Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 29, Num. 01 (2022) 28-41

RESEARCH ARTICLE

System for Identifying Pests and Diseases in Soybean Crop through Natural Language Processing

Sistema para Identificação de Pragas e Doenças na Cultura da Soja por meio de Processamento de Linguagem Natural

Caroline Roque e Faria[1]*,Cinthyan Renata Sachs C. de Barbosa[1]

Resumo: The presence of technologies in the agronomic field has the purpose of proposing the best solutions to the challenges found in agriculture, especially to the problems that affect cultivars. One of the obstacles found is to apply the use of your own language in applications that interact with the user in Brazilian Agribusiness. Therefore, this work uses Natural Language Processing techniques for the development of an automatic and effective computer system to interact with the user and assist in the identification of pests and diseases in soybean crop, stored in a non-relational database repository to provide accurate diagnostics to simplify the work of the farmer and the agricultural stakeholders who deal with a lot of information. In order to build dialogues and provide rich consultations, from agriculture manuals, a data structure with 108 pests and diseases with their information on the soybean cultivar and through the spaCy tool, it was possible to pre-process the texts, recognize the entities and support the requirements for the development of the conversacional system.
Keywords: Digital Agriculture — Intelligent systems — Natural Language Processing

Resumo: A presença das tecnologias no campo agronômico tem a finalidade de propor as melhores soluções para os desafios encontrados na agricultura, especialmente aos problemas que afetam as cultivares. Um dos obstáculos encontrados é aplicar o uso de sua própria linguagem em aplicativos que interajam com o usuário no Agronegócio Brasileiro. Assim, o presente trabalho utiliza técnicas de Processamento de Linguagem Natural para o desenvolvimento de um sistema computacional automático e efetivo para interagir com o usuário e auxiliar na identificação de pragas e doenças na cultura da soja, armazenados em um repositório de banco de dados não relacional para fornecer diagnósticos precisos para simplificar o trabalho do agricultor e das partes agrícolas interessadas que lidam com muitas informações. Com o intuito de construir diálogos e proporcionar consultas ricas, a partir de manuais da agricultura, uma estrutura de dados com 108 pragas e doenças com as suas informações na cultivar da soja e por meio da biblioteca spaCy foi possível pré-processar os textos, reconhecer as entidades nomeadas e suportar os requisitos para o desenvolvimento do sistemas conversacional.
Palavras-Chave: Agricultura digital — Sistemas inteligentes — Processamento de Linguagem Natural

[1] *Programa de Pós-Graduação em Ciência da Computação, Universidade Estadual de Londrina (UEL), Londrina - Paraná, Brazil*
***Corresponding author**: caroline.rf@outlook.com
DOI: http://dx.doi.org/10.22456/2175-2745.107149 • **Received:** 01/09/2020 • **Accepted:** 25/11/2021

1. Introdução

O mundo está passando por revoluções tecnológicas, principalmente no campo agronômico, o qual tem como objetivo proporcionar melhores soluções para o desenvolvimento e aumentar a produtividade da cultivar.

Devido à importante participação do Brasil na produção de soja e ao aumento de produtividade, ter uma ferramenta que auxilie o produtor sobre o patógeno, garante a sanidade das plantas, significa melhor produção e, desse modo, o agricultor pode lidar com as necessidades da lavoura.

Atualmente, existem inúmeros invasores que prejudicam a cultura da soja e que estão distribuídos em regiões produtoras, que em decorrência da falta de preparo em fazer uma análise do solo antes de realizar o plantio ou por fazer o uso de agrotóxicos de maneira desequilibrada ou aração incorreta, não rotacionam culturas e usam de forma irracional os produtos para o manejo das pragas.

Dificilmente uma tecnologia resolverá todos os problemas no campo, mas a combinação de diferentes ferramentas de controle permite manter o equilíbrio na lavoura. Sendo assim, a identificação tradicional das pragas e doenças na plantação é feita por um especialista (pragueiro), a partir da observação da planta, qualidade dos grãos, lesões na haste, coloração da

folhagem, podridão das raízes etc.

