Um modelo híbrido para o Reconhecimento de Entidades Nomeadas em português

DOI: http://dx.doi.org/10.21165/el.v51i3.3271

Andressa Vieira e Silva[1]

Marcos Lopes[2]

Resumo

O Reconhecimento de Entidades Nomeadas (REN) é uma tarefa computacional voltada para a classificação automática de termos chamados de Entidades Nomeadas em um texto, como os nomes de pessoas, lugares e organizações. Nesta pesquisa, propomos um modelo híbrido para o REN em português, que combina representações *word embeddings* e traços baseados em representações linguísticas explícitas (como regras morfossintáticas e pronomes de tratamento) aplicados a uma rede neural BiLSTM-CRF. O modelo foi treinado no *corpus* Harem (SANTOS; CARDOSO, 2007), obtendo 81,06% de medida-F, o que representa uma melhora estatisticamente significativa em relação ao modelo treinado somente com representações *word embeddings*. A BiLSTM-CRF também superou os resultados obtidos pelo módulo spaCy (HONNIBAL; MONTANI, 2017) e ficou um pouco acima do modelo BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020). Esses resultados sugerem que a incorporação de traços linguísticos pode melhorar o desempenho de modelos de redes neurais no reconhecimento de entidades nomeadas em textos.

Palavras-chave: Linguística Computacional; Reconhecimento de Entidades Nomeadas; representação do conhecimento linguístico; modelos híbridos.

1 Universidade de São Paulo (USP), São Paulo, São Paulo, Brasil; andressa.silva94v@gmail.com; https://orcid.org/0000-0001-7402-2702

2 Universidade de São Paulo (USP), São Paulo, São Paulo, Brasil; marcoslopes@usp.br; https://orcid.org/0000-0002-6147-7369

A hybrid model for Named Entity Recognition in Portuguese

Abstract

Named Entity Recognition (NER) is the computational task of retrieving and classifying nouns referring to certain classes of entities such as person, location, and organization names. In this paper we put forth a hybrid model for NER in Portuguese incorporating word embeddings and some explicit linguistic features (eg. morphosyntactic rules and honorifics) into a BiLSTM-CRF neural network. The model achieved 81.06 F-score in Harem corpus (SANTOS; CARDOSO, 2007). The outcomes reveal statistically significant improvements compared to the performance of the same model based exclusively upon word embeddings, without any linguistic attributes. Results were also significantly better than spaCy module's (HONNIBAL; MONTANI, 2017) on the same dataset, and slightly superior to BERTimbau model (SOUZA; NOGUEIRA; LOTUFO, 2020) trained and evaluated in Harem. These outcomes suggest that incorporating linguistic features can potentially improve the performance of neural network models in recognizing named entities in texts.

Keywords: Computational Linguistics; Named Entity Recognition; linguistic knowledge representation; hybrid models.

Introdução

O termo "Entidade Nomeada" foi cunhado para a Sexta Conferência sobre Compreensão de Mensagens (MUC-6), em 1996 (NADEAU; SEKINE, 2009). Essa conferência era voltada ao estudo de métodos de Extração de Informação (IE), um conjunto de técnicas voltadas a localizar conteúdos específicos, como resultados de buscas, em grandes conjuntos de textos não-estruturados (isto é, não previamente anotados, classificados ou tabulados). No contexto da IE, foi particularizada a tarefa de identificar quais dos conteúdos textuais poderiam denotar "entidades", inicialmente concebidas como sendo de três tipos: pessoas, locais ou organizações.

A partir desse momento, o Reconhecimento de Entidades Nomeadas (REN) passou a ser contemplado como uma das tarefas mais recorrentemente utilizadas no tratamento computacional dos textos em línguas naturais. As principais razões para a importância atribuída ao tema residem em seu potencial heurístico enquanto parte da solução de uma série de tarefas: muito do conhecimento enciclopédico presente em qualquer texto, a respeito de virtualmente qualquer assunto, está associado a entidades; na busca automática por respostas em dados textuais, as entidades nomeadas correspondem a uma grande parte da informação pesquisada em motores de busca, seja de um simples catálogo bibliográfico local até uma busca no Google ou em outros grandes sistemas; em algoritmos de representação do conhecimento por regras gramaticais, nos quais são atribuídos papéis semânticos aos termos (por exemplo, "sujeito-agente", "objeto-

beneficiário" etc.), as entidades nomeadas podem servir de filtro de pertinência (pessoas ou organizações tendem a ser agentes, por exemplo).

Como todas as tarefas propostas ao tratamento computacional de dados linguísticos, o REN passou por diferentes propostas de solução, inicialmente com sistemas baseados em regras ligadas a traços distintivos de busca no texto (letras maiúsculas, presença de palavras indicadoras de classes de entidades como "Rua", "Sr.", "Ltda." etc.) a classificadores probabilísticos baseados em aprendizado de máquina. De forma geral, os sistemas baseados em regras funcionam bem quando não há *corpora* previamente anotados para o aprendizado de máquina de tipo supervisionado (aquele em que os dados são previamente categorizados, idealmente por um humano, e a máquina deve "compreender" a associação entre o dado e a categoria anotada) ou quando há relativamente poucos dados nesses *corpora*. O desempenho dos sistemas baseados em aprendizado de máquina, por sua vez, tende a ser melhor quando há um *corpus* com grande quantidade de material anotado para o treinamento do classificador. Nos anos recentes, os resultados obtidos com a introdução dos modelos de aprendizado de máquina baseados em redes neurais profundas (isto é, com múltiplas camadas de representação) passaram a apresentar diferenças de desempenho cada vez mais acentuadas por relação aos sistemas de representação de conhecimento por regras, levando a questionamentos frequentes sobre a relevância ou mesmo a utilidade da codificação de traços linguísticos na solução da tarefa.

Nesse sentido, o objetivo deste trabalho é avaliar em que medida a incorporação de atributos linguísticos poderia impactar o desempenho de modelos atuais de redes neurais profundas. O conhecimento linguístico incorporado ao modelo é simples e faz parte da competência de qualquer falante do português, constituindo-se basicamente de regras morfossintáticas e da valoração contextual dos pronomes de tratamento para a classificação das entidades nomeadas. Tais atributos foram combinados a uma rede neural BiLSTM com uma camada de saída CRF, formando um modelo híbrido para o Reconhecimento de Entidades Nomeadas em português.

