Proposta recebida em Março 2023 e aceite para publicação em Julho 2023.

Uma revisão para o Reconhecimento de Entidades Nomeadas aplicado à língua portuguesa

A survey of Named Entity Recogntion applied to Portuguese

Andressa Vieira e Silva
Universidade de São Paulo

Resumo

O Reconhecimento de Entidades Nomeadas (REN) é a tarefa de identificação e classificação automática de entidades em um texto, tais como nomes de pessoas, lugares e organizações. Essa é uma tarefa importante em Processamento de Língua Natural, servindo como base de diversas aplicações, como tradução automática e sistemas de pergunta-e-resposta. Desde seu surgimento na década de 90, a tarefa passou por diversos fases com relação à abordagem computacional, indo dos sistemas baseados em regras manuais aos modelos de redes neurais.

Este artigo traz uma revisão da tarefa de REN considerando aplicações em textos de língua portuguesa. Apresenta-se um panorama geral da tarefa, traçando um histórico das principais iniciativas para promovê-la, dos recursos linguísticos e computacionais disponíveis e das abordagens já avaliadas para REN para o português. Por fim, apresenta-se uma discussão do cenário geral em que a tarefa se encontra e as considerações finais de análise.

Palavras chave

reconhecimento de entidades nomeadas, português

Abstract

Named Entity Recognition is the task of identifying and classifying Named Entities in a text, such as names of people, places and organizations. This is an important task in Natural Language Processing (NLP), serving as the basis for several tasks, such as automatic translation and question answering systems. Since its emergence in the 1990s, the task has gone through several periods in relation to the computational approach, ranging from rule-based systems to the neural network models.

This article presents a review of the NER task considering applications in Portuguese language texts. It presents an overview of the task, tracing a history of the main initiatives to promote the task, the linguistic and computational resources available and the approaches already applied to NER for Portuguese. Finally, a discussion of the general scenario in which the task is and the final analysis considerations is given.

Keywords

named entity recognition, Portuguese

1. Introdução

O Reconhecimento de Entidades Nomeadas[1] (REN) é uma tarefa voltada para a identificação e classificação de termos referentes a entidades em um texto, como nomes de lugares, pessoas, instâncias temporais etc. A tarefa surgiu na década de 90 como um dos tópicos de investigação da 6ª *Message Understanding Conference* (MUC-6) (Grishman & Sundheim, 1996), cujo objetivo era recuperar informações relevantes a partir de dados não-estruturados, como textos jornalísticos.

Desse modo, o REN surgiu como um dos ramos de Extração de Informação (EI) focado para a coleta das entidades de um texto. Diferente de termos com função puramente sintática, como preposições e artigos, as Entidades Nomeadas (EN) fornecem muitas pistas a respeito do conteúdo de um texto, podendo ser usadas para identificar os personagens em um livro, os nomes de países citados em uma notícia, o autor e ano de publicação de um texto etc.

Por essas razões, o reconhecimento de entidades se tornou uma etapa essencial em diversas tarefas de Processamento de Língua Natural (PLN), como tradução automática (Babych & Hartley, 2003; Li et al., 2020b), pergunta-e-resposta (Toral et al., 2005; Mollá et al., 2006) e resolução de correferências entre sintagmas (Dai et al., 2019; Gao et al., 2020). Para citar um exemplo, em tradução automática não é comum que os nomes de pessoas e países sejam traduzidos, então é importante identificá-los para que o modelo possa processá-los corretamente.

Outra tarefa associada ao REN é a Extração de Relações entre Entidades, do inglês *Entity Re-*

[1] Em inglês "Named Entity Recognition" (NER).

DOI: 10.21814/lm.15.2.396

lation Extraction, que consiste em identificar correlações entre entidades em um texto, como entre pessoa-e-organização e organização-e-lugar (Bach & Badaskar, 2007). Um exemplo, em "A sede da Apple fica na Califórnia", há uma relação de organização-e-lugar entre "sede da Apple" e "Califórnia". Essa não é uma tarefa simples, pois o seu desempenho depende da classificação correta das entidades do texto. Para o português, ela foi investigada no Segundo HAREM (Mota & Santos, 2008) e no IberLEF 2019 (Collovini et al., 2019).

O REN passou por diversas mudanças de paradigma no quesito de abordagem computacional, indo dos modelos baseados em regras ou léxicos aos modelos de aprendizado estatístico e, por fim, às rede neurais profundas. Ao longo desse período, produziu-se uma considerável literatura de revisão sobre tarefa. Um exemplo é o célebre artigo de Nadeau & Sekine (2007), em que são apresentadas discussões sobre a definição da tarefa, além de técnicas e algoritmos de aprendizado comuns na época.

Na última década, observou-se um aumento no número de artigos de revisão, principalmente aqueles voltados para ramos específicos do REN, como métodos de aprendizado de profundo (Yadav & Bethard, 2019; Li et al., 2020a) e aplicações na área de Biomedicina (Campos et al., 2012; Alshaikhdeeb & Ahmad, 2016), muitas dessas dedicadas a sistemas e recursos voltados para a língua inglesa.

Para o português, existe uma literatura de trabalhos investigativos sobre o REN conduzidos pelos pesquisadores da Linguateca[2] (Santos & Cardoso, 2007; Mota et al., 2007; Freitas et al., 2010; Mota & Santos, 2008), além de artigos para a comparação de algoritmos de aprendizado de máquina (Milidiú et al., 2007; Pellucci et al., 2011) e ferramentas (Amaral et al., 2014; Pires et al., 2017) aplicados ao REN. Mas trabalhos de revisão são datados e as pesquisas na área continuam avançando, portanto são necessárias novas revisões para a validação e comparação de técnicas mais recentes que surgiram, como os modelos de redes neurais profundas, e para a compreensão do que mudou e do que precisa ser trabalhado para novos avanços.

Este artigo tem por objetivo fornecer um panorama geral do desenvolvimento da tarefa de Reconhecimento de Entidades Nomeadas tendo em vista sua aplicação no processamento de dados em língua portuguesa. Serão apresentados os eventos promovidos para o REN, os recursos linguísticos e computacionais disponíveis e as pesquisas que têm sido feitas na área. Ademais, apresenta-se uma discussão a respeito dos desafios e caminhos futuros para pesquisas e, por fim, as considerações finais.

[2] https://www.linguateca.pt/

2. Definições da tarefa

Quando foi apresentada no MUC-6 (Grishman & Sundheim, 1996), definiu-se a tarefa como o reconhecimento de nomes de pessoas, lugares e organizações em textos, sendo assim a classificação de tipos de entidades definidos *a priori*. Essas três categorias acabaram se tornando as mais usuais entre os trabalhos em REN (Nadeau & Sekine, 2007), chamadas coletivamente de "ENAMEX". O MUC-6 também abrange categorias temporais, como data e tempo, e numéricas, como expressões monetárias e percentuais.

Na literatura em Português, a tarefa pode ser encontrada com duas nomenclaturas distintas: "Reconhecimento de Entidades Nomeadas" e "Reconhecimento de Entidades Mencionadas". O primeiro termo advém de uma tradução literal do nome adotado em inglês "Named Entity Recognition" e tem aparecido em diversos trabalhos na área (Amaral, 2017; Júnior et al., 2016; Mota et al., 2021; Pellucci et al., 2011). O segundo foi a adaptação inicialmente proposta para se referir à tarefa, em que "entidade mencionada" se refere às "entidades com nome próprio" (Santos & Cardoso, 2007).

As primeiras discussões a respeito do REN para a língua portuguesa estão ligadas ao HAREM, que será apresentado em mais detalhes na Seção 3.1. Diferente do MUC, o HAREM propõe um esquema de classificação para nomes próprios em geral, sem restrição a determinadas categorias (Santos, 2007). Isto é, parte-se da busca das entidades presentes nos textos em português, para depois definir-se o conjunto de categorias a partir dos exemplos encontrados.

O modelo de classificação do HAREM se baseia na ideia de vagueza da língua, em que um conceito não tem uma denotação fixa, mas depende do contexto para ser definido (Santos & Cardoso, 2007, p. 45). Sendo assim, não temos uma relação de um-para-um entre nome e objeto denotado. Em outras palavras, o mesmo nome pode se referir a mais de um objeto e é necessário o contexto para desambiguar sua referência. Por exemplo, temos nomes próprios que podem se referir a uma pessoa ou a um lugar cujo nome homenageia uma pessoa, como acontece com diversos nomes de ruas e prédios. Isso tem impacto no esquema de anotação do corpus, já que um mesmo nome pode ter mais de uma classificação possível pelo contexto.

