BOLETIM DE PESQUISA E DESENVOLVIMENTO

52

ISSN 2764-2623
Dezembro/2022

Uma estratégia para a identificação de citações geográficas em textos técnico-científicos da área agrícola na língua portuguesa

Embrapa

ISSN 2764-2623
Dezembro/2022

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Agricultura Digital
Ministério da Agricultura, Pecuária e Abastecimento

BOLETIM DE PESQUISA E DESENVOLVIMENTO 52

Uma estratégia para a identificação de citações geográficas em textos técnico-científicos da área agrícola na língua portuguesa

Maria Fernanda Moura
Clara Mattos Medeiros

Embrapa Agricultura Digital
Campinas, SP
2022

Embrapa Agricultura Digital
Av. Dr. André Tosello, 209 - Cidade Universitária
Campinas, SP, Brasil
CEP. 13083-886
Fone: (19) 3211-5700
www.embrapa.br
www.embrapa.br/fale-conosco/sac

Comitê Local de Publicações
da Unidade Responsável

Presidente
Stanley Robson de Medeiros Oliveira

Secretária-Executiva
Maria Fernanda Moura

Membros
Adriana Farah Gonzalez, Alexandre de Castro, Carla Cristiane Osawa, Debora Pignatari Drucker, Ivan Mazoni, João Camargo Neto, João Francisco Gonçalves Antunes, Magda Cruciol

Revisão de texto
Adriana Farah Gonzalez

Normalização bibliográfica
Carla Cristiane Osawa

Projeto gráfico da coleção
Carlos Eduardo Felice Barbeiro

Editoração eletrônica
Letícia Mathias do Amaral Campos

Imagem da capa
Letícia Mathias do Amaral Campos

1ª edição
Publicação digital - PDF (2022)

Todos os direitos reservados.
A reprodução não autorizada desta publicação, no todo ou em parte, constitui violação dos direitos autorais (Lei nº 9.610).

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Agricultura Digital

Moura, Maria Fernanda,
Uma estratégia para a identificação de citações geográficas em textos técnico-científicos da área agrícola na língua portuguesa / Maria Fernanda Moura, Clara Mattos Medeiros. - Campinas : Embrapa Agricultura Digital, 2022.
PDF (19 p.) : il. color. - (Boletim de pesquisa e desenvolvimento / Embrapa Agricultura Digital, ISSN 2764-2623 ; 52).

1. Mineração de texto. 2. Reconhecimento de entidades nomeadas. 3. SpaCy. I. Medeiros, Clara Mattos. II. Título. III. Embrapa Agricultura Digital. IV. Série.

CDD (21. ed.) 658.382

Carla Cristiane Osawa (CRB-8/10421)
© Embrapa, 2022

Sumário

Uma estratégia para a identificação de citações geográficas em textos técnico-científicos da área agrícola na língua portuguesa

Maria Fernanda Moura[1]
Clara Mattos Medeiros[2]

Resumo – A proposta deste trabalho é utilizar um reconhecedor de entidades nomeadas para a língua portuguesa, a fim de extrair metadados de citações a localizações geográficas brasileiras em publicações técnico-científicas do domínio agrícola. A estratégia adotada consistiu na identificação de padrões de citação a localidades de interesse, a partir da criação de uma coleção dourada, e a consequente personalização do reconhecedor de entidades nomeadas da biblioteca SpaCy. Os experimentos conduzidos, com os novos treinamentos da SpaCy, mostram uma revocação média de 0,92 e uma precisão média de 0,95, permitindo aceitar que a acurácia para a identificação das localidades nos textos seja bastante confiável.

Termos para indexação: Reconhecimento de Entidades Nomeadas; Processamento de Linguagem Natural; SpaCy; Geolocalização; Mineração de Textos.

A strategy for the identification of geographical citations in technical-scientific texts in the agricultural area in Portuguese

Abstract – The purpose of this work is to use a named entity recognizer for the Portuguese language in order to extract metadata of citations to Brazilian geographic locations in technical-scientific publications in the agricultural domain. The strategy adopted consisted of identifying patterns for locality cita-

[1] Estatística, doutora em Ciências da Computação, pesquisadora da Embrapa Agricultura Digital, Campinas, SP.