Além da avaliação da contribuição potencial dos atributos linguísticos, a importância de se ter um modelo híbrido está ligada à possibilidade de ajuste do desempenho em função do conjunto de regras que permite a representação de intuições linguísticas de diversos tipos (ortográficas, morfossintáticas e lexicais) a respeito do reconhecimento das entidades, sem deixar de lado os modelos de aprendizado de máquina que apresentam boa acurácia quando lidam com dados complexos, de difícil formulação com regras.

O modelo proposto foi treinado no *corpus* do Primeiro Harem e testado no Mini-Harem (SANTOS; CARDOSO, 2007), ambos bastante presentes na literatura sobre o tema em língua portuguesa. Os resultados foram comparados com modelos gerados pelo módulo spaCy (HONNIBAL; MONTANI, 2017), muito usado no processamento de dados

do português, e com o modelo BERTimbau, uma rede neural pré-treinada por Souza, Nogueira e Lotufo (2020).

Os resultados mostram que o modelo híbrido aqui proposto supera por larga margem o desempenho do módulo spaCy no *corpus* de teste e por pequena margem o modelo BERTimbau, o que permite argumentar em favor da relevância de regras que representem o conhecimento linguístico na tarefa de Reconhecimento de Entidades Nomeadas.

Revisão da Literatura Ligada ao REN em Português

O Reconhecimento de Entidades Nomeadas (REN) é uma tarefa que envolve, basicamente, dois procedimentos. Recebendo como entrada uma sequência de palavras, o algoritmo deve identificar quais delas nomeiam entidades; as demais, que formam a maioria das palavras nos textos convencionais, são etiquetadas como "O", de "Out". Já as palavras identificadas como entidades devem ser etiquetadas em um número pré-definido de categorias, sendo as mais comumente usadas Pessoa, Local e Organização. Além disso, as palavras associadas à nomeação são etiquetadas posicionalmente com B (para o começo do nome da entidade) ou I (para as posições seguintes). Assim, como exemplo, ao receber a sequência textual "O novo disco de Milton Nascimento", espera-se que o algoritmo forneça ("O", O), ("novo", O), ("disco", O), ("de", O), ("Milton", B-Pessoa), ("Nascimento", I-Pessoa).

As primeiras tentativas de modelagem computacional para o REN eram baseadas na elaboração manual de regras para a identificação e a classificação das Entidades Nomeadas (EN). As regras podem contemplar aspectos ortográficos, como a identificação de letra inicial maiúscula na palavra, linguísticos, por meio de análises no nível morfológico, sintático, semântico etc., ou ligados à frequência das palavras em *corpus*. Um exemplo de sistema desse tipo é o PALAVRAS-NER (BICK, 2006), um célebre programa baseado em regras desenvolvido para a língua portuguesa. Ele classifica as entidades através de módulos de análise morfológica, sintática e semântica do texto, também considerando padrões contextuais para desambiguação de sentido. Apesar de obterem boa precisão na tarefa, modelos baseados em regras tendem a ser muito aderentes ao *corpus* para o qual foram desenvolvidos, sendo assim difíceis de se adaptar a novos domínios de aplicação.

Atualmente, os modelos baseados em aprendizado de máquina tendem a ser a opção mais adotada na tarefa. Os algoritmos de aprendizado de máquina, em geral, são baseados em modelos estatísticos que mapeiam traços extraídos das palavras nas respectivas classificações. Tais modelos são divididos em tipos de aprendizado: supervisionados, semi-supervisionados e não-supervisionados. Aqui, serão focados os métodos de aprendizado supervisionado, por serem os mais utilizados no reconhecimento de entidades. Neles, o algoritmo aprende a classificar as palavras a partir de um *corpus*

anotado com uma classificação prévia para cada uma delas. Alguns algoritmos comuns para o REN são *Conditional Random Field* (CRF), *Support-Vector Machine* (SVM) e muitas das redes neurais.

Assim como nos modelos de regras, no aprendizado supervisionado clássico é necessário definir *a priori* os traços que serão utilizados para representar as palavras. No trabalho de Ratinov e Roth (2009), os autores analisam diversos tipos de traços para a tarefa, entre eles, a codificação de conhecimento externo usando *gazetteers*, isto é, bancos de dados contendo exemplos de entidades para cada categoria alvo (por exemplo, uma lista de nomes próprios de pessoas). Os *gazetteers* são recursos amplamente utilizados no REN e podem trazer uma melhoria no desempenho dos modelos (RATINOV; ROTH, 2009; CHIU; NICHOLS, 2016).

Nos últimos anos, as redes neurais profundas têm ganhado espaço entre os modelos baseados em aprendizado supervisionado, tendo alcançado resultados destacados em inúmeras tarefas, incluindo o REN. As redes neurais são treinadas para extrair padrões a partir dos dados sem que um humano precise indicar que características linguísticas o modelo deve observar. Desse modo, é difícil para um humano identificar o que o modelo aprendeu, se foram traços morfossintáticos, semânticos etc., de forma que a rede acaba funcionando como uma espécie de "caixa preta".

Outro desafio relacionado à análise de dados linguísticos pelas redes neurais está relacionado às formas de representar as expressões da língua natural para o processamento computacional. As redes neurais operam exclusivamente com dados numéricos. Para codificar informações textuais em números, foram propostas técnicas como os chamados *word embeddings*, que mapeiam as palavras em vetores numéricos. Recentemente, tem havido bastante interesse em investigar a combinação de outros métodos de representação de palavras em redes neurais para melhorar o desempenho no REN. Chiu e Nichols (2016) combinam *word embeddings*, *character embeddings* (traços baseados em ortografia) e *gazetteers* para alimentar um tipo de rede neural, a LSTM.