De acordo com Chowdhury [1], fazer uma análise visual da cultivar para detectar o vilão é ineficiente e difícil para plantações de grandes proporções, visto que o ataque acontece de forma inesperada e agressiva, além do procedimento fornecer uma acurácia relativamente baixa que requer um profissional qualificado e bem treinado para a realização de tal função [2].

O gerenciamento de dados agrícolas depende de informações adquiridas de diversas tecnologias e sistemas que sejam capazes de auxiliar na tomada de decisão. Portanto, é fundamental mapear os dados (principais características de identificação de pragas e doenças na cultura da soja) e padronizá-los.

Com o enorme fluxo de informações, essas tecnologias necessitam elaborar soluções para entender a linguagem e uma alternativa é o uso de Processamento de Linguagem Natural (PLN) aplicados para grandes volumes de dados. Essas técnicas pretendem analisar e extrair as informações de uma determinada cultura para proporcionar um diagnóstico com a finalidade de potencializar a produção dos agrônomos, agricultores, especialistas, estudantes e interessados na área.

Para Jurafsky e Martin [3], PLN tem como objetivo extrair textos em linguagem natural (LN) e executar tarefas relevantes, permitindo o diálogo entre homem e máquina, melhorando a comunicação humano-computador ou fazer processamento de texto ou fala (discurso).

Dada a complexidade das linguagens naturais, este trabalho tem o propósito de elaborar um sistema de pré-consulta ao usuário que, a partir da classificação dos vilões na planta torna mais eficaz o acesso à informação e o auxilia a tomar a melhor decisão sem a necessidade da presença de um especialista, com tempo mínimo e menos risco ao utilizar agrotóxico.

Problemas linguísticos dessas perguntas em LN foram levantados e estão sendo estudadas soluções para a implementação de um sistema conversacional, que usa possíveis perguntas relacionadas aos profissionais da área agronômica para pesquisas quanto às pragas e doenças na cultura da soja.

Este trabalho possui a seguinte estrutura: a seção 2 apresenta os trabalhos relacionados; a seção 3 é destinada à fundamentação teórica; a seção 4 aborda a importância de uma Interface em Linguagem Natural para Banco de Dados; a Arquitetura de um Sistema de Interpretação é apresentada na seção 5; a metodologia e desenvolvimento do trabalho encontra-se na seção 6; o banco de dados desenvolvido está detalhado na seção 7; a seção 8 exibe o sistema conversacional para Identificação das Pragas e Doenças da Soja. Por fim, na seção 9 tem as considerações finais.

2. Trabalhos relacionados

A agricultura digital está cada vez mais presente no campo e tem como objetivo aumentar a produção em um menor espaço, ampliando os lucros e diminuindo os prejuízos. O gerenciamento de dados agronômicos depende de informações adquiridas de diversas tecnologias e sistemas que sejam capazes de auxiliar na tomada de decisão nas técnicas agrícolas. Assim, é fundamental mapear os dados (principais características de identificação de pragas e doenças na cultura da soja) e padronizá-los.

Devido à grande quantidade de informações armazenadas e disponibilizadas em rede aberta, esta seção apresenta uma pesquisa focada na aplicação de métodos computacionais na soja e ontologias no campo.

Dentre os trabalhos nessa área, conta-se com o trabalho para a detecção e predição da ferrugem asiática da soja [4] , associados às técnicas de processamento digital de imagens e de mineração de dados, visando a obtenção de modelos preditivos de severidade nos diferentes estágios de desenvolvimento da soja.

Foram realizados estudos por Rosa [5] para a classificação das pragas mais comuns do estado do Paraná, que são a *Anticarsia gemmatalis, Helicoverpa armigera* e *Spodoptera cosmioide*, e utiliza rede neural convolucional profunda, na qual um sistema computacional do tipo cliente-servidor foi criado a fim de prover a classificação de pragas mediante serviço, baseada na arquitetura *Inception V3*. Foi obtido um índice de acerto de 92,5%.

Um estudo da cultura da soja [6] foi realizado pela elaboração de um banco de imagens com mais de 15 mil imagens do solo, soja e ervas daninhas de folhas largas e gramíneas. Com base nas imagens, uma Rede Neural Convolucional foi instruída para detectar as ervas daninhas e os resultados foram comparados aos Algoritmos de Máquina de Vetor de Suporte, das quais, as Redes Neurais Convolucionais foram capazes de alcançar uma precisão de mais de 98% na detecção das ervas daninhas de folhas largas e gramíneas.