[2] Estudante de Engenharia da Computação, bolsista de Iniciação Científica na Embrapa Agricultura Digital, Campinas, SP.

tions of interest, from the creation of a golden collection, and the consequent customization of the named entity recognizer from the SpaCy library. The experiments carried out with the new SpaCy trainings show an average recall of 0.92 and an average precision of 0.95, allowing us to accept that the accuracy for the identification of localities in the texts is quite reliable.

Index terms: Named Entities Recognition; Natural Language Processing; SpaCy; Geolocation; Text Mining

Introdução

A Empresa Brasileira de Pesquisa Agropecuária (Embrapa) possui um rico acervo de dados, informações e conhecimentos agropecuários resultantes de suas pesquisas, tecnologias e publicações desenvolvidas desde a década de 70, que podem ser recuperados por meio de ferramentas de busca, como as ferramentas do Repositório Alice[3], que permitem acesso à informação científica produzida pela Embrapa.

Em um de seus projetos para mapear o conhecimento, as informações e os dados gerados sobre tecnologias e produção científica associada ao tema "pastagens degradadas", decidiu-se por utilizar técnicas de mineração de textos para extrair e relacionar as informações de interesse, bem como observar tendências de temas e subtemas no tempo e espaço. Essa decisão deve-se à existência de mais de 6 mil obras técnico-científicas nesse tema, disponíveis em repositórios da Embrapa, além de tecnologias disponíveis e patentes relacionadas a esse tema.

Embora o acervo da Embrapa conte com metadados qualificados, a informação de cobertura geográfica encontra-se apenas no contexto da obra ou na descrição do uso da tecnologia. Por exemplo, em que região ou microrregião geográfica uma tecnologia agrícola pode ser aplicada. Para obter essa informação foi necessário identificá-la no contexto das publicações técnico-científicas na língua portuguesa utilizando processos de Reconhecimento de Entidades Nomeadas (REN), ou *Named Entity Recognition* (NER), em inglês. NER refere-se a uma tarefa de identificação seguida pela classifica-

[3] Disponível em: https://www.alice.cnptia.embrapa.br/.

ção das várias entidades nomeadas em um texto (Nasar et al., 2022), como nomes de pessoas, organizações, locais, expressões de tempos e quantidades. Para treinar uma ferramenta de NER usa-se um córpus previamente anotado, que é um conjunto de documentos onde são anotados padrões para a identificação de entidades nomeadas de interesse, de acordo com a revisão analítica de Marrero et al. (2013).

Há algumas ferramentas e metodologias de domínio público disponíveis para o reconhecimento de entidades nomeadas. A ferramenta TopExtract (Takemura et al., 2016) foi desenvolvida e utilizada pela Embrapa em alguns projetos, porém ela utilizava uma versão disponível do OpenCalais que era própria para a língua inglesa. Assim, era necessário traduzir os textos em português para o inglês, identificar as entidades nomeadas de locais e aplicar uma desambiguação de topônimos com base em dicionários geográficos. O Apache OpenNLP é uma biblioteca Java de Processamento de Linguagem Natural (PLN) que suporta diversas tarefas, como tokenização, etiquetagem morfossintática e reconhecimento de entidades nomeadas, mas, até o momento, não há um reconhecedor de entidades nomeadas para o português (Fonseca et al., 2015). O FreeLing é uma ferramenta *open source* escrita em C++ com dois métodos para o reconhecimento de entidades nomeadas: via análise de padrões morfossintáticos e via aprendizagem de máquina (Fonseca et al., 2015). Por não ter recebido evolução desde 2020, ocorreram alguns problemas que inviabilizaram sua instalação. O Stanza, uma biblioteca em Python desenvolvida por um grupo de Stanford, é uma ferramenta de PLN com modelos pré-treinados para mais de 50 línguas (Gonçalves et al, 2020). Em versões recentes, porém, o REN para a língua portuguesa foi excluído. O Natural Language Toolkit (NLTK), em Python, uma das plataformas mais utilizadas para PLN (Gonçalves et al., 2020), apresenta alguns erros no REN em português, também por não ter constante manutenção. O NLPyPort é um pipeline montado a partir do pipeline NLTK, adicionando e alterando seus elementos para melhor processamento do português (Gonçalves et al., 2020). Por falta de evoluções desde 2020, foram encontrados erros na sua instalação. A ferramenta PAMPO (Rocha et al., 2016), cujo objetivo era ser uma solução para estruturas semânticas não triviais na língua portuguesa, embora apresente bons resultados ao identificar entidades nomeadas, falha em não classificá-las, como indicar localidades e pessoas, na última versão disponível. A SpaCy, uma biblioteca de software de código aberto para pro-

cessamento avançado de linguagem natural, escrita nas linguagens de programação Python e Cython (Gonçalves et al., 2020), mostrou-se bastante adequada para localizar entidades nomeadas em textos na língua portuguesa, uma vez que possui um córpus bem anotado e, principalmente, é uma biblioteca que tem recebido constante evolução, ao contrário de várias outras que não têm recebido evoluções e muitas vezes apresentam erros de execução ou instalação.