Tratando especificamente do Reconhecimento de Entidades Nomeadas para o português, Santos e Guimarães (2015) treinaram uma rede neural no *corpus* Harem (SANTOS; CARDOSO, 2007) que recebe como entrada representações *word embeddings* somadas a dois traços manuais: um ortográfico e, outro, baseado no sufixo da palavra. O modelo obteve 71,23% de medida-F na avaliação de cinco categorias de entidades: Pessoa, Local, Organização, Tempo e Valor. Mais recentemente, Souza, Nogueira e Lotufo (2020) treinaram o modelo BERT (DEVLIN *et al.*, 2018) para o português, chamando-o de BERTimbau[3]. Os autores refinaram o treinamento do modelo para a tarefa de REN utilizando o Harem como *corpus*, obtendo 83,24% de medida-F na avaliação de

3 https://github.com/neuralmind-ai/portuguese-bert

cinco categorias de entidades, seguindo o trabalho de Santos e Guimarães (2015). É importante notar, entretanto, que os autores aplicam ao modelo uma etapa adicional de pós-processamento, ou seja, posterior à classificação propriamente dita, na qual as etiquetagens inconsistentes de entidades (por exemplo, quando o algoritmo anota o começo de uma entidade (B) depois de uma etiqueta I para a mesma entidade) são automaticamente reclassificadas para a categoria Out, que tem a maior probabilidade de ocorrência *a priori*.

Embasamento teórico

Uma das inspirações para a modelagem dos *word embeddings* vem da ideia de distribuição contextual das palavras, amplamente defendida na Semântica Distribucional. A Semântica Distribucional foi fundamentalmente inspirada por trabalhos da década de 1950 em Linguística Distribucional, que teve entre seus autores mais destacados o nome de Zellig Harris. Em um de seus artigos mais célebres, Harris (1954) discute a distribuição estrutural nas línguas postas em uso, fazendo uma análise a partir da associação entre a ocorrência de um elemento linguístico (por exemplo, um fonema) em relação a outros. Segundo o autor, a distribuição de um elemento é dada pela soma dos ambientes em que ele ocorre, sendo estes representados pelos demais elementos que coocorrem em determinada ordem com o elemento sob análise. Harris vai além, argumentando que o significado de um elemento pode ser dado em função de sua distribuição. Nesse sentido, elementos com uma distribuição semelhante tendem a ter significados semelhantes. Para citar um exemplo, os adjetivos "belo" e "bonito" aparecem frequentemente em contextos parecidos, sendo substituídos um pelo outro sem muita perda de significado em grande parte dos casos. Portanto, membros de uma mesma categoria distribucional (por exemplo, adjetivos, pronomes etc.) devem, analogamente, receber representações contextuais em comum. Generalizando, quanto mais parecida a distribuição, mais características as expressões sob análise compartilhariam.

Muitas dessas ideias viriam a compor os fundamentos do que posteriormente ficou conhecido como Semântica Distribucional. Alguns desses trabalhos foram aproveitados em um domínio específico, o PLN, que tinha interesse em avaliar a aplicação dessas análises em dados linguísticos. Um dos resultados dessas investigações foram as representações *word embeddings*, que trouxeram inúmeros avanços para o processamento computacional de línguas humanas.

Diversas técnicas são adotadas para gerar *word embeddings*. Algumas das mais comuns são Word2Vec (MIKOLOV *et al.*, 2013), GloVe (PENNINGTON; SOCHER; MANNING, 2014) e BERT (DEVLIN *et al.*, 2018). Entre esses, o BERT é um tipo de rede neural profunda capaz de gerar representações *word embeddings* contextuais, em que se representa uma palavra a partir dos contextos em que ela ocorre nos textos. Por comparação, os métodos de *word embeddings* não-contextuais, como Word2Vec e GloVe, funcionam como um conjunto

estático de representações de palavras, descontextualizado do texto de ocorrência da palavra. Já o BERT cria a representação da palavra representando, igualmente, o contexto em que ela ocorreu. Uma das vantagens do método é que ele é capaz de desambiguar os diversos sentidos que uma palavra pode ter, a partir das pistas fornecidas pelo contexto.

A importância do contexto para o REN

Para classificar as entidades nomeadas, é importante codificar as pistas contextuais contidas no próprio texto. Aqui, "contexto" refere-se a palavras coocorrendo à direita e à esquerda de uma determinada palavra-alvo, considerando-se os limites de uma sentença. A partir do contexto é possível extrair diversas informações sobre a categoria de uma entidade. Por exemplo, é comum encontrar pronomes de tratamento ou suas abreviações, como "Sr." e "Dr.", precedendo o nome de uma pessoa, o que pode ser codificado como uma pista de que aquela entidade pertence à categoria Pessoa.

Além disso, as informações contextuais são essenciais para desambiguar entidades possuidoras da mesma forma superficial, mas classificações distintas, como nos exemplos (1) e (2).

1. A cidade de São Paulo sofre com a poluição do ar.

2. O São Paulo ganhou o jogo de domingo.

Em (1), "São Paulo" se refere a uma cidade e, em (2), a um time de futebol. Em um sistema de REN, a palavra "cidade" pode ajudar a identificar que ali se trata de um Local e a palavra "jogo" pode ajudar na classificação da entidade como Organização. Essas palavras, portanto, podem ser utilizadas em um modelo de REN como traços classificadores para essas categorias.

Essa análise vai além de entidades isoladas, podendo se estender a categorias de entidades. Com efeito, observando-se a distribuição contextual dos tipos de entidades, é possível identificar contextos que coocorrem frequentemente com determinados tipos de EN. No exemplo (1), a entidade "São Paulo" poderia ser substituída por outros nomes de cidades e, em (2), outros nomes de times de futebol caberiam. Assim, identifica-se um critério que pode ser aplicado como regra em um modelo de classificação de entidades.

Assim como o conhecimento dos usos linguísticos mostram a importância das pistas contextuais para a categorização de entidades, os *word embeddings*, que buscam codificar informações da distribuição de uma palavra, são igualmente relevantes para a resolução da tarefa de REN. Nos estudos de Seok *et al.* (2016), a utilização de *word embeddings* resultou em melhoras no desempenho de modelos para REN em relação aos mesmos modelos considerando somente traços ortográficos, morfossintáticos e

palavras isoladas. Além disso, os autores mostraram que *word embeddings* representando entidades da mesma categoria tendem a aparecer agrupados. No trabalho de Augenstein, Derczynski e Bontcheva (2017), os autores ressaltam que os *word embeddings* podem ajudar a lidar com o problema de generalização de aprendizado de máquina, capturando traços para a classificação de exemplos desconhecidos na etapa de treinamento do modelo. Parece consensual na literatura, portanto, que os *word embeddings* são úteis para a representação de palavras na tarefa de REN.

