa e europeia da língua portuguesa, utilizando o *corpus* de *benchmark* HAREM [Santos and Cardoso 2007].

O restante do trabalho está organizando da seguinte forma: na Seção 2, são apresentados conceitos e trabalhos relacionados à esta pesquisa. A Seção 3 apresenta o método adotado na condução da avaliação. Na Seção 4, são apresentados os resultados da análise comparativa para cada uma das variantes estudadas. Por fim, a Seção 5 apresenta a conclusão e os trabalhos futuros.

## 2. Fundamentação Teórica

A Mineração de texto pode ser amplamente definida como um processo intensivo de descoberta de conhecimento a partir de uma coleção de documentos, utilizando um conjunto de ferramentas de análise [Aggarwal and Zhai 2012]. Assim como o processo de mineração de dados, a mineração de texto busca extrair informações úteis de fontes de dados através da identificação e exploração de padrões [Ronen Feldman 2006].

A Mineração de texto ocupa-se com a análise textual com suporte de máquina. Nela, são utilizadas técnicas de Recuperação de Informação (RI), Extração de Informação e também Processamento de Linguagem Natural (PLN), bem como algoritmos utilizados em mineração de dados [Hotho et al. 2005].

### 2.1. Reconhecimento de Entidades Nomeadas

Uma entidade nomeada é uma sequência de palavras que designam alguma entidade do mundo real, por exemplo: “California”, “Steve Jobs” e “Apple Inc”. A tarefa de Reconhecimento de Entidades Nomeadas (REN) ou Reconhecimento de Entidades Mencionadas (REM) é usada para identificar entidades nomeadas de textos escritos em forma

livre e classificá-las em um conjunto pré-definido de tipos, como: pessoa, organização e localização. Na tarefa de reconhecimento de entidades nomeadas, vários algoritmos de aprendizado de máquina podem ser utilizados. Dentre eles, se destacam: *Conditional Random Fields* (CRF), *Maximum-entropy Markov Model* (MEMM), *Support Vector Machine* (SVM) e Árvore de Decisão [Jiang et al. 2012].

### 2.2. Extração de Informação na Língua Portuguesa

São vários os esforços para o desenvolvimento do estado da arte na área de extração de informação utilizando a Língua Portuguesa. Dentre as iniciativas, destaca-se a ferramenta utilizando CRF para extração de informação chamada NERP-CRF [do Amaral and Vieira 2014]. Também, há a Linguateca [Liguateca 2015], um centro de recursos, que fornece *corpora*, ferramentas computacionais, repositório de teses na área de processamento de texto e organização de avaliações conjuntas. Seu acesso é público e gratuito e sua manutenção e desenvolvimento são feitos através da colaboração entre diversos pesquisadores de PLN para a Língua Portuguesa. Outros esforços a serem destacados são os eventos de avaliação conjunta HAREM [Linguateca 2013]. Esses eventos tiveram como objetivo avaliar métodos de reconhecimento de entidades nomeadas e suas relações [Santos and Cardoso 2007]. A primeira edição do HAREM ocorreu em 2004 e a segunda em 2008. Ambos eventos geraram contribuições, como a primeira e segunda coleção dourada do HAREM, os quais são *corpora* anotados, podendo ser utilizados na implementação, treinamento e avaliação de sistemas de reconhecimento de entidades nomeadas [Santos and Cardoso 2007].

### 2.3. Trabalhos relacionados

O trabalho de [Pires et al. 2017] utiliza um *benchmark* de ferramentas de reconhecimento de entidades nomeadas aplicado à Língua Portuguesa. Seu objetivo foi avaliar a performance de ferramentas estabelecidas de REN: *Stanford CoreNLP*, *OpenNLP*, *spaCy* e NLTK. Para isso, foi utilizada a coleção dourada do HAREM. Para realização das atividades de REN, utilizou-se o comportamento padrão das ferramentas, ou seja, não houve qualquer interferência ou mudança nos parâmetros. Como resultado, observou-se que a ferramenta *Stanford CoreNLP* teve o melhor desempenho, com 56,10% de F-*measure*, seguida por *Apache OpenNLP*, *spaCy* e NLTK. A Tabela 1 apresenta os resultados de forma mais detalhada.