Além disso, o HAREM considera os casos de metonímia, em que um nome tipicamente usado para designar determinado objeto é usado no lugar de outro com o qual mantém uma relação, como substituições de lugar-por-povo, empresa-por-produto (Santos & Cardoso, 2007, p. 47). No exemplo "O Brasil vai jogar na próxima semana na Copa do Mundo", "Brasil" não se refere ao país, mas aos jogadores da seleção brasileira, portanto seria ideal classificá-lo como "Pessoa". Isso não ocorre no MUC, em que uma entidade permanece com sua classificação prototípica, mesmo que o contexto forneça outra interpretação.[3] Desse modo, a perspectiva dada à tarefa mudou do MUC em relação ao HAREM, que expandiu o número de categorias de entidades e permitiu variações de classificação de acordo com o contexto de ocorrência.

A tarefa também mudou em outros aspectos conforme foi sendo aplicada em áreas específicas, como Medicina e Química, para a classificação nomes de substâncias, doenças, medicamentos, entre outros. Para citar um trabalho, Ferreira et al. (2010) buscam entidades relacionadas a diagnósticos médicos, classificando expressões como "diabetes controlado" e "alto nível de colesterol". Portanto, a tarefa não está mais restrita à classificação de nomes próprios, tendo uma aplicação muito mais ampla e diversa.

No trabalho de Marrero et al. (2013), os autores examinam e comparam diversas propostas de definição de Entidade Nomeada da literatura, apresentando análises baseadas em cunho gramatical, semântico e filosófico. No entanto, nenhuma das definições encontradas é boa o suficiente para delimitar o escopo da tarefa, o que os faz chegar à conclusão de que as Entidades Nomeadas serão definidas em razão do propósito de aplicação da tarefa.

Baseado nessa análise, não haveria uma definição pré-estabelecida para o que seriam "Entidades Nomeadas", mas *propostas de classificação*. Essas podem variar de acordo com a quantidade de categorias, o tipo (classes genéricas ou especializadas) e a organização (hierárquica ou não-hierárquica). Isso fica evidente quando se compara diferentes corpora: o CoNLL (Tjong Kim Sang, 2002; Tjong Kim Sang & De Meulder, 2003) considera quatro categorias de entidades (Pessoa, Local, Organização e Diversos) enquanto o HAREM (Santos & Cardoso, 2007; Mota & Santos, 2008) adota uma classificação hierárquica, com dez categorias principais divididas em subcategorias. Em abordagens de domínio aberto, como a Web, que visam classificar uma grande quantidade e diversidade de entidades, o número de classes pode ser ainda maior, como é o caso do modelo proposto por Sekine & Nobata (2004), que estabelece uma ontologia contendo cerca de 200 categorias de entidades.

[3] No MUC, o exemplo citado teria "Brasil" sendo classificado como "Lugar".

3. Iniciativas de fomento do REN

Desde seu surgimento, o REN ganhou muito espaço nas pesquisas em PLN. Entre 2000 e 2008, várias conferências importantes direcionaram mesas especificamente para trabalhos sobre a tarefa, entre elas o CoNLL (Tjong Kim Sang, 2002; Tjong Kim Sang & De Meulder, 2003) e o ACE (Doddington et al., 2004). No cenário atual, o REN é tópico na maioria dos eventos em PLN. Aqui, destacam-se duas iniciativas importantes que para a discussão e produção de recursos para o Reconhecimento de Entidades Nomeadas em língua portuguesa: o HAREM e o IberLEF 2019.

3.1. HAREM

A primeira grande iniciativa que fomentou o crescimento de pesquisas em REN aplicado ao Português foi o HAREM (Avaliação de sistemas de Reconhecimento de Entidades Mencionadas), uma parceria entre pesquisadores promovida pela equipe da Linguateca com o intuito de viabilizar encontros voltados para desenvolver e avaliar técnicas e sistemas para a classificação de nomes próprios em língua portuguesa. O HAREM teve duas edições, a primeira delas foi dividida em duas etapas: o Primeiro HAREM e o MiniHAREM (Santos & Cardoso, 2007), que ocorreram em 2005 e 2006, respectivamente; a segunda edição ocorreu em 2008, ficando conhecida como Segundo HAREM (Mota & Santos, 2008).

Ao todo, dez equipes participaram do HAREM submetendo seus sistemas para a avaliação. Essa foi feita com base nas Coleções Douradas (CD) do HAREM, um conjunto de corpora compostos de textos em português de vários países lusófonos que foram anotados manualmente para a tarefa. As CDs do HAREM possuem anotação hierárquica, com dez tipos de entidades principais (Pessoa, Local, Organização, Tempo, Valor, Obra, Acontecimento, Abstração, Coisa e Outro), cada qual com suas respectivas subcategorias. Os sistemas foram avaliados de acordo com diferentes métricas para medir o desempenho na identificação de entidades, na classificação morfológica (como gênero e número) e na classificação semântica (correspondente à categoria da entidade).

Os resultados apresentados ao fim do Primeiro (Santos & Cardoso, 2007) e Segundo HAREM (Mota & Santos, 2008) mostraram que a tarefa de classificação semântica foi mais desafiadora em comparação à de identificação, ficando pouco acima de 50% nos corpora avaliados. Para as ENAMEX (Pessoa, Lugar e Organização), aquela com pior desempenho geral foi Organização, o que também foi verificado em outras pesquisas (Amaral & Vieira, 2014; Santos et al., 2019). As categorias menos frequentes no corpus, como "Obra" e "Coisa", foram mais difíceis de classificar, talvez porque elas sejam mais ambíguas ou menos homogêneas em termos de padrões linguístico-ortográficos.

Além da organização desses encontros, os autores do HAREM publicaram uma extensa documentação a respeito do REN, oferecendo uma discussão sobre as dificuldades e soluções encontradas para a anotação de um corpus para o Reconhecimento de Entidades Nomeadas, as métricas de avaliação das ferramentas e metodologias para modelagem da tarefa. As Coleções Douradas produzidas foram disponibilizadas online com acesso livre, o que foi uma contribuição valorosa para a comunidade científica trabalhando com REN em português.

3.2. IberLEF 2019

O IberLEF (*Iberian Languages Evaluation Forum*) é uma campanha de avaliação conjunta voltada para diversas tarefas de processamento e compreensão de textos em línguas ibéricas. Em 2019, o IberLEF (Collovini et al., 2019) abordou a tarefa de Reconhecimento de Entidades Nomeadas como um dos tópicos do evento.

Os organizadores do IberLEF-2019 abriram uma chamada para equipes submeterem seus sistemas para avaliação. Os modelos submetidos para a competição foram avaliados em corpora de três domínios distintos: (I) textos gerais, como blogs e entrevistas, classificando cinco categorias de entidades (Pessoa, Local, Organização e Tempo e Valor), (II) dados clínicos de pacientes e (III) relatórios policiais, sendo (II) e (III) anotados apenas para Pessoa.

O IberLEF-2019 contou com a participação de cinco equipes para a competição na tarefa de REN. Os modelos variaram entre baseados em regras, baseados em aprendizado de máquina clássico, redes neurais e híbridos. Esses foram avaliados somente em relação à classificação, desconsiderando a identificação. Os resultados obtidos no IberLEF-2019 dependeram muito do tipo de corpus. No corpus policial, três modelos obtiveram bons resultados, acima de 80% medida-F. Já o desempenho no corpus clínico não foi tão promissor, uma vez que todos os modelos obtiveram menos que 50% de medida-F. No corpus geral, o melhor modelo alcançou 66,66% de medida-F, o que está longe de ser um resultado excelente.

Uma vez que os modelos foram avaliados em corpora de diferentes domínios, foi possível identificar quais foram os mais desafiadores. Como verificado, o corpus de dados clínicos se mostrou como o mais difícil, apesar de ter sido avaliado apenas para a categoria de entidade Pessoa. Isso reforça a dificuldade de aplicação da tarefa para determinados domínios. O IberLEF-2019 permitiu validar modelos recentes, como as redes neurais, para a aplicação de REN em corpora de diferentes domínios, mostrando que houve melhoria de desempenho para a classificação de entidades em textos de domínio geral. Entretanto, ainda há muito o que ser feito para que a tarefa possa ser considerada resolvida, principalmente em relação à classificação de entidades em domínio clínico.

4. Recursos linguísticos e computacionais

Existem diversos recursos linguísticos e computacionais disponíveis em língua portuguesa que podem ser aplicados sem necessidade de muitos ajustes por aqueles interessados em REN. Nesta revisão, foram selecionados recursos (ferramentas, corpora e léxicos) disponíveis em formato aberto.

4.1. Recursos linguísticos

Os corpora estão entre os principais recursos linguísticos para o processamento automático de línguas naturais. Aqui, corpus refere-se a um conjunto de textos digitais que pode ser processado por um computador. A tarefa de REN é tipicamente aplicada em corpora não-estruturados, como artigos de texto e postagens em redes sociais. Desse modo, há uma grande quantidade de conteúdos disponíveis para análise, seja em livros digitais, na Web, em revistas, dicionários etc. Contudo, os corpora mais valiosos são aqueles que possuem anotação linguística, classificando palavras, frases ou trechos de textos para uma tarefa específica.