Metodologia

Neste trabalho, adotamos como modelo de classificação uma rede neural LSTM bidirecional, com uma camada de saída CRF, abreviado como BiLSTM-CRF. Para testar o impacto da representação das palavras no desempenho, treinamos duas versões do modelo: (I) somente com representações *word embeddings* e (II) com *word embeddings* combinados a traços linguísticos. O desempenho do modelo foi avaliado em comparação a duas propostas para REN em português: spaCy (HONNIBAL; MONTANI, 2017) e BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020).

Corpus

O Primeiro Harem foi utilizado como *corpus* de treinamento. O Mini-Harem serviu como *corpus* de teste. Ambos fazem parte das Coleções Douradas (CD) do Harem (SANTOS; CARDOSO, 2007). As CDs são compostas por documentos extraídos de diversos gêneros textuais, como jornalístico, *web* e oral, originários principalmente do português do Brasil e de Portugal. O Primeiro Harem contém 129 documentos e o Mini-Harem, 128. Os *corpora* foram anotados manualmente por especialistas para dez categorias de entidades, a saber: Pessoa, Local, Organização, Tempo, Valor, Acontecimento, Coisa, Abstração, Obra e Outro. Cada categoria é dividida em um conjunto de subcategorias. Por exemplo, a categoria Tempo se subdivide em Data, Hora, Período e Cíclico.

Para fins desta pesquisa, consideramos somente cinco categorias para o treinamento e a avaliação do modelo: Pessoa, Local, Organização, Tempo e Valor. As subcategorias não foram consideradas. Além disso, os documentos foram divididos em sentenças delimitadas pelos símbolos de ponto final, interrogação, exclamação e ponto e vírgula. A Tabela 1 traz as estatísticas de cada *corpus* após o pré-processamento.

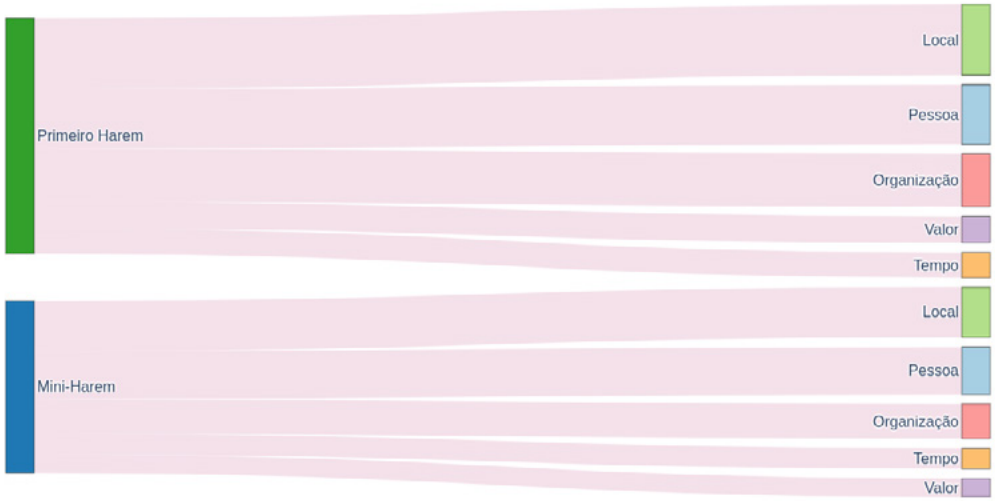
Tabela 1. Número de sentenças, *tokens* e entidades no Primeiro Harem e Mini-Harem

Corpus	**Sentenças**	***Tokens***	**Entidades Nomeadas**	**Maior sentença**	**Menor sentença**
Primeiro Harem	4482	94356	4081	226	1
Mini-Harem	3092	63650	2989	581	1

Fonte: Elaboração própria

Há um total de 7.070 entidades nomeadas, sendo 4.081 do Primeiro Harem e 2.989 do Mini-Harem. A Figura 1 traz uma distribuição dessas entidades por categoria em cada um dos *corpora*.

Figura 1. Distribuição das Entidades Nomeadas nos *corpora* Primeiro Harem e Mini-Harem



Fonte: Elaboração própria

A distribuição de exemplos por categoria não é balanceada em nenhum dos *corpora*. Contudo, ambos seguem padrões de distribuição semelhantes, com Local, Pessoa e Organização sendo as categorias com mais ocorrências.

Analisando o Harem mais de perto, surgem algumas questões sobre as motivações para determinados tipos de anotação, como mostram os exemplos (3) e (4) extraídos do *corpus*.

3. […] depois teve os outros aqui, minha {tia Rosinha, PESSOA}, meu {tio Pedro, PESSOA} […]

4. […] como dizia seu primo {Alves Redol, PESSOA} […]

Na sentença (3), os anotadores indicaram "tia Rosinha" e "tio Pedro" como sendo entidades únicas, isto é, incluindo o nome de parentesco e o nome próprio. Na sentença (4), contudo, "primo" não é considerado parte da entidade, o que não condiz com a anotação feita no exemplo anterior. Encontram-se diversos casos semelhantes no *corpus*, considerando pronomes de tratamento, entre eles, "senhora" e "professor", e alguns topônimos, como "rio" e "ilha", como parte da EN. Tais inconsistências da anotação acabam confundindo os modelos computacionais, dificultando na identificação das fronteiras de início desses tipos de entidades.

BiLSTM-CRF

A arquitetura do modelo BiLSTM-CRF é composta por uma rede neural com uma camada de entrada, duas camadas de rede LSTM e uma camada de saída CRF. Umas das camadas LSTM processa os dados da esquerda para a direita (*forward)* e a outra o faz da direita para esquerda (*backwards)*. A BiLSTM-CRF foi alimentada com as representações *word embeddings* das palavras, que foram combinadas com traços baseados em aspectos linguísticos. Os *word embeddings* foram obtidos a partir do modelo base do BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), sem passar por nenhum outro treinamento.