No trabalho de [Amaral et al. 2014], foi realizado um estudo comparativo com quatro sistemas de reconhecimento de entidades nomeadas com base na Língua Portuguesa, sendo elas: Freeling, LanguageTasks, PALAVRAS e NERP-CRF. O objetivo dessa comparação foi estimar a competitividade de cada sistema em termos de efetividade e eficiência. Foi usado o *corpus* HAREM como fonte de dados, porém utilizando apenas três das dez categorias de entidades possíveis: pessoa, local e organização. Na Tabela 1 é possível identificar os resultados deste trabalho.

No trabalho de [Fonseca et al. 2015], foi apresentado o processo de construção de um modelo para reconhecimento de entidades nomeadas usando a classe *NameFinder* do *Apache OpenNLP*. O objetivo do trabalho foi reconhecer e classificar entidades nomeadas em Português, quando não havia um modelo para a língua até dado momento. Para treinamento e avaliação, foram utilizados os *corpora* Amazônia, para treinamento do modelo, e HAREM, para avaliação. Esse trabalho não realizou uma comparação entre ferramentas.

**Tabela 1. Comparação dos trabalhos relacionados**

| Trabalho | Métodos | Pré-processamento | Métricas | Corpora | Entidades reconhecidas | Ranking Geral (F-measure) |
|---|---|---|---|---|---|---|
| Pires, Devezas e Nunes | CoreNLP<br>OpenNLP<br>SpaCy<br>NLTK | Tokenização<br>Segmentação de Sentenças<br>POS tagging | Precisão<br>Recall<br>F-measure | HAREM | Pessoa, organização, tempo, local, obra, acontecimento, abstração, coisa, valor, variado | 1. CoreNLP (56.10%)<br>2. OpenNLP (53.63%)<br>3. SpaCy (46.81%)<br>4. NLTK (30.97%) |
| Amaral, Fonseca Lopes e Vieira | FreeLing<br>LanguageTasks<br>PALAVRAS<br>NERP-CRF | POS tagging | Precisão<br>Recall<br>F-measure | HAREM | Pessoa, local, organização | 1. PALAVRAS (57%)<br>2. LanguageTasks (55%)<br>3. FreeLing (54%)<br>4. NERP-CRF (53%) |
| Fonseca, Chiele, Vieira e Vanin | OpenNLP | Remoção das POS<br>Adição de marcação das categorias | Precisão<br>Recall<br>F-measure | Amazônia<br>HAREM | Pessoa, organização, tempo, local, obra, acontecimento, abstração, coisa, valor, variado | 1. OpenNLP (38,06%) |

### 2.3.1. Análise comparativa

A análise comparativa entre os trabalhos relacionados está sumarizada na Tabela 1. É possível perceber que os resultados são parecidos entre os trabalhos de extração de informação: aproximadamente entre 30% e 57%. Também, foi possível identificar que abordagens que utilizam aprendizado de máquina tiveram bons resultados, como *Stanford CoreNLP*/CRF (56.10%) e NERP-CRF/CRF (53%). Entretanto, a ferramenta com maior índice foi a PALAVRAS, a qual não utiliza AM.

O trabalho de [Fonseca et al. 2015] utilizou uma maior quantidade de dados da variante brasileira do Português para treino, pela presença do *corpus* Amazônia. Por ter sido treinado com base nele, o modelo apresentou uma maior tendência para a variante BR. Esse pode ser um dos motivos pelos quais seu resultado (F-*measure* de 38,06%) se distanciou do resultado do trabalho de [Pires et al. 2017] (F-*measure* de 53,63%).

Portanto, por conta da existência desse impacto do uso das variantes da Língua Portuguesa, este trabalho propõe uma análise comparativa visando observar como as diferenças das variantes linguísticas podem afetar o desempenho do modelo de aprendizado de máquina treinada em outra variante.

## 3. Método

Para a realização da análise comparativa, foi adotado o processo apresentado na Figura 1, o qual está organizado em cinco etapas.

**Figura 1. Processo adotado neste trabalho para REN.**


### 3.1. Mapeamento de métodos de extração de informação

Para a seleção dos métodos de extração de informação, foi realizado um levantamento a partir de trabalhos publicados que fazem uso de ferramentas e algoritmos de extração de informação utilizando os seguintes critérios: 1) suporte para a Língua Portuguesa; 2) acesso aberto e gratuito; e 3) utilização de aprendizado de máquina. A Tabela 2 representa os métodos selecionados.