Em REN, a anotação consiste em identificar e classificar todos os termos correspondentes a Entidades Nomeadas. Um esquema de anotação comum para a tarefa é o chamado BIO (Begin-Inside-Outside), proposto por Ramshaw & Marcus (1999), em que a primeira palavra de uma

entidade nomeada é marcada por "B-", as demais palavras da entidade (se houver) são anotados com "I-" e as palavras considerados não-entidades são marcadas por "O". A Figura 1 traz um exemplo desse tipo de anotação, em que "PES" é abreviação para "PESSOA".

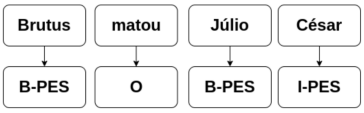

Figura 1: Representação do método de anotação BIO.

Outros esquemas de anotação para o REN são o IO (Inside-Outside), adotado no trabalho de Pirovani & Oliveira (2018), e o BILOU (Begin-Inside-Last-Outside-Unit), que aparece em Amaral & Vieira (2014). O IO distingue apenas entidades (I) de não-entidades (O), enquanto o BILOU diferencia início (B), meio (I) e fim (L) de entidades, entidades compostas de uma palavra (U) e não-entidades (O). Mas o esquema de anotação mais comum entre os trabalhos encontrados foi o BIO.

Um corpus anotado pode ser classificado em: (I) dourado, quando a sua anotação é feita manualmente e (II) prateado, quando sua anotação é feita automaticamente, sem auxílio ou revisão humana. Neste trabalho, refere-se corpora compostos de textos integral ou parcialmente escritos em Português anotados para o REN. A Tabela 1 apresenta uma comparação deles de acordo com o número de categorias de entidades, o número de entidades do corpus, o domínio e o tipo de anotação (dourado ou prateado).

Os corpora do HAREM (Santos & Cardoso, 2007; Freitas et al., 2010) são compostos principalmente de textos jornalísticos e da Web relacionados a diferentes temas, como ficção e política. O WikiNER[4] (Nothman et al., 2013), o Paramopama (Júnior et al., 2015) e o SESAME (Menezes et al., 2019) foram construídos a partir de textos de páginas da Wikipédia e anotados por meio de ferramentas computacionais. O SIGARRA (Pires, 2017) é constituído de dados da Web extraídos do sistema de informações da Universidade do Porto. As entidades são classificadas em Pessoa, Local, Organização, Data, Hora, Evento, Curso e Unidade Orgânica (por exemplo, nomes de institutos). Entre os corpora selecionados, o único composto exclusivamente de texto com linguagem da internet é o do Twitter (Peres et al., 2017), que é anotado para as categorias Pessoa, Local e Organização.

[4]O WikiNER é um corpus multilíngue, portanto apenas uma parte dele está em português.

Voltado para o domínio jurídico, o LeNER-Br (Araujo et al., 2018) é um corpus constituído de textos coletados de tribunais brasileiros e documentos legislativos. As entidades são categorizadas em Pessoa, Local, Organização, Tempo, Legislação e Jurisprudência. Já o SemClinBr (Oliveira et al., 2022) foi produzido a partir de dados clínicos de hospitais brasileiros, englobando diversas áreas de especialidade médica (cardiologia, neurologia etc.). As categorias semânticas foram baseadas no sistema UMLS,[5] que é hierárquico. Por exemplo, a categoria "Transtornos" contém subcategorias como "doença e síndrome" e "sinal ou sintoma".

Além dos corpora, a produção de léxicos, como os *gazetteers*, pode auxiliar os modelos de REN, principalmente os de abordagem híbrida ou de regras. Os *gazetteers* são repositórios contendo conjuntos de nomes próprios, por exemplo, nomes de pessoas, de doenças, de empresas etc. Eles são usados como fonte de conhecimento externo, fornecendo informações não contidas no texto que podem ser úteis para a classificação da entidade. O REPENTINO (Sarmento et al., 2006) é um léxico estruturado composto de nomes próprios extraídos do corpus WPT03.[6] Contém mais de 45.0000 exemplos de entidades divididas em 11 categorias e 97 subcategorias. Outro léxico importante é o HDBP (*Historical Dictionary of Brazilian Portuguese*), produzido por Vale et al. (2008), um dicionário de abreviações históricas.

4.2. Recursos computacionais

As ferramentas computacionais para a classificação de entidades são recursos úteis para a produção de novas ferramentas e avaliação do estado-da-arte em uma tarefa. Aqui, selecionou-se ferramentas com base em dois critérios (I) possuir modelos pré-treinados para o REN em português e (II) ser aberto. A relação de ferramentas é dada na Tabela 2.

As ferramentas encontradas estão baseadas em duas linguagens de programação muito usadas (C++ e Python). Entre essas, o spaCy é uma das mais conhecidas, projetado como uma ferramenta de aplicação industrial, conta com modelos pré-treinados em diferentes línguas, incluindo o português. O Polyglot e o Freeling também são bibliotecas com modelos pré-treinados em inúmeras línguas e tarefas. Já o

[5]https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

[6]https://www.linguateca.pt/WPT/WPT03.html

Corpus	Categoria	Entidade	Domínio	Anotação
Primeiro HAREM	10	5.132	geral	dourado
Segundo HAREM	10	7.847	geral	dourado
Mini-HAREM	10	3.758	geral	dourado
Paramopama	4	42.769	geral	prateado
WikiNER	4	330.286	geral	prateado
SESAME	3	6.411.479	geral	prateado
SIGARRA	7	12644	geral	dourado
Twitter-NER	3	935	geral	dourado
LeNER-Br	6	44.513	Direito	dourado
SemClinBr	100	65.129	Biomedicina	dourado

Tabela 1: Corpora em língua portuguesa anotados para o REN.

Corpus	Linguagem	Referência
spaCy	Python	https://spacy.io/
NLPyPort	Python	Ferreira et al. (2019)
Freeling	C++	Carreras et al. (2004)
Polyglot	Python	Al-Rfou et al. (2015)

Tabela 2: Ferramentas para Reconhecimento de Entidades Nomeadas em português.

NLPyPort é uma biblioteca baseada no NLTK[7] que foi desenvolvida especificamente para o português. Não foram encontradas ferramentas com interface gráfica que não exijam conhecimento em programação para a utilização, mas existem alguns grupos de investigação NLX que oferecem recursos online gratuitos para o REN, como o LX-Center.[8]

A acessibilidade a modelos pré-treinados — fornecida por repositórios como o HuggingFace[9] — e a disponibilização de bibliotecas para *deep learning* — como Keras, Pytorch e Tensorflow — vêm trazendo um crescimento de interesse nas pesquisas em PLN e impulsionado a área. De acordo com Li et al. (2020a), muitos dos trabalhos reportando avanços do estado-da-arte no REN têm sido obtidos por redes neurais. Por essa razão, os recursos pré-treinados em português são essenciais, já que esses podem ser ajustados para aplicações em tarefas específicas. O HuggingFace já possui modelos de redes neurais pré-treinados em dados do português, como o BERTimbau[10] e o BioBERT.[11] O BERTimbau foi treinado no brWaC (Wagner Filho et al., 2018), um corpus sem anotação linguística composto de textos em português de diversos domínios. O BioBERT (Schneider et al., 2020) foi treinado a partir de dados clínicos de hospitais brasileiros de diversas áreas médicas, como cardiologia, neurologia e endocrinologia.

O repositório de *word embeddings* pré-treinados, disponibilizado pelo Núcleo Interdisciplinar de Linguística Computacional da Universidade de São Paulo (NILC-USP)[12] também representa um recurso valioso para pesquisas com modelos de redes neurais em português. No repositório encontram-se disponíveis modelos pré-treinados Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), Wang2Vec (Ling et al., 2015) e FastText (Bojanowski et al., 2017).

5. Abordagens para o REN

Esta seção apresenta os trabalhos encontrados a respeito de REN aplicado à língua portuguesa. Para a seleção dos artigos utilizou-se o buscador Google Scholar,[13] pesquisando pelas palavras-chave *"reconhecimento de entidades nomeadas"*, *"named entity recognition in portuguese"* e *"portuguese named entity recognition"*. Os artigos estão divididos em quatro categorias de acordo com o método de abordagem aplicado: (I) baseados em regras, (II) modelos estatísticos tradicionais, (III) redes neurais profundas e (IV) modelos híbridos.

5.1. Modelos baseados em regras

Modelos baseados em regras são projetados com base em um conjunto de diretrizes para a tomada de decisões a partir de aspectos extraídos do texto. As regras podem ter por base informações

[7] https://www.nltk.org/
[8] http://lxcenter.di.fc.ul.pt/services/en/LXServicesNer.html
[9] https://huggingface.co/
[10] https://github.com/neuralmind-ai/portuguese-bert
[11] https://github.com/HAILab-PUCPR/BioBERTpt
[12] http://www.nilc.icmc.usp.br/embeddings
[13] https://scholar.google.com.br/

linguísticas (por exemplo, padrões sintáticos e semânticos), frequência de termos, similaridade com palavras em um dicionário de referência, entre outros.