Devido à necessidade de reconhecer referências específicas a regiões brasileiras, por exemplo, “litoral norte de São Paulo” ou “zona de caatinga”, que dificilmente são reconhecidas como entidades nomeadas em textos, foi necessário personalizar o NER da SpaCy, treinando-a com um córpus do domínio agrícola brasileiro e marcado de forma específica. Neste trabalho apresenta-se a metodologia utilizada na identificação das citações a localizações geográficas brasileiras em textos da área agrícola, a criação de uma base de textos dourada para uso na validação dos resultados e o processo de marcação de um córpus para o treino da SpaCy, além da discussão dos experimentos e os resultados obtidos.

A solução, aqui apresentada, mostrou-se eficaz, apresentando uma revocação média de 0,92 e uma precisão média de 0,95 com os textos analisados, permitindo que a informação extraída possa ser utilizada como um metadado confiável para as obras científicas e resultados tecnológicos no domínio agrícola do território brasileiro. Alguns casos são mais difíceis de serem solucionados, como “área de produção de babaçu”, pois não se restringem à questão geográfica. Seria necessário fazer um trabalho específico para áreas de predomínio de alguma atividade econômica, o que será possível transformar-se em trabalho futuro. A despeito de algumas restrições, como a personalização do NER de acordo com as necessidades específicas do projeto em desenvolvimento, e o modo como são referenciadas algumas regiões brasileiras, a solução aqui apresentada pode ser utilizada em problemas similares.

Material e Métodos

O processo utilizado baseia-se na necessidade de encontrar as citações a locais geográficos, particularmente do território brasileiro, que nem sempre são explicitamente citados nos textos e nos formatos esperados por fer-

ramentas de REN. Na Figura 1, ilustra-se a metodologia empregada neste trabalho, que envolve a delimitação do problema, isto é, quais entidades nomeadas interessam ao processo e a consequente criação de uma coleção dourada, testes realizados com a configuração disponível da SpaCy, análise de padrões encontrados versus esperados e a personalização passo a passo do NER da SpaCy. Cada passo da Figura 1 é explicado nas próximas subseções.

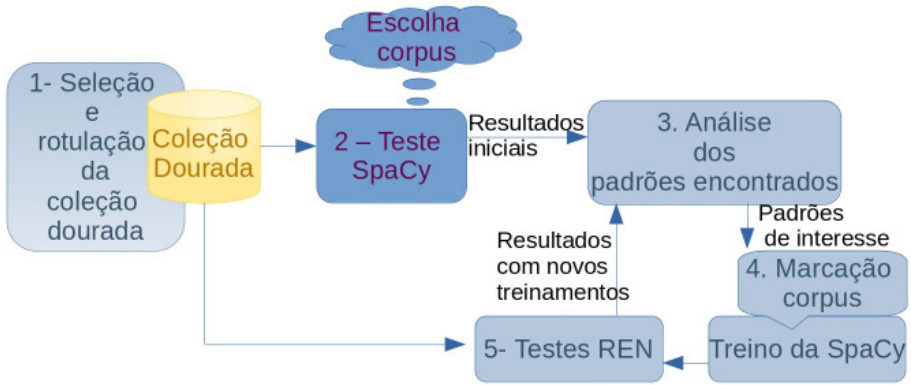

Figura 1. Estratégia proposta para personalizar o NER da SpaCy

Seleção e rotulação da coleção dourada

O primeiro passo é identificar o domínio aos quais pertencem os textos de interesse e o tipo de informação, isto é, os vários padrões em que uma entidade nomeada geográfica de interesse é apresentada nesses textos, do modo como se deseja reconhecê-las. Assim, para o tratamento e rotulação dos dados, foi selecionada uma planilha com 350 excertos de textos do banco de Tecnologias, Produtos e Serviços (TPS) da Embrapa , todos relacionados à questão de pastagens em áreas brasileiras. Os textos referem-se ao título da tecnologia, produto ou serviço e uma breve descrição, todos na língua portuguesa. Por vezes a descrição contém a cobertura geográfica, porém na maioria dos textos essa cobertura precisa ser inferida por meio da sua leitura. Cada texto foi lido e interpretado por duas pessoas e, manualmente, foram anotados os locais citados em um novo campo da planilha. Esta planilha pas-