Foram avaliados três tipos de traços linguísticos: ortográficos, morfossintáticos e lexicais. Os traços ortográficos identificam dois aspectos: forma e tipo. O tipo classifica uma palavra em alfabética (composta somente de letras), alfanumérica (combinação de letras e números), numérica e não-alfanumérica (tais como palavras contendo pontuações e símbolos). A forma ortográfica contempla estas possibilidades: primeira letra maiúscula, todas as letras maiúsculas, todas minúsculas e outros casos (como "CNPq", "spaCy" etc.).

Já o traço morfossintático classifica uma palavra a partir de sua categoria morfossintática (substantivo, adjetivo, pronome etc.), obtida com a ferramenta de POS-tagging da biblioteca nlpnet[4] (FONSECA; ROSA, 2013).

Por fim, consideramos um traço lexical inferido a partir do contexto de ocorrência de uma entidade. Esse traço foi codificado a partir de um conjunto de listas de palavras associadas a categorias de entidades, usadas como pistas para sua classificação. Para exemplificar, selecionamos uma lista de pronomes de tratamento para a classificação de pessoas. Esse traço é ativado se a palavra imediatamente à esquerda de uma EN é um pronome de tratamento, como em "senhor José", "Dra. Paula" etc. A Tabela 2 apresenta uma descrição das categorias e subcategorias dessas listas.

4 http://nilc.icmc.usp.br/nlpnet/

Tabela 2. Descrição das listas de palavras usadas para a classificação de entidades

Categoria	Subcategoria	Exemplos	Palavras
Pessoa	parentesco	prima, tio	64
	pronome de tratamento	senhora, excelência	63
	profissão	professora, reitor	743
Local	logradouro	rua, avenida	45
	topônimo	floresta, ilha	26
Organização	organização	companhia, farmácia	80
Total			1021

Fonte: Elaboração própria

Os traços lexicais foram projetados somente para três das cinco categorias (Pessoa, Local e Organização), totalizando 1.021 palavras classificadoras. Cada uma das subcategorias é uma lista de palavras gerada manualmente, podendo conter a mesma palavra em sua forma no singular e no plural. Quando uma palavra no plural é identificada no contexto, mais de uma entidade pode ser classificada pelo traço. Como ilustração, na sentença "Os professores Marcelo e João estão ministrando a disciplina", "professores" está classificando "Marcelo" e "João", o que faz com que ambos sejam detectados como Pessoa pelo traço. O contexto plural só anota mais de uma entidade se encontrar uma sequência de duas ou mais entidades consecutivas cujo último conectivo da lista seja "e". Em casos como "os supermercados X estão com promoções", mesmo que exista a interpretação de plural, há somente uma entidade sendo classificada pelo traço.

Treinamento e avaliação dos modelos

Durante os experimentos, avaliamos diversas combinações de parâmetros para a rede BiLSTM-CRF. Ao final, a arquitetura da rede ficou com 256 neurônios em cada uma das camadas LSTM e o modelo foi treinado por 30 épocas, com uma taxa de aprendizado de 0,001. Para lidar com possíveis problemas de *overfitting*, aplicamos a técnica de *dropout* sobre as camadas de entrada e saída, com uma taxa de 0,2, e nas camadas LSTM, uma taxa de 0,6. A validação utilizada foi *10-fold validation*, que envolve o treinamento do modelo por dez rodadas distintas, embaralhando os exemplos do conjunto de treino. Os resultados finais são dados pela média dos resultados obtidos.

A BiLSTM-CRF foi treinada em duas configurações: (I) com cinco categorias (Pessoa, Local, Organização, Tempo e Valor) e (II) com três categorias (Pessoa, Local e Organização). O segundo cenário foi avaliado para comparação de desempenho da BiLSTM-CRF com

a ferramenta spaCy, que só classifica essas três categorias[5] no modelo em português, e com o BERTimbau, disponível na biblioteca transformers[6]. No caso do BERTimbau, fizemos o treinamento do modelo no Primeiro Harem, utilizando 30 épocas e uma taxa de aprendizado de 0,00005. Os modelos treinados no segundo cenário (BERTimbau e BiLSTM-CRF) foram validados em somente cinco rodadas.

As implementações foram criadas em linguagem Python[7], baseadas nos módulos de aprendizado de máquina scikit-learn (PEDREGOSA *et al.*, 2011)[8], TensorFlow (ABADI *et al.*, 2015)[9], keras (CHOLLET *et al.*, 2015)[10] e PyTorch (PASZKE *et al.*, 2017)[11], executadas nas GPUs da plataforma Google Colab. O tempo médio de treinamento dos modelos foi de 25 minutos.

Resultados

A Tabela 3 traz os resultados comparando a rede neural somente com os *word embeddings* (BiLSTM-CRF) com a rede que utiliza *word embeddings* combinados aos traços linguísticos (BiLSTM-CRF+traços).

Tabela 3. Resultados obtidos pela BiLSTM-CRF com e sem os traços linguísticos

Modelo	Precisão	Cobertura	Medida-F
BiLSTM-CRF+traços	81,95	**80,19**	**81,06**
BiLSTM-CRF	**82,23**	79,17	80,66

Fonte: Elaboração própria

O desempenho geral (medida-F) obtido pela BiLSTM-CRF+traços foi melhor, ficando um pouco abaixo da BiLSTM-CRF somente na precisão (que é um dos fatores componentes da medida-F). Aplicando o teste estatístico U de Mann-Whitney sobre os valores de medida-F, obtivemos que a diferença entre os modelos é significativa ($n = 10$; $p = 0{,}444$; $\alpha = 0{,}05$). Isso mostra que a codificação de traços linguísticos voltados especificamente