É comum dividir o REN em duas etapas nos modelos de regras: (I) identificação e (II) classificação. Na primeira, as entidades nomeadas do texto são identificadas e separadas dos termos que não constituem entidades; na segunda, as entidades identificadas são classificadas de acordo com as categorias pré-definidas pelo modelo. Essa divisão é feita porque as regras para identificação e classificação são específicas de cada etapa.

As abordagens baseadas em regras foram muito exploradas nos primeiros sistemas para Reconhecimento de Entidades Nomeadas. Em razão do Primeiro HAREM, Bick (2006) propõe o PALAVRAS-NER, um modelo de gramática contextual que aplica uma série de regras baseadas em padrões sintático-semânticos e regras de desambiguação contextual. O PALAVRAS-NER foi o sistema que obteve melhor desempenho na avaliação do Primeiro HAREM, com 58,29% de medida-F na classificação de entidades. O SIEMES (Sarmento, 2006) também participou do Primeiro HAREM, alcançando o segundo lugar com 53,30% de medida-F. Sua abordagem é baseada na similaridade de possíveis entidades presentes em um texto com aquelas listadas no *gazetteer* REPENTINO. O algoritmo faz uma busca por correspondências completas ou parciais e desambigua entre possíveis classificações a partir de um conjunto de regras contextuais.

Outro sistema de regras é o PAMPO (Rocha et al., 2016), que utiliza informações morfossintáticas das palavras para a aplicação de regras de identificação de entidades. O PAMPO aplica conhecimento externo extraído de um conjunto de listas de termos contendo, por exemplo, palavras comuns que ocorrem na fronteira de entidades. Esses termos podem estar associados a categoria da entidade, como "senhor" e "doutor", que coocorrem com nomes de pessoas. O PAMPO chegou a 73,36% de medida-F para a identificação de nomes de organizações, lugares e pessoas no corpus do HAREM.

Por sua vez, Ferreira et al. (2010) propõem um modelo baseado em ontologias e regras linguísticas para detectar entidades em cartas médicas, buscando por informações como a condição, tratamento e evolução do quadro do paciente. Os autores testaram o modelo no MedAlert, um corpus próprio composto de 90 relatórios médicos de um hospital de Portugal, reportando uma precisão de 95,0% na classificação de entidades. Outros sistemas de regras são o Rembrandt (Cardoso, 2008), o REMMA (Ferreira et al., 2008) e o CAGE (Martins et al., 2007). Os dois primeiros são guiados por páginas da Wikipédia para identificar e extrair informações para a classificação de uma EN e o último é voltado para a identificação de entidades geográficas.

Como apontado por Li et al. (2020a), um dos pontos fracos dos modelos baseados em regras é que eles tendem a obter baixa cobertura[14] por serem projetados com base em um conjunto de regras restrito e específico. Isso faz com que eles tenham baixa portabilidade, ou seja, é difícil adaptá-los a outros domínios de aplicação.

5.2. Modelos estatísticos tradicionais

Os sistemas baseados em aprendizado de máquina estatístico tratam o REN como uma tarefa de classificação multi-classes em que o objetivo é classificar cada palavra com uma etiqueta correspondendo a uma categoria de entidade ou não-entidade. Esses modelos não costumam ter módulos distintos para identificação e classificação, processando ambas etapas em um único módulo de classificação.

Assim como em modelos de regras, os algoritmos de aprendizado de máquina precisam extrair características do texto para classificá-lo, isto é, padrões relevantes que forneçam pistas de classificação para o modelo. Existem diferentes tipos de características (traços) que podem ser analisadas. Na revisão apresentada por Nadeau & Sekine (2007), os autores definem três tipos de traços: no nível da palavra (por exemplo, pistas ortográficas), baseados em listas (como os *gazzetteers*) e no nível do documento (por exemplo, contagem de palavras no texto).

Existe mais de tipo de aprendizado de máquina, mas o mais adotado em REN é o supervisionado, em que o treinamento é realizado a partir de um corpus anotado. Alguns dos algorítimos usuais são: *Hidden Markov Model* (HMM), *Decision Tree* (DT), *Support Vector Machine* (SVM) e *Conditional Random Fields* (CRF).

O NERP-CRF (Amaral & Vieira, 2014) é um modelo CRF treinado com quinze traços, divididos em ortográficos, morfossintáticos e de contexto. A Coleção Dourada do HAREM serviu para treinamento e avaliação do modelo, que mostrou alta precisão (80,77% de medida-F), mas baixa cobertura (34,59%). As autoras verificam

[14] A cobertura se refere ao número de entidades capturadas pelo modelo.

que muitos dos erros do sistema foram na delimitação de fronteiras de uma EN, o que foi causado principalmente pela preposição "de", muito recorrente em nomes próprios em português.

Júnior et al. (2015) treinam o Starford-NER,[15] uma ferramenta pré-treinada baseada em um CRF, para a classificação de entidades. São selecionados traços ortográficos e morfossemânticos. Os autores também usam *gazetteers* para nomes de pessoas, lugares e organizações. O modelo foi treinado em diferentes corpora para comparação de desempenho, tendo obtido 82,34% de medida-F com o Paramopama combinado com o HAREM.

Já o trabalho de Solorio (2007) apresenta um sistema SVM treinado para o português a partir do conhecimento de um modelo desenvolvido para o espanhol. A autora considera traços internos (como informação ortográfica) e externos, obtidos a partir de um classificador de entidades para o espanhol. Esse classificador atribui a etiqueta morfossintática e uma pré-classificação da palavra seguindo o esquema de anotação BIO. Solorio (2007) mostrou que a combinação dos traços internos e externos retornou melhores resultados do que os traços internos sozinhos, indicando que é vantajoso aproveitar de conhecimento externo advindo de modelos treinados para línguas semelhantes.

Lopes et al. (2019) treinam um algoritmo CRF para identificar nomes de doenças, sintomas, genes, entre outros, com base em um corpus próprio de artigos científicos da revista portuguesa Sinapse.[16] O sistema desenvolvido considera traços contextuais com uma janela de cinco palavras, cujas informações ortográficas e morfossintáticas são extraídas. A classificação média do modelo ficou em 72,86% de medida-F. Também utilizando um CRF, de Souza et al. (2019) classificam entidades nomeadas a partir registros de saúde de hospitais anotados manualmente. Os autores adotam o esquema de classificação UMLS, mas optam por aglomerar determinadas classes devido a um desbalanceamento do corpus, englobando as categorias em três grandes grupos (Doenças, Procedimentos e Medicamentos). Treinou-se o CRF com traços ortográficos e morfossintáticos, chegando a uma média de 55,66% de medida-F.

[15] https://nlp.stanford.edu/software/CRF-NER.shtml

[16] https://www.sinapse.pt/

5.3. Modelos de redes neurais profundas

As redes neurais profundas têm ganhado muito espaço em inúmeras aplicações em PLN. Um dos motivos para o sucesso desses modelos são as representações vetoriais *word embeddings* (Bengio et al., 2003). Os *word embeddings* forneceram uma forma eficiente para a representação de texto em modelos computacionais que só processam informações numéricas, como é o caso das redes neurais. Eles têm sido aplicados não somente em palavras, mas também em caracteres (*character embeddings*) (Santos & Guimarães, 2015; Fernandes et al., 2018). Além disso, as representações vetoriais podem representar outros tipos de traços, como ortográficos e lexicais.

As arquiteturas de redes neurais são diversas, mas as mais adotadas em PLN são as redes neurais recorrentes, como *Long Short-Term Memory* (LSTM) (Hochreiter & Schmidhuber, 1997), e, mais recentemente, as redes *Transformer* (Vaswani et al., 2017), como BERT (Devlin et al., 2018) e RoBERTa (Liu et al., 2019).

Um dos primeiros trabalhos a testar uma rede neural para o REN em português foi Santos & Guimarães (2015). Os autores propõem uma rede neural que combina *word embeddings* e *character embeddings*, chamando o modelo de CharWNN. A rede foi treinada e testada em duas línguas: português e espanhol. Em suas discussões, os autores mostram que a combinação dos dois *embeddings* (de palavra e de caractere) é mais eficiente do que essas representações usadas isoladamente. A CharWNN alcançou 65,41% de medida-F no HAREM para a classificação em todas as categorias do corpus.

Por sua vez, Castro et al. (2018) aplicam uma rede neural LSTM bidirecional (BiLSTM) com camada de saída CRF para o REN em português. Foram comparados quatro tipos de *word embeddings* (FastText, GloVe, Wang2Vec e Word2Vec), tendo o Wang2Vec obtido o melhor desempenho nos experimentos testados. Também utilizando uma rede neural BiLSTM, Fernandes et al. (2018) avaliam diferentes tipos de classificadores para fazer o processamento na saída da rede neural. Os melhores resultados foram obtidos com uma arquitetura BiLSTM seguida de uma rede neural convolucional e uma classificador de saída CRF. Assim como Castro et al. (2018), os autores obtiverem melhores resultados com vetores Wang2Vec.