sou a ser utilizada como uma coleção dourada, ou seja, um gabarito do que o software desenvolvido deveria ser capaz de encontrar quando analisasse aqueles textos. Um pequeno extrato da planilha com anotações manuais é exibido na Tabela 1.

Tabela 1. Extrato da coleção dourada

Ano de lançamento	Nome do Ativo Tecnológico	Descrição do Ativo Tecnológico
2021	Trigo - BRS Tarumaxi	BRS Tarumaxi é uma cultivar de trigo que está sendo indicada tanto para a produção de forragem (adaptada ao período prolongado de pastejo), quanto para a produção de grãos, nos estados do Rio Grande do Sul e de Santa Catarina. É uma cultivar de ciclo tardio que pode ser semeada até 40 dias antes do período indicado para cultivares precoces, quando submetida a corte ou pastejo, com aptidão para múltiplos usos (somente pasto, pasto e grãos, pasto e feno ou pasto e silagem), com ênfase na produção de pasto em sistemas de integração lavoura e pecuária no Sul do Brasil. BRS Tarumaxi tem alta capacidade de rebrote e afilhamento, com médias de rendimento de massa seca de forragem e de grãos superiores às médias da cultivar BRS Tarumã (cultivar amplamente utilizada para forragem no RS), em 16,5% e 24% respectivamente, após dois cortes/pastejos, em três anos de avaliação (2017 a 2019).
2021	TropiCow: Processo de produção de bovino de corte, tropicalmente adaptado, com o uso de bovinos de raças brasileiras localmente adaptadas em cruzamento industrial	É um processo de cruzamento industrial para a obtenção do “Boi Tropical” que demonstra como utilizar recursos genéticos brasileiros adaptados aos trópicos para obter bom desempenho ponderal, bom rendimento de carcaça e carne macia, com melhores ganhos potenciais em sistemas de integração lavoura-pecuária-floresta, em pastagens artificiais e nativas de clima tropical do Brasil. A principal aplicação é a utilização como componente animal nos sistemas de integração lavoura-pecuária-floresta (ILPF) para produção de carne de qualidade de forma sustentável na região do Matopiba.
2020	Arroz - BRS A502	A Embrapa apresenta a BRS A502, uma cultivar de arroz para o sistema de terras altas (sequeiro). Essa cultivar destaca-se pela tolerância ao acamamento e alta estabilidade de rendimento de grãos inteiros, que permite ao produtor uma maior flexibilidade de colheita. Essas caracteristicas, associadas ao alto potencial produtivo e à excelente qualidade industrial e culinária de grãos, fazem dessa cultivar uma excelente opção para sistemas de produção de grãos e de renovação de pastagens.A BRS A502 é indicada para os seguintes estados: GO, MA, MT, PA, PI, RO. Para maiores informações, leia as publicações técnicas deste material e consulte sempre um Engenheiro Agrônomo.
2020	Práticas agropecuárias para validação de Sistemas de Integração Lavoura-Pecuária-Floresta em área de ocorrência de babaçu.	Ajuste e validação de modelos de sistemas de ILPF para recuperação de pastagem em área de ocorrência de babaçu. Aos modelos de pecuária praticados em áreas de ocorrência de babaçu, que em sua grande maioria são desenvolvidos em áreas de pastagens degradadas.

Continua

2019	Manejo do pastejo intensivo das cultivares BRS Paiaguás e BRS Ipyporã no bioma Mata Atlântica.	Trata-se de uma recomendação de manejo das cultivares BRS Paiaguás e BRS Ipyporã de *Brachiaria sp.* em sistemas intensivos de produção. A principal aplicação é para sistemas intensivos de produção de leite a pasto nas regiões sudeste, centro-oeste e norte do Brasil.
2018	Programa para valoração dos serviços ecossistêmicos das pastagens nativas (Sistema de informação ou análise)	Desenvolver um sistema de monitoramento e suporte a decisão para avaliar os serviços ambientais das pastagens nativas do Pantanal