**Tabela 2. Métodos de EI selecionados**

| Método | Tipo | Abordagem | Fonte | Link para download |
|---|---|---|---|---|
| Apache OpenNLP | Ferramenta/API | ML | [Baldridge 2005] | `https://opennlp.apache.org` |
| Stanford CoreNLP | Ferramenta/API | ML | [Manning et al. 2014] | `https://stanfordnlp.github.io/CoreNLP` |
| spaCy | Ferramenta/API | ML | [Honnibal and Montani 2017] | `https://spacy.io` |

## 3.2. Definição do corpus

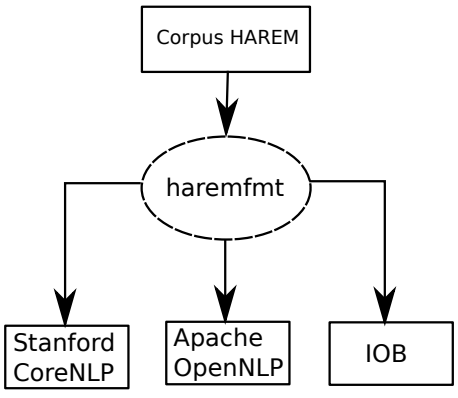
Para treinamento e teste dos métodos de extração de informação selecionados, é necessária a utilização de um *corpus* anotado. Para garantir uma boa segurança quanto aos resultados e a capacidade de comparação e replicação desta pesquisa, foi escolhida a coleção dourada do segundo HAREM. Essa escolha se deu pelo fato de este *corpus* ser altamente robusto e consolidado na área de extração de informação para a Língua Portuguesa, tendo sido utilizado, inclusive, pelos trabalhos relacionados [Santos et al. 2004, Amaral et al. 2013, Fonseca et al. 2015, Pires et al. 2017].

## 3.3. Pré-Processamento

Nesta etapa, foram removidas palavras que não possuem um significado relevante para seu uso na etapa de extração de informação e foram tratados dados faltosos ou inconsistentes.

Para o pré-processamento, foi desenvolvida a ferramenta *haremfmt* [Pinheiro 2020]. Com ela, é possível manipular o *corpus* HAREM, podendo, por exemplo: extrair segmentos do *corpus* em níveis de: parágrafo, documento e entidade; distinguir as variantes linguísticas (BR/PT) presentes no *corpus*; gerar saída das anotações nos formatos *Inside, outside, beginning* (IOB), *OpenNLP*, *CoreNLP*. A Figura 2 ilustra o funcionamento da ferramenta. Dos métodos de pré-processamento, foram realizados: tokenização e segmentação de sentenças, ambos implementados pela ferramenta desenvolvida.

**Figura 2. Funcionamento da *haremfmt*.**




Todas as ferramentas comparadas possuem técnicas para pré-processamento das bases de dados, porém a ferramenta de tratamento desenvolvida especialmente para o *corpus* HAREM permite uma maior flexibilidade. Além disso, com o método genérico de pré-processamento fornecido pelas ferramentas, não seria possível, por exemplo, dividir o *corpus* de acordo com as variantes linguísticas presentes, impedindo a comparação entre essas variantes, indo de encontro com o objetivo deste trabalho. Também não seria possível fazer o tratamento de anotações alternativas ou desconsiderá-las.

### 3.4. Reconhecimento de Entidades Nomeadas

Para a realização desta tarefa, foi necessária a definição das entidades a serem reconhecidas. A Tabela 3 traz as categorias de entidades presentes no *corpus* HAREM [Mota and Santos 2008].

**Tabela 3. Categorias do *corpus* HAREM.**

| Categoria | Quantidade Anotações |
|---|---|
| PESSOA | 2036 |
| LOCAL | 1311 |
| TEMPO | 1189 |
| ORGANIZACAO | 961 |
| OBRA | 449 |
| VALOR | 353 |
| COISA | 308 |
| ACONTECIMENTO | 300 |
| ABSTRACCAO | 286 |
| OUTRO | 79 |

As ferramentas selecionadas possuem interfaces em linha de comando para execução dessa tarefa. Além dessa interface, essas ferramentas também possuem APIs, permitindo que o processo fosse totalmente programável. Com a ferramenta desenvolvida no trabalho, a *haremfmt*, foi possível definir o nível de especificidade de tratamento do *corpus*: nos níveis de documento, sentença e entidade. Para este trabalho, foi usado o nível de sentença, pois permite a comparação com os trabalhos relacionados.