Santos et al. (2019) avaliam a eficiência de combinação de *word embeddings* não-contextuais e contextuais, utilizando *Flair embeddings* (Akbik et al., 2019). Esses *embeddings* codificam in-

formações a partir das palavras vizinhas e de seus caracteres, sendo portanto uma representação de palavras e de caracteres ao mesmo tempo. A rede neural implementada foi uma BiLSTM com camada de saída CRF e o corpus de avaliação foi o HAREM.

O BERTimbau foi testado para o REN em português por Souza et al. (2019), treinado no corpus HAREM para a classificação de entidades nomeadas. Os autores comparam o resultado do BERTimbau ao do BERT treinado em um corpus multilíngue e mostram que o primeiro se sai melhor.

Peres et al. (2017) testam variações de redes BLSTM para a classificação de entidades em *tweets* no Twitter-NER. O modelo apresentado por eles se saiu melhor do que o *baseline* avaliado, um modelo CRF, mas mesmo assim alcançou pouco mais de 50% de medida-F. Entre as classes de entidades avaliadas (Pessoa, Local e Organização), "Local" foi aquela com menor desempenho.

Recentemente, houve um crescimento na aplicação de REN para domínios específicos, o que se popularizou ainda mais com as redes neurais. No domínio jurídico, Araujo et al. (2018) treinam uma rede neural BiLSTM-CRF no corpus LeNER-Br utilizando *word embeddings* GloVe. Os resultados reportados foram excelentes, ficando com uma média acima de 90% de medida-F. Por sua vez, Mota et al. (2021) testam três redes neurais: duas redes recorrentes com combinações distintas de *word embeddings* e uma rede convolucional. Os modelos foram avaliados em relação a duas categorias (Legislação e Jurisprudência), tendo como corpus de treino e teste o LeNER-Br mais um conjunto de petições iniciais de processos. O modelo com melhor desempenho foi uma rede neural recorrente com *Flair embeddings*, que alcançou média de 76% de medida-F para a classificação de entidades.

Outra área que tem chamado a atenção dos pesquisadores é a Biomedicina. Lidando com a área de neurologia, Lopes et al. (2020) apresentam duas arquiteturas BiLSTM-CRF alimentadas com *word embeddings* pré-treinados em dois domínios distintos: na Wikipédia e em dados clínicos. O modelo com os *embeddings* de domínio clínico obtiveram 75,08% de medida-F, em comparação a 74,58% obtido pelo modelo treinado em domínio geral. Esse resultado corrobora que existem diferenças nas características de entidades de domínio específico que não podem ser desconsideradas.

Schneider et al. (2020) estavam interessados em avaliar a influência do treinamento de modelos em domínio específico para o REN. Para isso, comparam o desempenho do BioBERTpt ao do BERTimbau e ao do BERT treinado em um corpus multilíngue. O BioBERT obteve 60,4% de medida-F, superando os demais modelos em mais de 2%. Aponta-se novamente que as entidades nomeadas em domínio específico têm suas particularidades que não são capturadas por modelos treinados em textos genéricos. Em um trabalho posterior, de Souza et al. (2020) aplicam o BioBERT para a classificação multinomial no corpus SemClinBr, em que uma única entidade pode ter mais de uma etiqueta em determinados contextos. Comparado aos outros modelos, o BioBERT obtém o melhor desempenho na tarefa, alcançando 3,5% a mais em relação ao *baseline* (CRF).

5.4. Modelos híbridos

Os modelos híbridos são aqueles que combinam duas ou mais abordagens distintas, como aprendizado de máquina e regras manuais. Jiang et al. (2016) aponta que os modelos híbridos são boas escolhas quando se quer combinar alta precisão e cobertura e não há disponibilidade de grande quantidade de dados anotados para o treinamento.

As abordagens híbridas podem ser das mais variadas. O trabalho de Milidiú et al. (2007) compara algumas abordagens para o REN aplicados para o português. Eles selecionam três algoritmos (HMM, SVM e *Transformation based learning*), além de um modelo *baseline* baseado em *gazetteers* e regras. Os autores realizam testes combinando os modelos entre si, mostrando que eles se saem melhor combinados do que isolados.

Ferreira et al. (2007) propõem um sistema de REN que combina dois módulos: um de regras manuais baseadas em expressões regulares para a classificação de entidades numéricas e um de aprendizado de máquina para os nomes próprios. Além disso, o modelo conta com um conjunto de léxicos de nomes para ajudar na correção de possíveis erros de classificação. Outro modelo híbrido é o CRF-LG (Pirovani & Oliveira, 2018), um sistema que combina um CRF com um conjunto de regras manuais verificadas a partir de uma gramática local usada para pré-atribuir a classificação de entidades. O CRF-LG foi avaliado no HAREM com medida-F de 57,8%.

Utilizando o método de aprendizado profundo, Júnior et al. (2016) concatenam uma rede neural LSTM com uma rede neural convolucional (CNN). A CNN gera representações de caracteres das palavras, que são combinadas a *word embeddings*. A rede LSTM é alimentada pelo vetor final gerado. Os corpora usados para treino e

teste foram o HAREM, o WikiNER e o Paramopama. Na avaliação do HAREM, o modelo obtém medida-F de 71,35% para a classificação de cinco classes de entidades (Pessoa, Local, Organização, Tempo e Valor).

Em domínio especializado, Dias et al. (2020) aplicam a tarefa de REN a dados sensíveis, capturando informações como nomes próprios, telefone, endereço e profissão de indivíduos. Para isso, os autores propõem um modelo composto de inúmeros módulos: um baseado em regras, um baseado em léxico e um de modelos estatísticos. O modelo de regras classifica entidades como números telefônicos e e-mails a partir de expressões regulares, enquanto os modelos estatísticos focam na classificação de entidades mais complexas, como pessoa, local e organização. Após a avaliação de diversos algoritmos, os autores utilizam uma rede neural BiLSTM no módulo estatístico. O modelo foi avaliado no DataSense NER Corpus, um corpus próprio anotado para a pesquisa, obtendo 83,0% de medida-F na classificação de entidades.

5.5. Panorama geral da tarefa

A Seção 5 apresentou uma revisão de trabalhos a respeito de REN para a língua portuguesa, que foram divididos por tipo de técnica adotada na abordagem. Buscou-se distinguir os trabalhos de domínio geral e específico e o corpus a que o modelo foi aplicado, já que isso reflete muito no desempenho. Para uma visão geral dos resultados, as Tabelas 3 e 4 apresentam uma comparação dos trabalhos discutidos em relação ao método utilizado, corpus de teste e medida-F obtida. As abreviações "MR", "AM", "RN" e "MB" significam "Modelo de regras", "Aprendizado de máquina", "Rede neural" e "Modelo híbrido", respectivamente.

Analisando os trabalhos da Tabela 3, vemos que a maioria foi testado no HAREM, o que o torna o principal corpus de textos de domínio geral para a avaliação de desempenho em REN. O HAREM além de ser anotado manualmente, contém muitas categorias de entidade e subcategorias, que podem ser selecionadas de acordo com o objetivo de pesquisa. Comparando os trabalhos em relação ao método de abordagem, observa-se que os modelos de redes neurais são os que vem alcançando melhores resultados. Para a avaliação com dez categorias de entidades, o modelo com o estado-da-arte é o BERTimbau (Souza et al., 2019), com 78,6% de medida-F. Já o desempenho no HAREM para apenas quatro categorias (Pessoa, Local, Organização e Tempo) foi de 80,7%, reportado por Júnior et al. (2015).

Trabalho	Método	Corpus teste	F1
Bick	MR	HAREM	58,2
Sarmento	MR	HAREM	53,3
Rocha et al.	MR	HAREM	73,6*
Martins et al.	MR	HAREM	34,1
Cardoso	MR	HAREM	56,7
Ferreira et al.	MR	HAREM	45,2
Amaral & Vieira	AM	HAREM	57,9
Júnior et al.	AM	HAREM	80,7**
Solorio	AM	Lácio Web	–
Santos & Guimarães	RN	HAREM	65,4
Castro et al.	RN	HAREM	70,3
Fernandes et al.	RN	HAREM	67,5
Santos et al.	RN	HAREM	74,6
Souza et al.	RN	HAREM	78,6
Peres et al.	RN	NER-Twitter	52,7
Milidiú et al.	MB	próprio	88,1
Ferreira et al.	MB	HAREM	92,5*
Pirovani & Oliveira	MB	HAREM	57,8
Júnior et al.	MB	HAREM	71,3**
Dias et al.	MB	próprio	83,0

* indica que o modelo foi avaliado apenas para identificação.
** indica que o modelo avaliado apenas para um subconjunto de entidades do corpus.

Tabela 3: Comparação de resultados dos trabalhos em domínio geral.