Teste da SpaCy com córpus disponibilizado

A escolha da biblioteca SpaCy deu-se pela disponibilização de um mecanismo REN para várias línguas, especialmente a língua portuguesa e por estar recebendo constante evolução e manutenção, além de ser publicamente disponibilizada. A partir da construção da coleção dourada é possível verificar se os treinos disponíveis na SpaCy, com algum córpus na língua portuguesa, apresentam um resultado satisfatório em relação ao esperado. Esse passo é de interesse, mesmo que se saiba a priori que os resultados não sejam satisfatórios, pois auxilia a compreender como criar novos exemplos de treinamento. O córpus “pt_core_news_sm”[4] é bastante completo, por isso foi utilizado no momento em que este processo foi montado.

Análise dos padrões reconhecidos

A análise subjetiva dos resultados obtidos é fundamental para a construção de exemplos de treino. Quanto maior a massa de dados (número e tamanho dos textos), mais difícil é executar a análise. Porém, quanto maior o número de exemplos de treino, com variadas formas de se identificar as entidades, melhor será a acurácia dos resultados. Assim, além de verificar o número de acertos e erros, conferindo-os por leitura e interpretação humanas, optou-se por calcular:

• Precisão (P, precision): a relação entre o número de acertos (p, positivos) e o número total de entidades reconhecidas como corretas (p + fp, onde fp são as classificadas como corretas, porém incorretas). Esta medida nos

[4] Disponível em: https://spacy.io/models/pt.

dá indícios da capacidade do REN, mediante os exemplos de treinamento, de evitar a obtenção de falsos positivos.

• Revocação (R, recall): a relação entre o número de acertos (p) e o total de acertos esperado (número de entidades que deveriam ter sido reconhecidas nos textos). Esta medida indica o quão boa foi a escolha do REN, mediante os exemplos de treinamento.

• Medida F: indica a média harmônica entre P e R (2*(P*R)/(P+R)). Quanto mais próximo de um (1) estiver o valor da F, mais confiável é a acurácia do REN, com seus exemplos de treinamento.

O critério para aceitar o REN, com o córpus de treinamento fornecido, obrigatoriamente passa pela análise subjetiva e pelas medidas obtidas. Para a análise subjetiva, quando o número de textos for muito grande, sugere-se sortear aleatoriamente, entre os resultados que apresentarem maior número de classificações incorretas, um número razoável de textos para conferir e decidir se é necessário criar novos exemplos de treino.

Marcação de córpus e treino da Spacy

O NER da Spacy treinado com o córpus “pt_core_news_sm” apresentou alguns falsos positivos muito comuns, como classificar um nome de cultivar como localidade. Por exemplo, para o texto:

> BRS Tarumaxi é uma cultivar de trigo que está sendo indicada tanto para a produção de forragem (adaptada ao período prolongado de pastejo), quanto para a produção de grãos, nos estados do Rio Grande do Sul e de Santa Catarina. É uma cultivar de ciclo tardio que pode ser semeada até 40 dias antes do período indicado para cultivares precoces, quando submetida a corte ou pastejo, com aptidão para múltiplos usos (somente pasto, pasto e grãos, pasto e feno ou pasto e silagem), com ênfase na produção de pasto em sistemas de integração lavoura e pecuária no Sul do Brasil. BRS Tarumaxi tem alta capacidade de rebrote e afilhamento, com médias de rendimento de massa seca de forragem e de grãos superiores às médias da cultivar BRS Tarumã (cultivar amplamente utilizada para forragem no RS), em 16,5% e 24% respectivamente, após dois cortes/pastejos, em três anos de avaliação (2017 a 2019).,

são encontradas as seguintes entidades nomeadas:

BRS LOC
Tarumaxi MISC
Rio Grande do Sul LOC
Santa Catarina LOC
Sul do Brasil LOC
Tarumaxi MISC
BRS Tarumã LOC
RS LOC

Porém, “Tarumaxi” é parte do nome da cultivar “BRS Tarumaxi” e “Tarumã” também é parte do nome da outra cultivar “BRS Tarumã”. Os nomes dos estados estão corretos, bem como a região “Sul do Brasil”. No entanto, em outros exemplos, se estiver escrito “sul do Brasil” (a palavra “sul” com letra minúscula), a entidade nomeada não é reconhecida.