A ferramenta *OpenNLP* requer um formato de aquivo próprio como entrada. Neste arquivos, as entidades são anotadas a partir dos delimitadores $\langle start \rangle$ e $\langle end \rangle$. Então, o comando *TokenNameFinderTrainer* é executado com os parâmetros: arquivo em que o modelo será salvo, arquivo de entrada anotado, codificação do arquivo de entrada e língua.

A ferramenta *CoreNLP* também requer que o arquivo esteja em um formato específico, no qual é somente possível ter um *token* por linha, tendo as anotações em outra coluna, separada por um carácter de tabulação. Quando não há anotação, o carácter "O" é usado. Também, há um arquivo de configuração, no qual todos os parâmetros de treinamento do modelo, bem como a localização do aquivo de treinamento são definidos. Para executar a ferramenta, utiliza-se o comando *CRFClassifier* passando o arquivo de configuração como entrada.

A ferramenta *spaCy* aceita o formato JSON como entrada, porém também possui um comando que transforma formatos como *IOB*, *CoNLL* para o formato requerido pela ferramenta. Com o *haremfmt* foi possível gerar um arquivo no formato IOB, que depois foi convertido para JSON. A partir disso, foi realizado o treinamento do modelo, fornecendo, como argumentos, o idioma, o arquivo em que o modelo será salvo e o arquivo de treinamento em formato JSON.

### 3.5. Avaliação

Foram realizados dois tipos de avaliação: uma geral, utilizando toda a coleção dourada do segundo HAREM para treinamento e avaliação do modelo e outra fazendo uma

comparação entre variantes da Língua Portuguesa presentes no *corpus*: português europeu e português brasileiro.

Para a avaliação geral dos modelos treinados, foi utilizada a validação cruzada, ou *cross validation*, K-fold com $K = 10$. O tamanho do volume de dados para treinamento foi de 90% e testes 10%. A escolha desse tipo de avaliação se deu, pois, na literatura, ela é muito utilizada para avaliar modelos baseados em aprendizado de máquina. Assim, é possível obter informações como: frequência de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos. Essas informações dão base para calcular medidas de desempenho do modelo treinado. Neste trabalho, foram utilizadas três métricas para a avaliação do método de extração de informação: Precisão, *Recall* e F-*measure*.

Para a avaliação de variantes, foram definidos três conjuntos de dados: variante português europeu (PT), variante português brasileiro (BR) e a junção das duas variantes (MIX), a qual possui dados balanceados das duas variantes. Essa divisão é semelhante a utilizada no trabalho de [Vitório et al. 2017], no qual foi realizada uma avaliação do impacto das variantes do Português em mineração de opinião. O objetivo desta análise foi de verificar se modelos treinados com uma variante conseguem identificar entidades de outra, além de possíveis nuances e características específicas de cada variante.

## 4. Resultados

Como citado anteriormente, foi utilizada a técnica 10-fold para obtenção dos resultados. Dessa forma, os resultados apresentados são a média das 10 execuções da validação cruzada para cada cenário. A Tabela 4 apresenta os resultados obtidos pela avaliação geral das três ferramentas, considerando uma análise a nível de parágrafo.

Na avaliação por sentença e desconsiderando variantes, obteve-se o seguinte resultado: *Stanford CoreNLP* com o melhor F-*measure* (69,24%), seguido do *Apache OpenNLP* (63,50%) e *spaCy* (33,31%).

**Tabela 4. Avaliação geral a nível de parágrafo.**

| Ferramenta | Precisão | Recall | F-measure | Desvio padrão (F-measure) |
|---|---|---|---|---|
| Stanford CoreNLP | 71,65% | 66,99% | 69,24% | 2,36 |
| Apache OpenNLP | 65,39% | 61,74% | 63,50% | 2,17 |
| spaCy | 36,47% | 30,68% | 33,31% | 2,21 |

A ferramenta *OpenNLP* teve seu melhor desempenho na classe *pessoa*, com a F-*measure* de 70,12%, seguida de *tempo* (69,07%) e *local* (66,7%). Os resultados completos para essa ferramenta estão listados na Tabela 5.