Foi retratado por Name [7], um método computacional baseado em *software* livre para auxiliar na avaliação do crescimento radicular para amostras lavadas da cultura da soja. Tal método utiliza OpenCV para comparação de imagens, reduzindo o tempo para obtenção de atributos de raízes em 53% quando comparado ao método tradicional.

Métodos em visão computacional para a análise de folha de soja, com a finalidade de localizar e extrair características que possibilitem a detecção de doenças foliares, foram apresentados por Rocha e Sartin [8]. Esse trabalho analisa as folhas da soja pelo pré-processamento de imagens, utilizando-se de filtros de média, mediana e métodos de detecção de bordas e linhas para identificar a folha da soja, e faz uso de segmentação de imagens para extrair as mais importantes características foliares.

LeafDoctor [9] é um trabalho desenvolvido pela parceria da Universidade do Hawaii – Manoa, Universidade *Cornell e College of Tropical Agriculture and Human Resourcer* (UHCTAHR) que realiza avaliações quantitativas para doenças em folhas de plantas, a partir da submissão de fotografias de plantas doentes. O aplicativo utiliza um algoritmo para calcular a percentagem de tecido doente e classificá-la.

O projeto Método Computacional para Identificação do Fungo *Cercospora Kikuchii* em Sementes de Soja [10] faz uso de processamento de imagens para controle e identificação do

fungo *Cercospora kikuchii*, agente responsável pela mancha púrpura na folha da soja, capaz de provocar prejuízos, seja na produtividade ou na produção dos derivados. Foi utilizada a biblioteca de visão computacional *OpenCV* em conjunto com a linguagem Java e a ferramenta *Weka* para o desenvolvimento do sistema computacional. O resultado mostrou-se eficiente, visto que foram analisadas 150 sementes sadias e 150 sementes contaminadas, com um índice 86% de precisão.

PlantAI é um aplicativo desenvolvido [11] para a detecção e identificação de doenças em plantas que utiliza técnicas de *Deep Learning* (DL) para auxiliar na tomada de decisão do profissional, de forma a fornecer um diagnóstico e sugestão de tratamento, caso necessite.

O Agronomobot [12], *chatbot* para fins agrícolas foi desenvolvido para informar sobre as condições do campo, como temperatura do ar e do solo, umidade relativa do ar, umidade do solo, precipitação e velocidade do vento, para auxiliar na tomada de decisão sobre o gerenciamento da fazenda, de forma rápida e eficaz. A plataforma permite o uso de PLN durante a experiência de conversação.

O trabalho de Camargo et al. [13], utiliza-se de um ambiente virtual para disponibilizar um meio de estudo, de experimentação, análises e avaliações de situações da cultura da soja. No ambiente virtual, por meio de um formalismo matemático, uma gramática *L-Systems* e o desenvolvimento de sua representação gráfica no ambiente virtual, foi possível simular o comportamento da cultura da soja com alterações dos macronutrientes do solo, tendo como resultados um melhor acompanhamento da cultura da soja e manejo do solo.

Dentre os trabalhos da área de ontologias de domínio no campo da agricultura pode-se destacar o AgroPortal [14], que reutiliza as ferramentas e *insights* semânticos do domínio biomédico para atender a agronomia, ciências da alimentação e biodiversidade. O portal apresenta hospedagem, pesquisa, versionamento, visualização, comentário, recomendação de ontologias, além de permitir anotações semânticas. Seu repositório apresentou-se robusto e rico em recursos e com grande valor para o domínio agronômico.

O CGIAR (*Consultative Group for International Agricultural Research*) [15] sugeriu o desenvolvimento de um *software* de *big data* para a agricultura, com o uso da ontologia de domínio para a agricultura *Crop Ontology* (*Crop and Agronomy Ontology Community*) e o projeto AGROVOC [16] também utiliza um vocabulário voltado para agricultura, compartilha palavras e faz o reuso de ontologias.