Assim, com os exemplos de entidades nomeadas incorretamente encontradas e outras não encontradas que são de interesse para o tema em questão, formam-se exemplos de treinamento. Por isso, é importante criar a coleção dourada, pois a partir dela estabelecem-se os padrões que interessa reconhecer.

No exemplo a seguir, pode-se observar como são marcados os casos de treino para a SpaCy:

```
import spacy
import random
from spacy.util import minibatch, compounding
from pathlib import Path
from spacy.training.example import Example
nlp=spacy.load(‘pt_core_news_sm’)
# Getting the pipeline component
ner=nlp.get_pipe(“ner”)

TRAIN_DATA = [
(“BRS Tarumaxi é uma cultivar de trigo que está sendo
indicada tanto para a produção de forragem (adaptada ao
```

período prolongado de pastejo), quanto para a produção de grãos, nos estados do Rio Grande do Sul e de Santa Catarina. É uma cultivar de ciclo tardio que pode ser semeada até 40 dias antes do período indicado para cultivares precoces, quando submetida a corte ou pastejo, com aptidão para múltiplos usos (somente pasto, pasto e grãos, pasto e feno ou pasto e silagem), com ênfase na produção de pasto em sistemas de integração lavoura e pecuária no Sul do Brasil. BRS Tarumaxi tem alta capacidade de rebrote e afilhamento, com médias de rendimento de massa seca de forragem e de grãos superiores às médias da cultivar BRS Tarumã (cultivar amplamente utilizada para forragem no RS), em 16,5% e 24% respectivamente, após dois cortes/pastejos, em três anos de avaliação (2017 a 2019). “, {“entities”: [(188, 206, “LOC”), (211, 225, “LOC”), (557, 570, “LOC”), (785, 787, “LOC”)]}),

O conjunto “TRAIN-DATA” contém os exemplos de treino, isto é, cada texto componente do córpus anotado. No exemplo estão ilustrados apenas os dois primeiros textos. Note que, no primeiro texto, a especificação “{“entities”: [(188, 206, “LOC”), (211, 225, “LOC”), (557, 570, “LOC”), (785, 787, “LOC”)]},” corresponde às entidades de interesse no texto, com as respectivas posições de início e final de cada entidade de localidade no texto e o tipo da entidade - neste caso, todas são de localização (LOC).

Desta forma, a partir de textos da coleção dourada e resultados de erros que se vão apresentando, formam-se novos exemplos de treinamento. Inicialmente, dos 350 textos da coleção dourada, os 99 primeiros exemplos foram separados e anotados para formar o córpus de treino para a SpaCy.

Teste com o REN das localizações

Uma vez que o REN tenha sido treinado com os exemplos de interesse, geram-se os novos resultados e verifica-se o número de erros e acertos. Ressalta-se que os exemplos da coleção dourada utilizados para treinar o córpus não devem ser incluídos nesses testes, pois a máquina foi treinada com eles e a sua reapresentação aos testes causaria um *overfitting* (quando um modelo estatístico se ajusta muito bem ao conjunto de dados anterior-

mente observado, mas não necessariamente é eficaz para prever novos resultados). O programa que executa os testes é responsável por gravar uma nova planilha com as entidades encontradas e já classificadas como acertos (p) ou falsos positivos (fp), a partir dos quais calculam-se precisão e revocação. As comparações das entidades encontradas com as do gabarito da coleção dourada são feitas por funções que procuram a maior similaridade entre elas. Então, volta-se ao passo 3 (item 2.3), referente à análise de padrões.

Resultados e Discussão

Os experimentos, bem como a criação de novos exemplos de treino, foram conduzidos com base na coleção dourada criada, para a qual foram identificadas, por leitura humana, 369 localidades de interesse em 348 textos sobre tecnologias, produtos e serviços relacionados a pastagens no Brasil.

Observou-se que, rodando a SpaCy, com o córpus "pt_core_news_sm", foram encontradas 720 entidades do tipo LOC, sendo 320 acertos e 400 falsos positivos, de acordo com as definições de locais deste trabalho. Por exemplo, foram encontradas e classificadas entidades do tipo LOC, tais como:

> "BRS Tarumaxi", que corresponde a uma cultivar e não ao nome de um local;
>
> "Universidade Federal do Rio Grande do Sul", que corresponde ao nome de uma instituição, e não é interessante para este trabalho tratá-la como localidade;
>
> "Integração Lavoura-Pecuária-Floresta (ILPF)", que corresponde a uma estratégia de produção agrícola.