Quanto à comparação no reconhecimento dos tipos de entidades nomeadas, o *CoreNLP* obteve o melhor desempenho na identificação da entidade *tempo*, com 79,77% de F-*measure*, seguido de *pessoa* (77,20%) e *local* (71,53%). A Tabela 6 mostra os resultados alcançados por essa ferramenta para cada entidade.

A ferramenta *spaCy* não retorna o resultado para cada classe individual, logo não foi possível incluí-la nesta comparação.

A Figura 3 mostra o F-measure das ferramentas por classe.

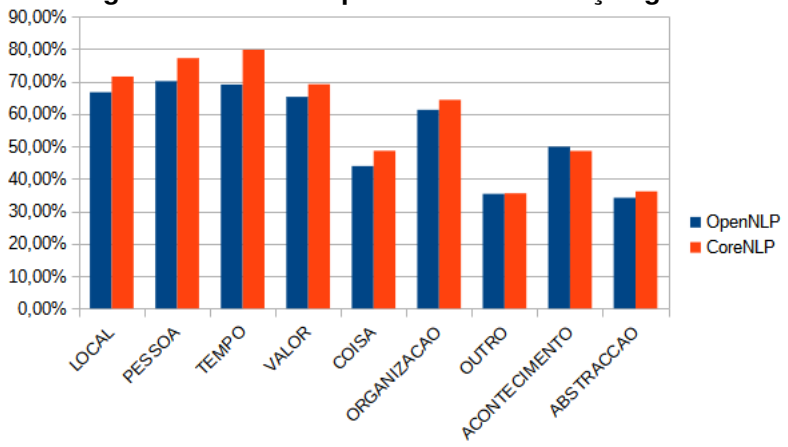
**Tabela 5. Comparação de classes do HAREM na ferramenta OpenNLP.**

| Classe | Precisão | Recall | F-measure |
|---|---|---|---|
| LOCAL | 65,57% | 68,49% | 66,70% |
| PESSOA | 66,65% | 74,27% | 70,12% |
| TEMPO | 75,54% | 63,78% | 69,07% |
| VALOR | 63,59% | 67,77% | 65,27% |
| COISA | 56,81% | 36,86% | 43,92% |
| ORGANIZACAO | 62,56% | 60,94% | 61,26% |
| OUTRO | 62,59% | 27,48% | 35,35% |
| ACONTECIMENTO | 66,91% | 40,76% | 49,89% |
| ABSTRACCAO | 50,09% | 26,72% | 34,15% |

**Tabela 6. Comparação de classes do HAREM na ferramenta CoreNLP.**

| Classe | Precisão | Recall | F-measure |
|---|---|---|---|
| LOCAL | 69,23% | 74,23% | 71,53% |
| PESSOA | 74,13% | 80,68% | 77,20% |
| TEMPO | 86,15% | 74,38% | 79,77% |
| VALOR | 77,61% | 72,89% | 69,17% |
| COISA | 73,71% | 37,01% | 48,64% |
| ORGANIZACAO | 64,29% | 64,73% | 64,34% |
| OUTRO | 76,18% | 24,79% | 35,53% |
| ACONTECIMENTO | 64,88% | 39,94% | 48,59% |
| ABSTRACCAO | 54,48% | 27,78% | 36,11% |

**Figura 3. F-measure por classe na avaliação geral**




## 4.1. Variantes

A seguir, nas Tabelas 7, 8 e 9, encontram-se os resultados do estudo do impacto das variantes brasileira e europeia em cada ferramenta. A coluna "Treino" indica a variante utilizada no treinamento do modelo. A coluna "Teste" indica a variante utilizada para avaliar o modelo. O conjunto "MIX" é uma união das duas variantes de forma balanceada, os conjuntos "BR" e "PT" remetem às variantes brasileira e portuguesa, respectivamente. É importante perceber, que há uma diferença entre a avaliação geral e entre variantes "MIX-MIX". Como o *corpus* teve de ser balanceado para essa avaliação, o conjunto "MIX" foi reduzido de modo que a quantidade de documentos fosse igual as das variantes "BR" e "PT".