5 Há também uma categoria composta de entidades mistas (Miscellaneous) que não foi considerada na avaliação.

6 https://huggingface.co/transformers/

7 https://www.python.org/

8 https://scikit-learn.org/

9 https://www.tensorflow.org/

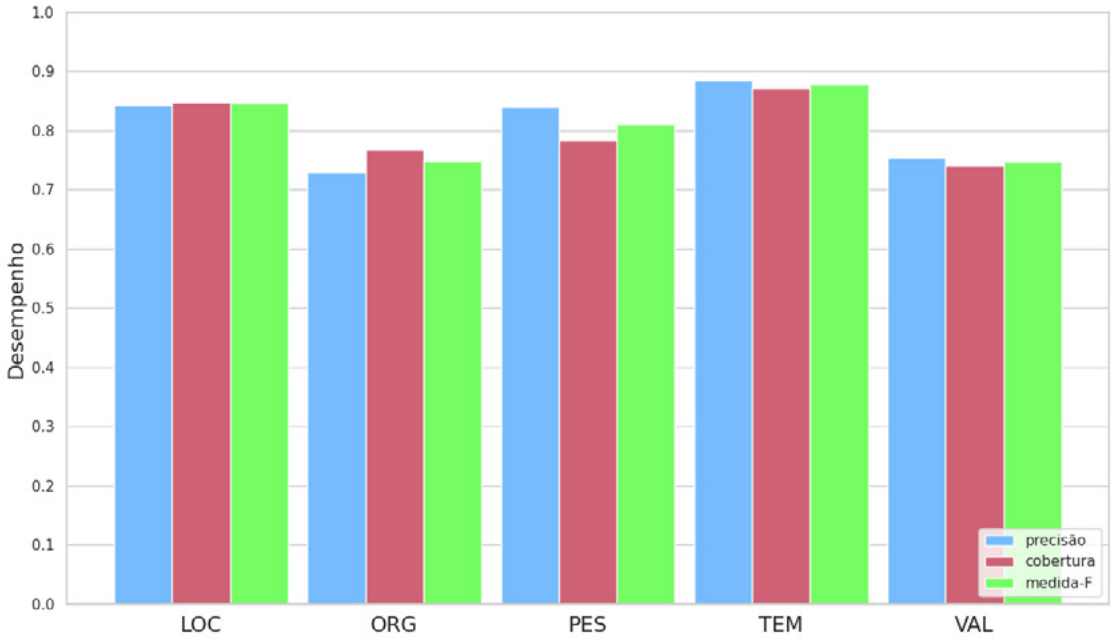
10 https://keras.io/

11 https://pytorch.org/

para uma tarefa pode gerar representações de palavras mais ricas quando combinadas a *word embeddings* e melhorar o desempenho de modelos neurais para o REN.

Na Figura 2, está o desempenho da BiLSTM-CRF+traços por categorias de EN. Entre as cinco categorias, aquelas com os melhores resultados foram Tempo, Local e Pessoa. No caso de Tempo, há grande regularidade nos formatos de entidades encontradas no *corpus*, tais como "01 de janeiro", "1995", o que facilita o aprendizado do modelo e se reflete diretamente na alta precisão obtida na categoria. Já Organização e Valor ficaram quase empatadas, com cerca de 75% de medida-F.

Figura 2. Desempenho da BiLSTM-CRF+traços em uma única rodada



Fonte: Elaboração própria

Com relação à Organização, vale dizer que o número de exemplos não foi decisivo para o desempenho nessa categoria, já que Tempo tem menos exemplos no *corpus* e apresentou melhores resultados. Diversos artigos na literatura (JIANG; BANCHS; LI, 2016; AUGENSTEIN; DERCZYNSKI; BONTCHEVA, 2017) já trataram Organização como a categoria de classificação mais difícil na tríade Pessoa, Local e Organização. Isso pode estar relacionado a uma dificuldade de extrair padrões contextuais para a ocorrência de organizações e à grande variabilidade lexical, ortográfica e morfossintática desses nomes, que podem se realizar como siglas ("IBM"), nomes comuns usados como próprios ("Apple"), sintagmas nominais compostos de nomes comuns e próprios ("Fundação Padre Anchieta") e assim por diante. Assim, a expressão dessas entidades costuma ser bem heterogênea, com ocorrências raras e específicas em cada *corpus*.

Para uma análise de erros do modelo, trazemos alguns dos exemplos de classificação de entidades pela BiLSTM-CRF+traços nos exemplos (5) e (6) extraídos do *corpus* de teste.

5. A partir de {1880, TEM} ensinou psicologia e filosofia em {Harvard, ORG}, universidade que abandonou em {1907, TEM}, proferindo conferências nas universidades de {Columbia, LOC} e {Oxford, LOC}.

6. Estive em casa do {Pinto, LOC}, em {Brandim, LOC} e depois em casa do {Miranda, LOC} cá em {Parafita, LOC}, estive lá muito anos.

No exemplo (5), o modelo classificou corretamente "1880", "Harvard" e "1907". Já "Columbia" e "Oxford" foram incorretamente classificadas como "Local", quando na verdade era esperada a categoria "Organização". Porém, o contexto em que essas entidades ocorrem pode ser interpretado como locativo, dado que se refere ao lugar em que as conferências foram realizadas. Portanto, esses são casos complexos para a anotação, já que são ambíguos para classificação mesmo por humanos.

Em (6), duas entidades foram classificadas corretamente ("Brandim" e "Parafita") e duas incorretamente ("Pinto" e "Miranda"). Nos erros, esperava-se a classificação de "Pessoa" em vez de "Local". É interessante que o modelo tenha classificado todas as entidades encontradas como lugares, o que talvez esteja relacionado à ocorrência da preposição "em" precedendo cada uma delas, o que pode ser uma pista forte de contextos locativos.

Na Tabela 4 estão os resultados dos modelos testados no cenário II.