Embora alguns dos falsos positivos considerados sejam localidades de acordo com uma definição mais geral, para essa pesquisa não são pertinentes, pois os locais significativos são aqueles nos quais ocorre produção agrícola no Brasil.

Observados os falsos positivos e os falsos negativos, isto é, tanto as entidades nomeadas que foram encontradas e não eram desejadas ou incorretas quanto as que eram desejadas e não foram encontradas, iniciou-se o processo de anotação de exemplos de treino. Os 99 primeiros exemplos da planilha dourada foram anotados e utilizados como exemplos de treino.

Após aplicar os testes do REN à coleção dourada reduzida (com exceção dos 99 primeiros excertos de texto), foi verificado se os locais encontrados, em letras minúsculas, eram iguais aos resultados esperados, também em minúsculas. Isso foi necessário para evitar que um local encontrado no texto como “sul da Bahia” fosse entendido como erro quando comparado com uma anotação do gabarito “Sul da Bahia”. Além disso, foi verificado também se a resposta encontrada e a resposta esperada possuem intersecções. Por exemplo, caso a resposta esperada fosse “região Norte do Brasil” e a Spacy encontrasse “Norte do Brasil”, também seria contado como um acerto, pois a resposta encontrada está contida na resposta esperada.

Dessa forma, os locais encontrados para cada excerto de texto na planilha reduzida foram comparados com os locais esperados para aquele trecho, e foram calculados o número de acertos (tp) e de falsos positivos (fp). Caso houvesse sido reconhecido como local pelo SpaCy algo que não estava no gabarito previamente anotado, configurava-se como um falso positivo. Já um falso negativo é algum local marcado no gabarito que não foi encontrado pela ferramenta. Finalmente, foram contados quantos locais existiam na planilha (via a quantidade de elementos do gabarito) e somado ao número de acertos, falsos positivos e falsos negativos.

Na Tabela 2 são apresentados os resumos dos resultados dos treinamentos realizados. Inicialmente foram usados 30 exemplos de treino, que foram sendo aumentados de acordo com os padrões que eram identificados como de interesse e não apresentados como acertos. Até os 99 primeiros exemplos de treinos, todos os exemplos vieram da planilha da coleção dourada; depois foram sendo inseridos exemplos mais específicos. O objetivo era chegar a uma revocação e precisão maiores que 90%.

Um padrão de erro comumente encontrado era a não identificação de uma sigla de estado, quando entre parênteses, após vírgula ou barra. Por isso, em exemplos como “município de Silva Jardim, RJ”, “município de Silva Jardim (RJ)” ou “município de Silva Jardim/RJ”, o local “RJ” não era encontrado. Para que isso se revertesse, foram adicionados exemplos que explorassem essas exceções. Assim, foram sendo otimizados os resultados obtidos com o REN, até se chegar aos 118 exemplos, que permitiram uma precisão de 0.94, revocação de 0.92 e, consequentemente, uma F de 0.93, o que indica uma acurácia bastante alta.

Tabela 2. Acompanhamento da evolução do algoritmo com o aumento de exemplos de treino

Quantidade de locais na planilha de teste	Quantidade de exemplos nos treinos	Número de acertos tp	Falsos positivos fp	Falsos negativos fn	Precisão	Revocação
396	30	310	85	59	0,785	0,840
	50	312	87	57	0,782	0,846
	70	315	60	54	0,840	0,854
	99	312	68	57	0,821	0,846
	110	326	34	43	0,906	0,883
	118	341	21	28	0,942	0,924

Além dos textos da coleção dourada, foram analisados mais trinta textos sobre pastagens aleatoriamente escolhidos do repositório Alice da Embrapa (Visoli et al. 2017), que oferece publicações científicas em língua portuguesa. Dos textos científicos utilizaram-se apenas o título e o resumo, a fim de anotar as regiões brasileiras citadas e futuramente anexá-las à coleção dourada. Também foram escolhidos, aleatoriamente, 60 textos de Tecnologias, Produtos e Serviços da Embrapa, em língua portuguesa, para identificação de localidades, sem intersecção com a coleção dourada. Os resultados da aplicação do REN treinado com os 118 exemplos podem ser observados na Tabela 3. Nota-se que a F mantém um valor acima de 0,9, o que indica que a acurácia dos resultados é bastante alta.