**Tabela 7. Avaliação de variantes na ferramenta OpenNLP**

| Treino | Teste | Precisão | Recall | F-measure |
|---|---|---|---|---|
| MIX | MIX | 58,37% | 54,02% | 56,11% |
| MIX | BR | 60,74% | 58,26% | 59,47% |
| MIX | PT | 89,11% | 87,17% | 88,13% |
| BR | MIX | 58,91% | 53,47% | 56,06% |
| BR | BR | 61,02% | 57,25% | 59,08% |
| BR | PT | 46,36% | 40,70% | 43,35% |
| PT | MIX | 86,69% | 85,53% | 86,11% |
| PT | BR | 45,69% | 44,06% | 44,86% |
| PT | PT | 61,41% | 56,35% | 58,77% |

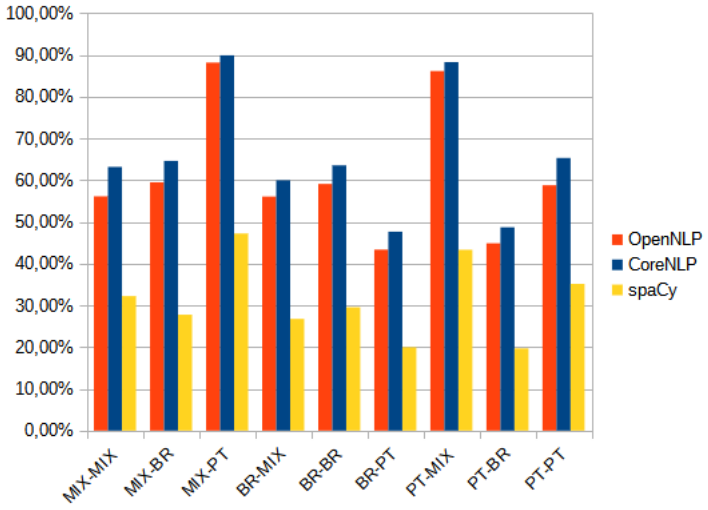
**Tabela 8. Avaliação de variantes na ferramenta CoreNLP.**

| Treino | Teste | Precisão | Recall | F-measure |
|---|---|---|---|---|
| MIX | MIX | 65,97% | 60,57% | 63,15% |
| MIX | BR | 67,12% | 62,24% | 64,59% |
| MIX | PT | 91,45% | 88,36% | 89,88% |
| BR | MIX | 65,10% | 55,57% | 59,95% |
| BR | BR | 66,77% | 60,60% | 63,53% |
| BR | PT | 53,26% | 43,10% | 47,64% |
| PT | MIX | 89,15% | 87,37% | 88,25% |
| PT | BR | 50,75% | 46,81% | 48,70% |
| PT | PT | 68,30% | 62,54% | 65,29% |

**Tabela 9. Avaliação de variantes na ferramenta spaCy.**

| Treino | Teste | Precisão | Recall | F-measure |
|---|---|---|---|---|
| MIX | MIX | 35,12% | 29,73% | 32,20% |
| MIX | BR | 28,91% | 26,68% | 27,75% |
| MIX | PT | 52,05% | 43,19% | 47,21% |
| BR | MIX | 30,11% | 24,07% | 26,75% |
| BR | BR | 31,64% | 27,68% | 29,53% |
| BR | PT | 21,86% | 18,37% | 19,96% |
| PT | MIX | 45,14% | 41,62% | 43,31% |
| PT | BR | 21,20% | 18,36% | 19,68% |
| PT | PT | 39,01% | 31,95% | 35,13% |

**Figura 4. F-measure da avaliação por variantes**




## 4.2. Discussão dos resultados

Dentre as ferramentas analisadas, pôde-se notar que a *Stanford CoreNLP* apresentou o melhor desempenho nos cenários definidos neste trabalho. Já a ferramenta *Apache OpenNLP*, apesar de ter um desempenho inferior à *CoreNLP*, mostrou-se uma boa concorrente, tendo um *gap* percentual de aproximadamente 6% em comparação com a melhor colocada. Além disso, apesar de tempo e a utilização dos recursos computacionais não terem sido considerados como critérios de avaliação, durante a execução dos experimentos, foi observado que esta ferramenta utiliza menos recursos e tempo que a *CoreNLP*. Assim, pode-se justificar o uso dessa ferramenta pelo *trade-off* entre menor tempo/recurso e desempenho dos seus modelos, embora inferiores aos da *CoreNLP*.