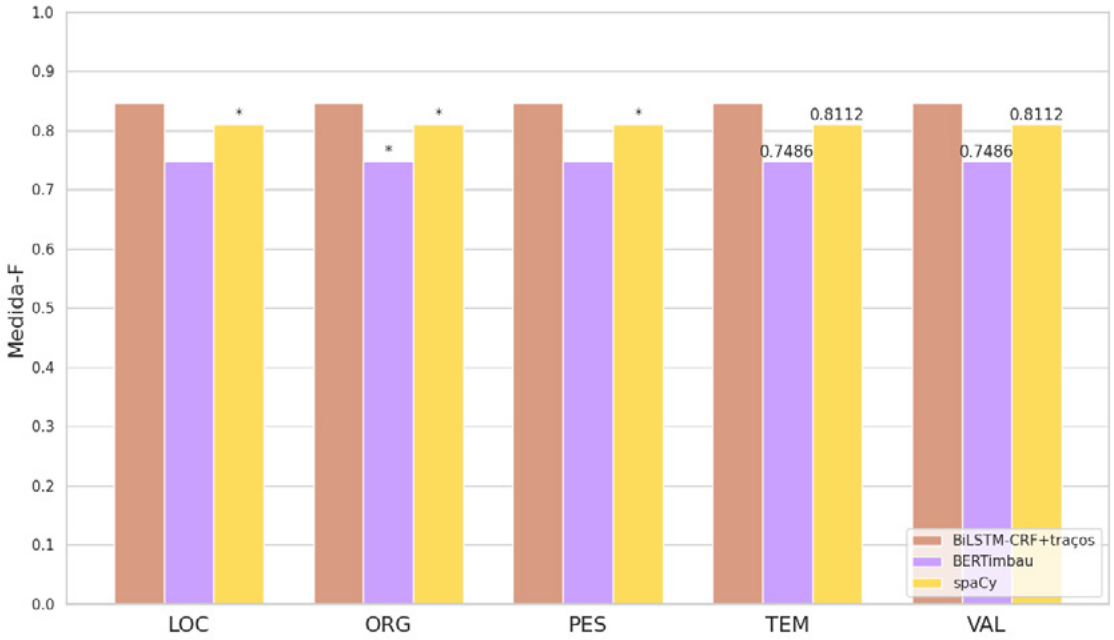
Tabela 4. Comparação de desempenho entre BERTimbau, BiLSTM-CRF+traços e spaCy

Modelo	Precisão	Cobertura	Medida-F
BERTimbau	80,00	78,94	79,47
BiLSTM-CRF+traços	81,56	79,12	80,31
spaCy	55,77	61,68	58,57

Fonte: Elaboração própria

A BiLSTM-CRF+traços superou os resultados tanto do spaCy quanto do BERTimbau. No caso do segundo, a diferença foi menor, mas cabe ressaltar que a precisão da BiLSTM-CRF+traços supera em 1,56% a do BERTimbau. Na Figura 3, compara-se o desempenho dos três modelos por categoria. O símbolo * acima da barra indica que há diferença significativa (teste de Mann-Whitney, com $\alpha = 0,05$) entre os desempenhos dos modelos naquela categoria.

Figura 3. Comparação de desempenho dos modelos por categoria de EN



Fonte: Elaboração própria

Nas categorias Local e Pessoa, o BERTimbau e a BiLSTM-CRF+traços estão quase empatados, sem diferença significativa no desempenho. Para a categoria Organização, a medida-F da BiLSTM-CRF+traços foi significativamente maior que nos demais modelos. Ademais, essa foi a categoria com os resultados mais baixos e com maior variação nos três modelos.

Comparando o desempenho de BiLSTM-CRF+traços nos cenários (I) e (II), vemos que o modelo saiu-se melhor quando avaliado com cinco categorias, com uma medida-F de 81,06%. Em parte, isso ocorre em função do desempenho da categoria Tempo, com uma medida-F individual de quase 90%, o que puxa a média de desempenho para cima.

Conclusões

Este trabalho comparou resultados de diferentes abordagens para o Reconhecimento de Entidades Nomeadas com dados de língua portuguesa. Foram apresentados testes com dois modelos já existentes, spaCy e BERTimbau, além de um novo modelo, um híbrido composto por uma rede BiLSTM com uma camada de saída CRF e representação de traços linguísticos. De forma global, os resultados mostraram a prevalência do terceiro modelo sobre os demais. Além disso, o efeito da camada de representação dos traços linguísticos na rede foi testado isoladamente. Diferenças significativas foram encontradas no desempenho do modelo enriquecido (BiLSTM-CRF+traços), o que permite afirmar que os traços linguísticos aprimoraram os resultados.

Atualmente, muitos debates vêm sendo travados acerca do uso de modelos pré-treinados em tarefas gerais e que demonstram excelente desempenho em tarefas específicas já de saída, ou seja, prescindindo de ajustes finos para a execução de tais tarefas, o que, por consequência, significaria também deixar de lado a representação do conhecimento linguístico específico ligado à realização das tarefas (BOMMASANI *et al*., 2021). Os resultados obtidos no presente trabalho pelo modelo que incorpora traços linguísticos, entretanto, indicam a direção inversa, isto é, a importância da representação do conhecimento linguístico nos modelos.

As representações *word embeddings* são capazes de representar traços contextuais linguísticos. Entretanto, a informação contextual é ampla demais quando não existem regras linguísticas de seleção, pois, nesse caso, toda palavra na vizinhança da entidade nomeada seria candidata a contexto relevante. Os traços linguísticos aqui contemplados representam, sobretudo, informações específicas de caráter contextual (regras morfossintáticas, anteposição de pronomes de tratamento a lexemas de certas categorias etc.). Portanto, a contribuição das regras é justamente essa, a de filtrar possibilidades por meio da aplicação de paradigmas lexicais e morfossintáticos simples, facilmente reconhecidos e utilizados pelos falantes da língua.

Por fim, percebe-se uma diferença para menos na classificação das Organizações quando comparada ao desempenho das demais categorias pelo modelo. A razão provável é a dificuldade de encontrar padrões na nomeação dessas entidades, cujos nomes variam em extensão (de uma a dez palavras) e em padrões de escrita (organizações admitem siglas, nomes inventados e outras idiossincrasias). Além disso, nomes de pessoas podem aparecer como nomes de Organizações, mas o inverso é mais improvável. Por sua vez, é muito comum que a categoria Local receba nomes de pessoas, mas a presença dos lexemas marcadores de logradouros e de preposições ligadas à espacialidade (“em”, “na”, “até”...) contribuem para a classificação dessa categoria, sendo que não existem equivalentes igualmente fortes para as organizações. Aqui, novamente, impõe-se a constatação de que traços da representação linguística das categorias são decisivos para a classificação das entidades nomeadas.