O único trabalho encontrado para a classificação de *tweets* foi o de Peres et al. (2017), que obteve 52,7% para classificação de nomes de pessoas, lugares e organizações. Esse resultado reforça a dificuldade da aplicar o REN a textos que usam linguagem da internet, como é o caso das redes sociais. Nesse tipo de meio, as entidades muitas vezes são pessoas da roda social do usuário (como amigos e família) e os nomes próprios não costumam ser escritos com letra inicial maiúscula, além de haver muitos apelidos, que são difíceis de detectar.

No domínio específico (Tabela 4), os resultados são mais difíceis de comparar, pois os trabalhos foram testados em corpora e domínios variados. Entre os corpora, o SemClinBR parece ser um dos mais desafiadores, sendo aquele com maior número de categorias e subcategorias classificadas, que podem ser confundidas entre si pelos modelos. De acordo com Schneider et al. (2020), as categorias mais difíceis foram aquelas com maior granularidade, especificidade e com vocabulários variáveis entre os hospitais, como a subcategoria "Laboratório". Na área de Direito, Araujo et al. (2018) alcançou um ótimo resultado no LeNER, com mais de 90% de medida-F.

Trabalho	Área	Método	Corpus teste	F1
Ferreira et al.	Médica	MR	MedAlert	–
Lopes et al.	Médica	AM	próprio	72,8
de Souza et al.	Médica	AM	próprio	55,6
Araujo et al.	Direito	RN	LeNER	92,5
Mota et al.	Direito	RN	próprio	76,0
Lopes et al.	Médica	RN	próprio	74,9
Schneider et al.	Médica	RN	SemClinBR	60,4
de Souza et al.	Médica	RN	SemClinBR	56,1

Tabela 4: Comparação de resultados dos trabalhos em domínio específico.

Vale ressaltar que no LeNER apenas duas categorias das seis são específicas do Direito (Jurisprudência e Legislação). Um dos pontos que favorecem na classificação de entidades em Direito é a padronização, isto é, os termos, abreviações e expressões seguem, em geral, uma fórmula, tendo textos com um estilo de escrita muito repetitivos que pode ajudar os modelos a capturarem as características das entidades.

6. Desafios e caminhos futuros da área

O Reconhecimento de Entidades Nomeadas tem alcançado bons resultados de avaliação em corpora de língua inglesa, com o estado-da-arte em 94,6% de medida-F no corpus ConLL-2003, reportado pelo trabalho de Wang et al. (2020). Já para o português, o estado-da-arte no MiniHAREM, um dos corpus mais utilizados para teste na tarefa, é de 78,6% de medida-F, obtido por Souza et al. (2019) com o BERTimbau. Isso mostra que ainda existe muito trabalho a ser feito na área. Aqui serão discutidos alguns desafios no REN em português e possíveis caminhos a serem explorados.

6.1. Criação de novos recursos

Os corpora anotados são recursos valiosos para o treinamento de algoritmos de aprendizado supervisionado. Quando se trata de redes neurais profundas, a quantidade de textos anotados precisa ser ainda maior para que o algoritmo obtenha melhores resultados. Para fins de comparação, o ConLL-2003 tem cerca de 23.499 entidades no corpus de treinamento, enquanto o Primeiro HAREM, usado geralmente para treino dos modelos, tem cerca de 5.132. Desse modo, é esperado que os modelos treinados no ConLL obtenham melhores resultados.

A criação de novos corpora anotados para o português é um caminho para avançar as pesquisas em REN. Existem diversos métodos de anotação automática através de ferramentas pré-treinadas ou de ontologias, como a DBpedia. Nesse sentido, seria interessante explorar tais recursos para a anotação, principalmente de textos em rede sociais. Como mostrado na Tabela 1, somente um dos corpora encontrados é composto completamente de textos da Web.

Li et al. (2020a) ressaltam que a classificação de entidades nomeadas em textos de redes sociais tende a ser mais desafiadora, com resultados pouco acima de 40% de medida-F. Sendo assim, é preciso produzir mais materiais para o treinamento e a avaliação em textos dessa natureza. Somente assim será possível obter modelos mais robustos, capazes de lidar com a diversidade estilística e aplicáveis no domínio da Web. Nesse sentido, a API do Twitter[17] pode ser avaliada como uma ferramenta de anotação automática de *tweets*, já que ela fornece alguns recursos prontos para a identificação de pessoas, lugares e produtos citados no texto.

Os corpora de domínios específicos também são importantes para o treinamento de modelos especializados. O domínio comercial tem recebido muita atenção em PLN nos últimos anos, principalmente para a análise de comentários de usuários a respeito de produtos. Modelos especializados em classificação de produtos já foram explorados na literatura estrangeira (Zhao & Liu, 2008; Luo et al., 2011), mas não foram encontrados trabalhos com textos em português, sendo esse um campo rico para a produção de corpora e ferramentas para REN.

6.2. Reutilização de recursos

Como a anotação manual de corpora de qualidade é um trabalho árduo e demorado, uma alternativa possível seria combinar diversos corpora existentes para o treinamento dos modelos, como foi feito no IberLef-2019. Isso pode aju-

[17] https://developer.twitter.com/en/docs/twitter-api

dar a aumentar o número de exemplos para treinamento e fornecer mais diversidade linguística e estilística, ainda mais se os textos forem de domínios e gêneros distintos. A maior dificuldade nessa estratégia é a unificação dos corpora, já que eles podem ter esquemas de anotação distintos e inconsistências que prejudiquem a classificação.

Além da quantidade de exemplos, o balanceamento do corpus é extremamente importante para que o modelo não fique enviesado pelas categorias com mais exemplos. Por isso, retirar ou concatenar categorias que não são representativas dentro do corpus também pode ajudar a melhorar o desempenho dos modelos. Por exemplo, o HAREM define 10 categorias de entidades, mas algumas delas são pouco representativas em relação às demais, o que prejudica muito o desempenho dos modelos quando avaliados em todas elas.

Por fim, a reutilização de modelos pré-treinados é um caminho promissor para avançar com as pesquisas em REN. As técnicas de *transfer learning* vêm sendo amplamente exploradas em PLN, principalmente em aplicações com redes neurais profundas (Malte & Ratadiya, 2019; Alyafeai et al., 2020). Essas técnicas exploram o conhecimento adquirido do treinamento de um modelo em um domínio para aplicá-lo em outro. Alyafeai et al. (2020) apontam que o uso de técnicas de *transfer learning* em redes neurais profundas pode diminuir a complexidade de treinamento, seja pela quantidade de parâmetros ou tempo necessário para o treinamento.

Diversos trabalhos discutidos nesta revisão fazem uso de recursos computacionais pré-treinados para a classificação de entidades nomeadas (Souza et al., 2019; Schneider et al., 2020; Fernandes et al., 2018; Castro et al., 2018; Santos et al., 2019). Mais de uma delas mostrou que modelos pré-treinados na mesma língua que a tarefa alvo apresentam melhores resultados do aqueles pré-treinados em outras línguas (Souza et al., 2019; Schneider et al., 2020), indicando que o conhecimento para a resolução do REN é muito associado à língua alvo.

Outra vantagem da reutilização de modelos pré-treinados é que eles podem ser adaptados a novos domínios de aplicação. Existem diversas maneiras de reutilizá-los, seja pelo treinamento em um corpus anotado através de *fine-tuning*, seja utilizando-os como *word embeddings* para um novo modelo ou combinados com outras técnicas para gerar modelos híbridos.

7. Considerações finais

Este artigo apresentou uma trajetória da tarefa de Reconhecimento de Entidades Nomeadas para língua portuguesa, oferecendo um panorama geral para aqueles interessados em saber mais sobre a tarefa, recursos e técnicas disponíveis. Os trabalhos foram apresentados de forma analógica ao desenvolvimento da área ao longo dos anos, partindo dos modelos de regras aos de aprendizado de máquina profundo. As pesquisas em PLN com língua portuguesa têm se dedicado ao REN há mais de 10 anos. Nesse sentido, as iniciativas de eventos sobre REN foram importantes para estabelecer a área, principalmente considerando o HAREM, que forneceu recursos computacionais e linguísticos importantes para o ponta-pé inicial da área.

No cenário atual, vê-se uma nova onda de impulsionamento nas pesquisas de REN em português, direcionada fortemente aos modelos de aprendizado profundo. Atribui-se isso à recente disponibilização de corpora anotados maiores e mais diversos, modelos computacionais pré-treinados e bibliotecas de PLN para treinamento de modelos de *deep learning*. Contudo, ainda há um longo caminho a ser percorrido até que essa seja dada como uma tarefa resolvida ou, ao menos, tenha alcançado resultados bons o suficiente para aplicação em mundo real. Apresentou-se alguns caminhos possíveis a serem adotados em via de se obter sistemas de REN mais robustos e adaptáveis para o português.