A ferramenta *spaCy* apresentou um fraco desempenho quando comparada às outras duas. Porém, como ponto positivo, pode-se apresentar que ela possui uma API, permitindo a utilização de outras APIs externas de aprendizado de máquina e forma programática.

Quanto à avaliação das entidades individuais, as reconhecidas mais facilmente foram: *tempo*, *pessoa* e *local*, em ambas as ferramentas que permitem essa avaliação (*OpenNLP* e *CoreNLP*). Isso pode ter relação com a quantidade de exemplos que o modelo recebe. É evidenciado na Tabela 3 e na Figura 3 que as entidades com maiores quantidades

são justamente as que apresentaram melhor desempenho. Isso mostra uma relação entre a quantidade de exemplos e a performance do modelo. Além disso, a performance das ferramentas *OpenNLP* e *CoreNLP* foram muito semelhantes nas entidades "OUTRO", "ACONTECIMENTO" e "ABSTRACCAO".

No estudo de variantes, foi observado que os melhores desempenhos foram entre as variantes MIX e PT, enquanto que os resultados envolvendo a variante BR tiveram um desempenho aproximado do geral. Na ferramenta *OpenNLP*, o conjunto treino-teste MIX/BR teve um F-*measure* de 59,47%, e, na avaliação geral, essa mesma ferramenta teve um desempenho de 63,50%.

Quanto as variantes PT e BR, houve uma grande diferença de desempenho, tendo os menores valores de F-*measure* registrados nessa comparação, o *OpenNLP*: BR/PT (43,35%), PT/BR (44,86%), *spaCy*: BR/PT (19,96%) e PT/BR (19,68%). A ferramenta *CoreNLP* obteve os melhores resultados entre todas as combinações de variantes, mas ainda seguindo o padrão das anteriores: BR/PT (47,64%) e PT/BR (48,7%). Isso pode ser um indicativo de que as variantes afetam no treinamento e uso de modelos, pois entidades como organizações, coisas, acontecimentos, locais possuem diferenças linguísticas e culturais entre essas variantes, prejudicando no desempenho desses modelos.

Isso vai de acordo com o trabalho de [Castro et al. 2017], no qual foi possível verificar que diferenças léxicas parecem ter uma maior relevância do que diferenças sintáticas quanto as variantes linguísticas do português.

É possível perceber que a quantidade de documentos também influencia no desempenho do modelo, dado que as avaliações "MIX-MIX" tiveram um pior desempenho em comparação com a avaliação geral, pois, apesar de possuírem o mesmo domínio, há uma diferença na quantidade de documentos usados para treinar e testar o modelo.

## 5. Conclusão e trabalhos futuros

Neste trabalho, foram mapeados e comparados métodos de reconhecimento de entidades nomeadas para a Língua Portuguesa, os quais utilizam aprendizado de máquina: *Apache OpenNLP*, *Stanford CoreNLP* e *spaCy*. Além disso, utilizando essas ferramentas, pôde-se avaliar se as diferenças existentes nas variantes brasileira e europeia do português afetam os modelos de reconhecimento de entidades nomeadas.

Com base nos resultados, pôde-se perceber que a ferramenta *Stanford CoreNLP* apresenta um desempenho superior, embora a *Apache OpenNLP* possa ser mais indicada caso tempo e recursos computacionais sejam um fator importante. Já com relação à análise das variantes linguísticas para as ferramentas apresentadas, pôde-se verificar que há um impacto no modelo, já que a utilização de uma variante para treino e outra para teste resultou em uma queda no desempenho.

Como trabalhos futuros, há a possibilidade de realizar essa avaliação em outros níveis, além do nível de sentença, utilizado nesta pesquisa. O treinamento e avaliação dos modelos foram realizados utilizando a interface de linha de comando das ferramentas, sem qualquer alteração nas configurações padrão. Porém, todas essas ferramentas possuem APIs em que é possível utilizar recursos disponibilizados por elas de forma programável, permitindo assim um maior controle sob o treinamento e avaliação de modelos. Por fim, é possível o desenvolvimento conjunto ou individual de modelos oficiais de REN para

a Língua Portuguesa, dado que nenhuma dessas ferramentas possui um modelo de REN oficial para língua portuguesa no momento da realização deste trabalho [OpenNLP 2020, CoreNLP 2020, spaCy 2020].