Tabela 3. Resultados com a utilização da versão final do córpus.

Planilha	Quantidade de Locais na Planilha	Acertos tp	Falsos Positivos fp	Falsos Negativos fn	precision tp/(tp+fp)	recall tp/ (tp+fp)	Medida F
Publicações da Embrapa	39	38 97%	3 7%	1 3%	0,93	0,97	0,93
TPS	118	103 87%	3 3%	15 13%	0,97	0,87	0,92

Considerações Finais

Nota-se pelos valores obtidos de precisão e revocação que há grande evidência de que os treinamentos realizados sobre o REN, do modo como colocados, ampliaram muito a capacidade do programa desenvolvido em identificar citações geográficas em textos técnico-científicos da área agrícola na língua portuguesa.

É importante ressaltar a especificidade do córpus anotado, construído com o propósito de reconhecer apenas locais em textos com termos técnicos agrícolas. Para reconhecer outras entidades nomeadas, é necessário que estas sejam também anotadas. Além disso, ainda é possível melhorar o trabalho de anotação para reconhecer outras entidades de interesse, como “área de produção de babaçu”, “áreas de clima equatorial” e outras similares, necessidade que será avaliada em trabalhos futuros.

O próximo trabalho a ser realizado é utilizar os resultados do REN treinado neste em um processo de desambiguação, como proposto por Takemura et al. (2016), e encontrar a cobertura geográfica completa de um texto, correspondente ao polígono de abrangência das citações de geolocalidades(?) no texto. Para uma base de tecnologias, produtos e serviços pode indicar a sua abrangência.

Referências

FONSECA, E. B.; CHIELE, G. C.; VIEIRA, R.; VANIM, A. A.;. **Reconhecimento de entidades nomeadas para o português usando o OpenNLP**. 2015. Disponível em: https://repositorio.pucrs.br/dspace/handle/10923/14040. Acesso em: 20 out. 2022.

GONÇALVES, M.; COHEUR, L.; BAPTISTA, J.; MINEIRO, A.. Avaliação de recursos computacionais para o português. **Linguamática**, v. 12, n,. 2, p. 51-68, 2020. DOI: 10.21814/lm.12.2.331.

MARRERO, M.; URBANO, J.; SÁNCHEZ-CUADRADO, S.; MORATO, J.; GÓMEZ-BERBÍS, J. M. Named entity recognition: fallacies, challenges and opportunities. **Computer Standards & Interfaces**, v. 35, n. 5, p. 482-489, Sept. 2013. DOI: 10.1016/j.csi.2012.09.004.

NASAR, Z.; JAFFRY, S. W.; MALIK, M. K. Named entity recognition and relation extraction: state-of-the-art. **ACM Computing Surveys**, v. 54, n. 1, p. 1-39, Jan. 2022. DOI: 10.1145/3445965.

ROCHA, C.; JORGE., A.; SIONARA., R.; BRITO, P.; PIMENTA, C.; REZENDE, S. **PAMPO**: using pattern matching and pos-tagging for effective Named Entities recognition in Portuguese. arXiv:1612.09535, 2016. DOI: 10.48550/arXiv.1612.09535.

TAKEMURA, C. M.; SILVA, G. B. S. da; OLIVEIRA, S. R. de M.; MOURA, M. F. Desambiguação de topônimos usando dicionários geográficos. In: SEMINÁRIO DA REDE AGROHIDRO, 4., 2016, Brasília, DF. **Água e agricultura: incertezas e desafios para a sustentabilidade frente às mudanças do clima e do uso da terra:** anais. Planaltina, DF: Embrapa Cerrados, 2016. 290 p. Editores técnicos: Lineu Neiva Rodrigues; Maria Fernanda Moura; Raimundo Cosme de Oliveira Júnior. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/152436/1/desambiguacao-CELINA.pdf. Acesso em: 20 out. 2020.

VISOLI, M. C.; GONZALES, L. E.; LEMOS, T. V. A.; ALVES, A. A.; BERTIN, P. R. B.; SILVA, A. R. da; SIMAO, V. P. M.; VACARI, I.; PRAXEDES, M. G. G. **Alice:** Repositório Acesso Livre à Informação Científica da Embrapa. Versão 2.0. Campinas: Embrapa Informática Agropecuária, 2017.

Embrapa

Agricultura Digital

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO

PÁTRIA AMADA BRASIL GOVERNO FEDERAL

CGPE 17927