Mais do que a criação e avaliação de recursos linguísticos e computacionais, será necessário voltar um olhar mais crítico ao que já foi produzido para tentar compreender melhor as principais dificuldades da tarefa e delinear possíveis soluções. Como discutido nos trabalhos de Santos et al. (2019), Júnior et al. (2015) e Amaral & Vieira (2014), a confusão entre classes foi um erro frequente cometido pelo reconhecedor de entidades, como classificar em "pessoa" um lugar que tem seu nome homenageado a alguém, tal qual "Raposo Tavares" se referindo à rodovia. Esse é tipo de ambiguidade de classificação é um dos tipos de erros mais recorrentes entre os modelos de REN. Desse modo, ressalta-se o trabalho colaborativo entre pesquisadores para investigar tais questões, a fim de propor soluções e avaliar técnicas para superar esses problemas.

Referências

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter & Roland Vollgraf. 2019. FLAIR: an easy-to-use framework for state-of-the-art NLP. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 54–59. 10.18653/v1/N19-4010.

Al-Rfou, Rami, Vivek Kulkarni, Bryan Perozzi & Steven Skiena. 2015. POLYGLOT-NER: Massive multilingual named entity recognition. Em *SIAM International Conference on Data Mining*, 586–594. 10.1137/1.9781611974010.

Alshaikhdeeb, Basel & Kamsuriah Ahmad. 2016. Biomedical named entity recognition: a review. *International Journal on Advanced Science, Engineering and Information Technology* 6(6). 889–895. 10.18517/ijaseit.6.6.1367.

Alyafeai, Zaid, Maged Saeed AlShaibani & Irfan Ahmad. 2020. A survey on transfer learning in natural language processing. ArXiv [cs.CL]. 10.48550/arXiv.2007.04239.

Amaral, Daniela. 2017. *Reconhecimento de entidades nomeadas na área da geologia: bacias sedimentares brasileiras*: Pontifícia Universidade Católica do Rio Grande do Sul. Tese de Doutoramento.

Amaral, Daniela, Evandro Brasil Fonseca, Lucelene Lopes & Renata Vieira. 2014. Comparative analysis of Portuguese named entities recognition tools. Em *9th International Conference on Language Resources and Evaluation (LREC)*, 2554–2558.

Amaral, Daniela & Renata Vieira. 2014. NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. *Linguamática* 6(1). 41–49.

Araujo, Pedro, Teófilo Campos, Renato Oliveira, Matheus Stauffer, Samuel Couto & Paulo Bermejo. 2018. LeNER-Br: a dataset for named entity recognition in brazilian legal text. Em *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 313–323.

Babych, Bogdan & Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. Em *7th International EAMT workshop on MT and other language technology tools*, 1–8.

Bach, Nguyen & Sameer Badaskar. 2007. A review of relation extraction. Unpublished University Work: Literature review for the "Language and Statistics II" class.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent & Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3. 1137–1155.

Bick, Eckhard. 2006. Functional aspects in Portuguese NER. Em *International Workshop on Computational Processing of the Portuguese Language (PROPOR)*, 80–89.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)* 5. 135–146. 10.1162/tacl_a_00051.

Campos, David, Sérgio Matos & José Luís Oliveira. 2012. Biomedical named entity recognition: a survey of machine-learning tools. *Theory and Applications for Advanced Text Mining* 11. 175–195. 10.5772/51066.

Cardoso, Nuno. 2008. REMBRANDT: reconhecimento de entidades mencionadas baseado em relaçoes e análise detalhada do texto. Em *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 195–211. Linguateca.

Carreras, Xavier, Isaac Chao, Lluís Padró & Muntsa Padró. 2004. FreeLing: An open-source suite of language analyzers. Em *4th International Conference on Language Resources and Evaluation (LREC)*, 239–242.

Castro, Pedro, Nádia Silva & Anderson Soares. 2018. Portuguese named entity recognition using LSTM-CRF. Em *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 83–92.

Collovini, Sandra, Joaquim Francisco Santos Neto, Bernardo Scapini Consoli, Juliano Terra, Renata Vieira, Paulo Quaresma, Marlo Souza, Daniela Barreiro Claro & Rafael Glauber. 2019. IberLEF 2019 Portuguese named entity recognition and relation extraction tasks. Em *Iberian Languages Evaluation Forum (IberLEF)*, 390–410.

Dai, Zeyu, Hongliang Fei & Ping Li. 2019. Coreference aware representation learning for neural named entity recognition. Em *28th International Joint Conference on Artificial Intelligence (IJCAI)*, 4946–4953. 10.24963/ijcai.2019/687.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv [cs.CL]. 10.48550/arXiv.1810.04805.

Dias, Mariana, João Boné, João C Ferreira, Ricardo Ribeiro & Rui Maia. 2020. Named entity recognition for sensitive data discovery in Portuguese. *Applied Sciences* 10(7). 2303. 10.3390/app10072303.

Doddington, George R, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel & Ralph M Weischedel. 2004. The Automatic Content Extraction (ACE) program-tasks, data, and evaluation. Em *4th International Conference on Language Resources and Evaluation (LREC)*, 837–840.

Fernandes, Ivo, Henrique Lopes Cardoso & Eugenio Oliveira. 2018. Applying deep neural networks to named entity recognition in portuguese texts. Em *5th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 284–289. 10.1109/SNAMS.2018.8554782.

Ferreira, Eduardo, João Balsa & António Branco. 2007. Combining rule-based and statistical methods for named entity recognition in Portuguese. Em *5th Workshop em Tecnologias da Informaçao e da Linguagem Humana*, 1615–1624.

Ferreira, João, Hugo Gonçalo Oliveira & Ricardo Rodrigues. 2019. Improving NLTK for processing Portuguese. Em *8th Symposium on Languages, Applications and Technologies (SLATE)*, 18:1–18:9. 10.4230/OASIcs.SLATE.2019.18.

Ferreira, Liliana, António Teixeira & Joao Paulo Silva Cunha. 2008. REMMA: Reconhecimento de entidades mencionadas do MedAlert. Em *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 213–229. Linguateca.

Ferreira, Liliana, António Teixeira & Joao Paulo da Silva Cunha. 2010. Information extraction from portuguese hospital discharge letters. Em *VI Jornadas en Tecnologia del Habla/Speech and Languages Technologies for Iberian Languages (FALA)*, 39–42.

Freitas, Cláudia, Paula Carvalho, Hugo Gonçalo Oliveira, Cristina Mota & Diana Santos. 2010. Second HAREM: advancing the state of the art of named entity recognition in Portuguese. Em *7th International Conference on Language Resources and Evaluation (LREC)*, 3630–3637.

Gao, Cheng-sheng, Jun-fu Zhang, Wei-ping Li, Wen Zhao & Shi-kun Zhang. 2020. A joint model of named entity recognition and coreference resolution based on hybrid neural network. *Acta Electonica Sinica* 48(3). 442–448. 10.3969/j.issn.0372-2112.2020.03.004.

Grishman, Ralph & Beth M Sundheim. 1996. Message understanding conference-6: A brief history. Em *International Conference on Computational Linguistics (COLING)*, 466–471.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8). 1735–1780. 10.1162/neco.1997.9.8.1735.

Jiang, Ridong, Rafael E Banchs & Haizhou Li. 2016. Evaluating and combining name entity recognition systems. Em *6th Named Entity Workshop*, 21–27. 10.18653/v1/W16-2703.

Júnior, C Mendonça, Hendrik Macedo, Thiago Bispo, Flávio Santos, Nayara Silva & Luciano Barbosa. 2015. Paramopama: a brazilian-portuguese corpus for named entity recognition. Em *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, SBC.

Júnior, Carlos, Luciano Barbosa, Hendrik Macedo & SE Sao Cristóvao. 2016. Uma arquitetura hıbrida LSTM-CNN para reconhecimento de entidades nomeadas em textos naturais em lıngua portuguesa. Em *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 241–252.

Li, Jing, Aixin Sun, Jianglei Han & Chenliang Li. 2020a. A survey on deep learning for named entity recognition. Em *International Conference on Data Engineering*, vol. 34 1, 50–70. 10.1109/ICDE55515.2023.00335.

Li, Zhen, Dan Qu, Chaojie Xie, Wenlin Zhang & Yanxia Li. 2020b. Language model pre-training method in machine translation based on named entity recognition. *International Journal on Artificial Intelligence Tools* 29(07n08). 2040021. 10.1142/S0218213020400217.

Ling, Wang, Chris Dyer, Alan W Black & Isabel Trancoso. 2015. Two/too simple adaptations of word2vec for syntax problems. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 1299–1304. 10.3115/v1/N15-1142.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. ArXiv [cs.CL]. 10.48550/arXiv.1907.11692.

Lopes, Fábio, César Teixeira & Hugo Gonçalo Oliveira. 2019. Named entity recognition in portuguese neurology text using CRF. Em *EPIA Conference on Artificial Intelligence*, 336–348. 10.1007/978-3-030-30241-2_29.

Lopes, Fábio, César Teixeira & Hugo Gonçalo Oliveira. 2020. Comparing different methods for named entity recognition in portuguese neurology text. *Journal of Medical Systems* 44(4). 1–20. 10.1007/s10916-020-1542-8.