## Referências


[Aggarwal and Zhai 2012] Aggarwal, C. C. and Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

[Akbik et al. 2016] Akbik, A., Chiticariu, L., Danilevsky, M., Kbrom, Y., Li, Y., and Zhu, H. (2016). Multilingual information extraction with PolyglotIE. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 268–272, Osaka, Japan. The COLING 2016 Organizing Committee.

[Amaral et al. 2014] Amaral, D., Fonseca, E., Lopes, L., and Vieira, R. (2014). Comparative analysis of Portuguese named entities recognition tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2554–2558, Reykjavik, Iceland. European Language Resources Association (ELRA).

[Amaral et al. 2013] Amaral, D. O. F. d. et al. (2013). O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa.

[Baldridge 2005] Baldridge, J. (2005). The opennlp project. *URL: http://opennlp. apache. org/index. html,(accessed 2 February 2012)*, page 1.

[Castro et al. 2017] Castro, D. W., Souza, E., Vitório, D., Santos, D., and Oliveira, A. L. (2017). Smoothed n-gram based models for tweet language identification: A case study of the brazilian and european portuguese national varieties. *Applied Soft Computing*, 61:1160–1172.

[CoreNLP 2020] CoreNLP, S. (2020). *Modelos oficiais do Stanford CoreNLP*.

[do Amaral and Vieira 2014] do Amaral, D. O. F. and Vieira, R. (2014). Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. *Linguamática*, 6(1):41–49.

[Fonseca et al. 2015] Fonseca, B. E., Chiele, C. G., Vieira, R., and Vanin, A. A. (2015). Reconhecimento de entidades nomeadas para o português usando o opennlp. In *XII National Meeting on Artificial and Computational Intelligence*. Brazilian Conference on Inteligent Systems.

[Gruhl et al. 2004] Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM.

[Honnibal and Montani 2017] Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

[Hotho et al. 2005] Hotho, A., Nürnberger, A., and Paass, G. (2005). A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20:19–62.

[Jiang et al. 2012] Jiang, J., Aggarwal, C. C., and Zhai, C. X. (2012). *Mining Text Data*. Springer Publishing Company, Incorporated.

[Liguateca 2015] Liguateca (2015). Linguateca.

[Linguateca 2013] Linguateca (2013). Harem: Reconhecimento de entidades mencionadas em português.

[Manning et al. 2014] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

[Mota and Santos 2008] Mota, C. and Santos, D., editors (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. ISBN: 978-989-20-1656-6.

[Nadeau and Sekine 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

[OpenNLP 2020] OpenNLP, A. (2020). *Modelos oficiais do Apache OpenNLP*.

[Pinheiro 2020] Pinheiro, B. (2020). Repositório com código-fonte haremfmt.

[Pires et al. 2017] Pires, A., Devezas, J. L., and Nunes, S. (2017). Benchmarking named entity recognition tools for portuguese.

[Ronen Feldman 2006] Ronen Feldman, J. S. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 1 edition.

[Santos and Cardoso 2007] Santos, D. and Cardoso, N. (2007). Reconhecimento de entidades mencionadas em português: Documentação e actas do harem, a primeira avaliação conjunta na área.

[Santos et al. 2004] Santos, D., Simões, A., Frankenberg-Garcia, A., Pinto, A., Barreiro, A., Maia, B., Mota, C., Oliveira, D., Bick, E., Ranchhod, E., et al. (2004). Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa.

[spaCy 2020] spaCy (2020). *Modelos oficiais do spaCy*.

[Vitório et al. 2017] Vitório, D., Souza, E., Teles, I., and Oliveira, A. L. I. (2017). Investigating opinion mining through language varieties: a case study of brazilian and european portuguese tweets. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 43–52, Porto Alegre, RS, Brasil. SBC.

[Weiss et al. 2004] Weiss, S., Indurkhya, N., Zhang, T., and Damerau, F. (2004). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. SpringerVerlag.