Em [17] pode ser encontrado um trabalho que utiliza metadados e ontologias para a melhoria do processo de intercâmbio de dados para processos de produção da cultura da erva-mate (*Ilex Paraguariensis St. Hill.*). O uso das ontologias de domínio resultou em estruturas de informação para realizar o intercâmbio e rastreabilidade de dados agrícolas, aprimorando os processos e gestão das informações. Na fase de desenvolvimento da ontologia foi utilizada a Ferramenta *Protégé*, aplicando a metodologia *Ontology Development 101*.

O Edubot [18] é um projeto que possui um processo de aquisição de um modelo de domínio baseado em ontologias, modelado em lógica de descrição para a extração de conhecimento em linguagem natural, por meio da construção e representação de diálogos entre pessoas e máquinas, capaz de aprender e raciocinar sobre os fatos dialogados. Os conteúdos mostram a usabilidade do AIML (*Artificial Intelligence Markup Language*) e, assim, capturam fatos para representá-los nas ontologias em DL.

Os recursos computacionais aliados aos meios de comunicação oferecem a seus usuários cada vez mais funcionalidades e alternativas de usabilidade, que foram desenvolvidas para melhorar a interação humano-computador.

A presente pesquisa é relevante porque não há um *software* que utiliza Processamento de Linguagem Natural para auxiliar o profissional da área agronômica na tomada de decisão para identificação das pragas e doenças da soja, baseado nas descrições dos patógenos e suas principais características armazenados em um repositório e que tenha eficiência de consulta por meio do diálogo.

3. Fundamentação teórica

Com a expansão da tecnologia no mundo, as interfaces estão inseridas no dia a dia entre os seres humanos e os sistemas informatizados. O Processamento de Linguagem Natural tem o intuito de desempenhar um papel fundamental para a comunicação com os usuários, de maneira que esses se sintam mais confortáveis ao fazerem suas consultas em Banco de Dados com sua própria língua de comunicação.

A Linguagem Natural (LN) é uma linguagem de comando nas quais as regras sintáticas são as da linguagem natural utilizada pelo usuário e, portanto, dispensa a aprendizagem de sintaxes muito específicas e inflexíveis, como as linguagens de programação tradicionais [19].

Em virtude da soja ser um produto agrícola em destaque mundial, os cuidados com a cultura para evitar a manifestação de pragas e doenças devem ser feitos corretamente para evitar prejuízos econômicos. O controle é feito a partir da identificação dos vilões na planta e para identificá-los em um sistema informatizado é necessário o uso de técnicas de PLN que armazenam suas características.

As ferramentas de PLN baseiam-se na aplicação de técnicas para analisar dados textuais e tem como objetivo extrair representações e significados mais completos de textos em linguagem natural, utilizando técnicas e padrões linguísticos para sistemas que necessitam tratar de modo mais amigável a entrada de dados do usuário [20].

Para Miura [21] é imprescindível o desenvolvimento de aplicações automatizadas para analisar e interpretar textos em linguagem natural para que ações sejam executadas como resposta, fornecendo informações.

Nesse cenário tem-se o exemplo de ferramentas que são projetadas para analisar consultas e recuperar informações, de maneira rápida e natural, sem que haja necessidade de conhecer a estrutura interna de implementação, como os assistentes virtuais.

Conforme o modelo acima, o receptor compreende uma mensagem, processa a relação entre as palavras que a compõe e as classifica quanto ao seu significado. Ou seja, para o sistema computacional analisar a sentença, necessita-se identificar os problemas enfrentados no PLN: a ambiguidade da mensagem, a dependência do contexto para ter a interpretação correta, a incompletude do conhecimento e a evolução do conhecimento [22].

Visto a complexidade do assunto, o PLN tem o intuito de desempenhar um papel fundamental para a comunicação com os usuários, de maneira que esses se sintam mais confortáveis ao fazerem suas consultas em banco de dados com sua própria língua de comunicação.

A Figura 1 apresenta as cinco fases utilizadas na análise de Dale [23] que é dividida em vários segmentos, baseado nos traços linguísticos. A presente seção irá conceituar as técnicas de PLN para que o sistema possa compreender e interpretar os dados gerados.