Luo, Fang, Han Xiao & Weili Chang. 2011. Product named entity recognition using conditional random fields. Em *4th International Conference on Business Intelligence and Financial Engineering*, 86–89. 10.1109/BIFE.2011.101.

Malte, Aditya & Pratik Ratadiya. 2019. Evolution of transfer learning in natural language processing. ArXiv [cs.CL]. 10.48550/arXiv.1910.07370.

Marrero, Mónica, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato & Juan Miguel Gómez-Berbís. 2013. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces* 35(5). 482–489. 10.1016/j.csi.2012.09.004.

Martins, Bruno, Mário Silva & Marcirio Chaves. 2007. O sistema CaGE no HAREM-reconhecimento de entidades geográficas em textos em língua portuguesa. Em *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM a primeira avaliação conjunta na área*, 98–112. Linguateca.

Menezes, Daniel, Ruy Milidiu & Pedro Savarese. 2019. Building a massive corpus for named entity recognition using free open data sources. Em *8th Brazilian Conference on Intelligent Systems (BRACIS)*, 6–11.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. ArXiv [cs.CL]. 10.48550/arXiv.1301.3781.

Milidiú, Ruy Luiz, Julio Cesar Duarte & Roberto Cavalcante. 2007. Machine learning algorithms for portuguese named entity recognition. *Inteligencia Artificial* 11(36). 67–75.

Mollá, Diego, Menno Van Zaanen & Daniel Smith. 2006. Named entity recognition for question answering. Em *Australasian Language Technology Workshop*, 51–58.

Mota, Caio, André Nascimento, Péricles Miranda, Rafael Ferreira Mello, Isabel Maldonado & José Coelho Filho. 2021. Reconhecimento de entidades nomeadas em documentos jurídicos em português utilizando redes neurais. Em *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 130–140.

Mota, Cristina & Diana Santos. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo HAREM*. Linguateca.

Mota, Cristina, Diana Santos & Elisabete Ranchhod. 2007. Avaliação de reconhecimento de entidades mencionadas: princípio de AREM. Em *Avaliação Conjunta: um novo paradigma no processamento computacional da língua portuguesa*, 161–175. IST Press.

Nadeau, David & Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30(1). 3–26. 10.1075/li.30.1.03nad.

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy & James R Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194. 151–175. 10.1016/j.artint.2012.03.006.

Oliveira, Lucas, Ana Carolina Peters, Adalniza da Silva, Caroline Gebeluca, Yohan Gumiel, Lilian Cintho, Deborah Carvalho, Sadid Al Hasan & Claudia Moro. 2022. SemClinBr: a multi institutional and multi specialty semantically annotated corpus for portuguese clinical NLP tasks. *Journal of Biomedical Semantics* 13. 13. 10.1186/s13326-022-00269-1.

Pellucci, Paulo Roberto Simões, Renato Ribeiro de Paula, Walter Borges de Oliveira Silva & Ana Paula Ladeira. 2011. Utilização de técnicas de aprendizado de máquina no reconhecimento de entidades nomeadas no português. *e-xacta* 4(1). 73–81.

Pennington, Jeffrey, Richard Socher & Christopher D Manning. 2014. Glove: Global vectors for word representation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 10.3115/v1/D14-1162.

Peres, Rafael, Diego Esteves & Gaurav Maheshwari. 2017. Bidirectional LSTM with a context input window for named entity recognition in tweets. Em *The Knowledge Capture Conference*, 1–4. 10.1145/3148011.3154478.

Pires, André, José Devezas & Sérgio Nunes. 2017. Benchmarking named entity recognition tools for portuguese. Em *9th INForum: Simpósio de Informática*, 111–121.

Pires, André Ricardo Oliveira. 2017. *Named entity extraction from Portuguese web text*: Universidade do Porto. Tese de Mestrado.

Pirovani, Juliana & Elias Oliveira. 2018. Portuguese named entity recognition using conditional random fields and local grammars. Em *11th International Conference on Language Resources and Evaluation (LREC)*, 4452–4456.

Ramshaw, Lance A & Mitchell P Marcus. 1999. Text chunking using transformation-based learning. Em *Natural language processing using very large corpora*, 157–176. Springer. 10.1007/978-94-017-2390-9_10.

Rocha, Conceição, Alípio Jorge, Roberta Sionara, Paula Brito, Carlos Pimenta & Solange Rezende. 2016. PAMPO: using pattern matching and pos-tagging for effective named entities recognition in portuguese. ArXiv [cs.IR]. 10.48550/arXiv.1612.09535.

Santos, Cícero Nogueira dos & Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. Em *5th Named Entity Workshop*, 25–33. 10.18653/v1/W15-3904.

Santos, Diana. 2007. O modelo semântico usado no primeiro HAREM. Em *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM a primeira avaliação conjunta na área*, 43–57. Linguateca.

Santos, Diana & Nuno Cardoso. 2007. *Reconhecimento de entidades mencionadas em português: Documentação e actas do harem a primeira avaliação conjunta na área*. Linguateca.

Santos, Joaquim, Bernardo Consoli, Cicero dos Santos, Juliano Terra, Sandra Collonini & Renata Vieira. 2019. Assessing the impact of contextual embeddings for portuguese named entity recognition. Em *8th Brazilian Conference on Intelligent Systems (BRACIS)*, 437–442. 10.1109/BRACIS.2019.00083.

Sarmento, Luis. 2006. SIEMÊS –a named-entity recognizer for portuguese relying on similarity rules. Em *International Workshop on Computational Processing of the Portuguese Language (PROPOR)*, 90–99. 10.1007/11751984_10.

Sarmento, Luís, Ana Sofia Pinto & Luís Cabral. 2006. Repentino – a wide-scope gazetteer for entity recognition in portuguese. Em *International Workshop on Computational Processing of the Portuguese Language (PROPOR)*, 31–40. 10.1007/11751984_4.

Schneider, Elisa Terumi Rubel, Joao Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro & Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt – a portuguese neural language model for clinical named entity recognition. Em *3rd Clinical Natural Language Processing Workshop*, 65–72. 10.18653/v1/2020.clinicalnlp-1.7.

Sekine, Satoshi & Chikashi Nobata. 2004. Definition, dictionaries and tagger for extended named entity hierarchy. Em *4th International Conference on Language Resources and Evaluation (LREC)*, 1977–1980.

Solorio, Thamar. 2007. MALINCHE: a NER system for portuguese that reuses knowledge from Spanish. Em *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM a primeira avaliação conjunta na área*, 123–136. Linguateca.

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. ArXiv [cs.CL]. 10.48550/arXiv.1909.10649.

de Souza, João Vitor Andrioli, Yohan Bonescki Gumiel, Lucas Emanuel Silva & Claudia Maria Cabral Moro. 2019. Named entity recognition for clinical portuguese corpus with conditional random fields and semantic groups. Em *XIX Simpósio Brasileiro de Computação Aplicada à Saúde*, 318–323. 10.5753/sbcas.2019.6269.

de Souza, João Vitor Andrioli, Elisa Terumi Rubel Schneider, Josilaine Oliveira Cezar, Lucas Emanuel Silva, Yohan Bonescki Gumiel, Emerson Cabrera Paraiso, Douglas Teodoro & Claudia Maria Cabral Moro Barra. 2020. A multilabel approach to portuguese clinical named entity recognition. *Journal of Health Informatics* 12. 366–372.

Tjong Kim Sang, Erik F. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. Em

6th Conference on Natural Language Learning, (CoNLL).

Tjong Kim Sang, Erik F & Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. Em *7th Conference on Natural Language Learning (CoNLL)*, 142–147.

Toral, Antonio, Elisa Noguera, Fernando Llopis & Rafael Munoz. 2005. Improving question answering using named entity recognition. Em *10th International Conference on Applications of Natural Language to Information Systems*, 181–191. 10.1007/11428817_17.

Vale, Oto, Arnaldo Candido, Marcelo Muniz, Clarissa Bengtson, Lívia Cucatto, Gladis Almeida, Abner Batista, Maria C Parreira, Maria Tereza Biderman & Sandra Aluísio. 2008. Building a large dictionary of abbreviations for named entity recognition in portuguese historical corpora. Em *International Conference on Language Resources and Evaluation (LREC)*, 47–54.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need 1. 5999–6008.

Wagner Filho, Jorge A, Rodrigo Wilkens, Marco Idiart & Aline Villavicencio. 2018. The brWaC corpus: a new open resource for Brazilian Portuguese. Em *11th International Conference on Language Resources and Evaluation (LREC)*, 4339–4344.

Wang, Xinyu, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang & Kewei Tu. 2020. Automated concatenation of embeddings for structured prediction. ArXiv [cs.LG] v1. 10.48550/arXiv.2010.05006.

Yadav, Vikas & Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. ArXiv [cs.CL]. 10.48550/arXiv.1910.11470.

Zhao, Jun & Feifan Liu. 2008. Product named entity recognition in Chinese text. *Language Resources and Evaluation* 42(2). 197–217. 10.1007/s10579-008-9066-8.