O código-fonte completo para treinamento e avaliação dos modelos implementados está disponível na plataforma GitHub no endereço https://github.com/andressa-vs/named-entity-recognition-pt.

Limites do trabalho atual

A opção de treinar e avaliar nossos modelos a partir dos *corpora* Harem e Mini-Harem deve-se, antes de tudo, às possibilidades de comparação com outros estudos na literatura sobre REN em português. Como consequência imediata, os resultados obtidos no trabalho atual estão restritos, portanto, às avaliações com esses dados. Ambos os

corpora possuem características próprias de etiquetagem (como os diferentes critérios anteriormente apresentados nos exemplos (3) e (4); a presença de categorias muito pouco representadas, como "Obra", "Acontecimento" e "Outro", que não são comumente encontradas em outros *corpora* para REN) e de pequeno volume de dados etiquetados disponíveis, características estas capazes de inibir a generalização dos resultados com outros conjuntos de dados. Como diferentes *corpora* utilizam-se de outras categorizações de entidades, que diferem quanto ao número de categorias consideradas e também quanto à informação a categorizar, é difícil prever a qualidade do desempenho dos modelos com esses outros conjuntos de dados. Não obstantes essas ressalvas, será futuramente interessante treinar e avaliar os modelos com outros *corpora* do português.

REFERÊNCIAS

ABADI, M. *et al*. *TensorFlow:* Large-scale machine learning on heterogeneous systems. 2015. Disponível em: https://www.tensorflow.org/. Acesso em: 19 ago. 2021.

AUGENSTEIN, I.; DERCZYNSKI, L.; BONTCHEVA, K. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, v. 44, p. 61-83, jul. 2017. Disponível em: https://doi.org/10.1016/j.csl.2017.01.012. Acesso em: 31 ago. 2021.

BICK, E. Functional aspects in Portuguese NER. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 80-89. Disponível em: http://dx.doi.org/10.1007/11751984_9. Acesso em: 19 ago. 2021.

BOMMASANI, R. *et al*. On the Opportunities and Risks of Foundation Models. *arXiv preprint*. arXiv:2108.07258. 2021. Disponível em: https://arxiv.org/abs/2108.07258. Acesso em 20 ago. 2021.

CHIU, J. P. C.; NICHOLS, E. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, v. 4, p. 357-370, dez. 2016.

CHOI, Y.; CHA, J. Korean Named Entity Recognition and Classification using Word Embedding Features. *Journal of KIISE*, v. 43, n. 6, p. 678-685, 15 jun. 2016.

CHOLLET, F. *et al*. *Keras*. Disponível em: https://keras.io. Acesso em: 19 ago. 2021.

DEVLIN, J. *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*. arXiv:1810.04805. 2018. Disponível em: https://arxiv.org/abs/1810.04805. Acesso em: 29 ago. 2021.

DOS SANTOS, C.; GUIMARÃES, V. Boosting named entity recognition with neural character embeddings. 2015a, Stroudsburg, PA, USA: *Association for Computational Linguistics*, 2015. Disponível em: http://dx.doi.org/10.18653/v1/w15-3904. Acesso em: 19 ago. 2021.

FONSECA, E. R.; ROSA, J. L. G. Mac-Morpho revisited: towards robust part-of-speech tagging. *In*: *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*. STIL, 2013.

HARRIS, Z. S. Distributional structure. *WORD*, v. 10, n. 2-3, p. 146-162, 1954.

HONNIBAL, M.; MONTANI, I. Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *spaCy*. 2, 2017. Disponível em: https://spacy.io/. Acesso em: 29 ago. 2021.

JIANG, R.; BANCHS, R. E.; LI, H. Evaluating and Combining Name Entity Recognition Systems. 2016, Stroudsburg, PA, USA: *Association for Computational Linguistics*, 2016. Disponível em: http://dx.doi.org/10.18653/v1/w16-2703. Acesso em: 29 ago. 2021.

MIKOLOV, T. *et al.* Efficient Estimation of Word Representations in Vector Space. *arXiv preprint*. arXiv:1301.3781. 2013. Disponível em: https://arxiv.org/abs/1301.3781. Acesso em: 29 ago. 2021.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. *Benjamins Current Topics*. Amsterdam: John Benjamins, 2009. p. 3-28. Disponível em: http://dx.doi.org/10.1075/bct.19.03nad. Acesso em: 29 ago. 2021.

PASZKE, A. *et al.* Automatic differentiation in Pytorch. *NIPS 2017 Workshop.* 2017. Disponível em: https://openreview.net/forum?id=BJJsrmfCZ. Acesso em: 29 ago. 2021.

PEDREGOSA, F. *et al*. Scikit-learn: Machine Learning in Python. *JMLR*, v. 12, p. 2825-2830, 2011. Disponível em: https://scikit-learn.org/. Acesso em: 29 ago. 2021.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. 2014, Stroudsburg, PA, USA: *Association for Computational Linguistics*, 2014. Disponível em: http://dx.doi.org/10.3115/v1/d14-1162. Acesso em: 19 ago. 2021.

RATINOV, L.; ROTH, D. Design challenges and misconceptions in named entity recognition. Morristown, NJ, USA: *Association for Computational Linguistics*, 2009. Disponível em: http://dx.doi.org/10.3115/1596374.1596399. Acesso em: 19 ago. 2021.

SANTOS, D.; CARDOSO, N. A Golden Resource for Named Entity Recognition in Portuguese. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 69-79. Disponível em: http://dx.doi.org/10.1007/11751984_8. Acesso em: 29 ago. 2021.

SEOK, M. *et al*. Named entity recognition using word embedding as a feature. *International Journal of Software Engineering and its Applications*, v. 10, n. 2, p. 93-104, 2016.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT models for Brazilian Portuguese. *Intelligent Systems*. Cham: Springer International Publishing, 2020. p. 403-417. Disponível em: http://dx.doi.org/10.1007/978-3-030-61377-8_28. Acesso em: 19 ago. 2021.