Figura 1. Estágios da Análise no PLN

Tokenização: com a tokenização foi possível estudar das palavras (isoladamente) segundo sua classificação. Essa faz o reconhecimento das letras maiúsculas e minúsculas, palavras compostas ou abreviadas e quebra de caracteres;

Análise Léxico-morfológica: essa etapa é responsável por fazer a verificação ortográfica e classificação léxico-morfológica, podendo ser classificada como: substantivo, verbo, advérbio, pronome, numeral, preposição, conjunção, interjeição, artigo e adjetivo; segundo sua estrutura e formação, identificando as partes: radical, tema, vogal temática, dentre outras [24].

Análise sintática: conjunto de tarefas que definem a função sintática do *token* na frase. A tarefa da Análise Sintática (*parsing*) é de extrair informações de uma frase representada por meio de uma gramática e árvores sintáticas.

Para Domingues [24], a finalidade do *parsing* é analisar e gerar sentenças corretas de acordo com a estrutura de cada palavra. O sistema de perguntas e respostas objetivam analisar uma pergunta formulada em linguagem humana e determinar sua resposta. Comumente atuam em um domínio restrito, de maneira que os sistemas de PLN podem explorar o conteúdo do domínio específico na construção de suas bases de conhecimento.

Análise Semântica: consiste em analisar os significados das palavras, ou seja, interpretar as expressões fixadas, sentenças inteiras e enunciados no contexto [25], pois as frases podem ser ambíguas.

Análise Pragmática: o estudo fundamenta-se em reconhecimento de palavras dentro de um contexto [26]. A estrutura não parte de apenas uma frase, visto que busca nas frases do texto para compreender a frase analisada. Chomsky [27] exemplifica essa fase com o exemplo de fala em um diálogo *"Já é muito tarde?"*, podendo o autor da frase se referir ao tempo/às horas do dia ou à ausência de pontualidade. Um exemplo do domínio deste trabalho temos: *"A mancha-púrpura implica a planta. Essa apodreceu."*

A aplicação desse conhecimento em PLN é indispensável, principalmente quando o destaque do processamento interpreta as informações do usuário, permitindo o acesso à base de dados e os são textos esquemáticos, com uma parte variável que só é determinada no decorrer do discurso (ex.: essa, isso, aquela etc.).

A arquitetura de um sistema computacional [28] que executa a língua natural pode variar de acordo com as especificidades da aplicação. O tradutor automático é um possível exemplo em um sistema mais completo que deverá ser capaz de:

I - Identificar, ou seja, extrair cada uma das palavras da sentença;

II - Analisar sintaticamente a sentença, isso é, relacionar a cada palavra suas propriedades e funções sintáticas;

III - Construir uma nova sentença para retomar o sentido das informações levantadas anteriormente, ou seja, extrair um significado absoluto da mesma, a partir dos significados das palavras e das relações entre elas;

IV - Associar o significado extraído em uma representação adequada. Essa representação pode ser independente da língua destino (interlíngua);

V - Transformar a representação anterior em uma sentença na língua destino, isso é, traduzir ou associar as palavras da língua origem para a língua destino, desde que sejam equivalentes.

A partir disso, é possível determinar as inúmeras maneiras de escrever ou expressar um certo material informacional, fazendo com que o sistema controle automaticamente a geração da tarefa e que o processo dessa seja mais simples do que a de interpretação, como sistemas que têm a função específica de transmitir informações constantes em uma base de dados

em PLN, ou seja, de sistemas que a missão seja declarar ou informar, como é sugerido na Figura 2.

Ainda assim, visar a simplificação do sistema pode prejudicar o resultado, dado que o texto implica na modificação do conteúdo informacional e no domínio dos pontos comunicativos, conforme as intenções do autor do texto, como a tradução automática, que propõe uma correspondência mais verdadeira entre o texto de origem e o texto de destino.

4. Interface em Linguagem Natural para Banco de Dados

A interface consiste na interação entre o usuário e o sistema, ou seja, um sistema de comunicação.

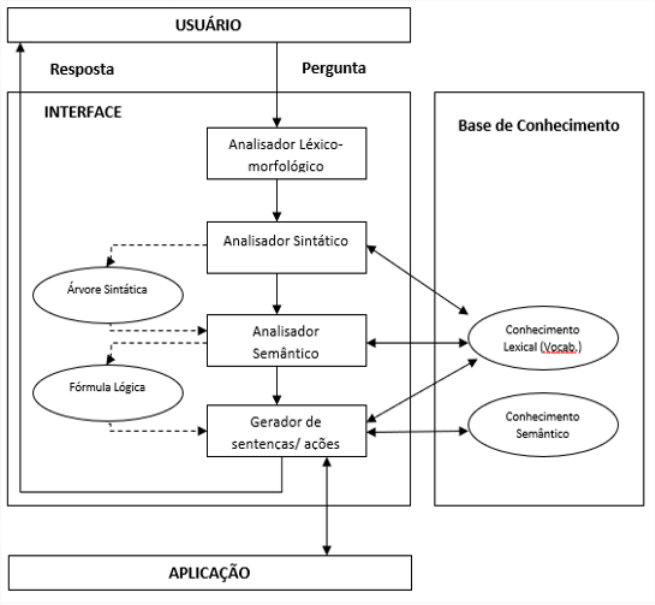

Figura 2. Arquitetura de uma Interface em Linguagem Natural para Banco de Dados

A interface de usuário deve ser entendida como sendo a parte de um sistema computacional com a qual uma pessoa entra em contato – física, perceptiva ou conceitualmente [29].

As características dessa interação e interface avaliam a qualidade de um sistema [30] e são chamadas de critérios de qualidade pela:

• **Usabilidade:** facilidade de aprendizado do uso do sistema devido à interface;

• **Experiência do usuário:** medida de satisfação do usuário em relação ao sistema;

• **Acessibilidade:** capacidade dos usuários interagirem com o sistema;

• **Comunicabilidade:** capacidade de comunicação entre o sistema e o usuário por meio da interface.

De acordo com Jain, Lund e Wixon [31], interfaces naturais possuem uma interação semelhante ao contato do usuário com o mundo, ou seja, não percebe a tecnologia no momento da utilização.

O objetivo da interface natural baseada na experiência que o usuário possui ao utilizá-la depende de uma interface clara para o usuário entender como manusear, autêntica e com indicações visuais para facilitar o acesso.

A preocupação dessa é aproximar cada vez mais o comportamento das ações humanas em LN no idioma português (no caso em questão) para que os usuários apenas pensem nas ações que querem expressar, sem investir tempo para aprender uma linguagem de programação para consulta em base de dados.

As vantagens de uma Interface em Linguagem Natural para Banco de Dados (ILNBD) são que o usuário não precisa compreender nenhuma linguagem artificial de comunicação, ou seja, as consultas são formuladas na linguagem nativa do usuário. Alguns tipos de perguntas podem ser facilmente expressos em linguagem natural e podem ser interpretadas pelo contexto, e tornam-se difíceis ou não são suportadas quando são usadas interfaces gráficas ou baseadas em formulários [32]

Ainda em Barbosa [32] são apresentados alguns problemas linguísticos na construção de ILNBDs, como:

• **Ligação de modificadores:** constituintes de uma frase modificam o significado de outros constituintes sintáticos. Para o sistema escolher a leitura correta deve saber o domínio da aplicação e apresentá-la ou imprimir respostas correspondentes à ambas leituras e iniciar quais respostas se referem a quais leituras. Em alguns casos a ligação de modificadores é ambígua. Por exemplo em consultas que envolvem Ligação de Modificadores, como no caso: *As pragas atacaram rapidamente.*

• **Escopo de quantificadores:** problema enfrentado quando as frases são convertidas para declarações lógicas (um, cada, todo, algum), como no seguinte exemplo: *Algumas pragas prejudicam a cultura da soja.*

• **Conjunção e disjunção:** a palavra "*e*" pode denotar disjunção ao invés de conjunção. Exemplo: *A Broca-da-Vagem provoca danos antes e depois da colheita.*

• **Nominais compostos:** são "substantivo-substantivo" e "adjetivo-substantivo". Esses nominais dificultam a determinação de seus significados, por exemplo: *O professor agricultor diagnosticou a doença.*

• **Anáforas:** são encontradas em frases que denotam implicitamente entidades mencionadas no discurso, como em: *A praga foi considerada uma ameaça quando ela não foi totalmente erradicada.*

• **Sentenças elípticas:** são sentenças incompletas usadas em discurso. O tratamento de elipses é a capacidade de entender o contexto da frase facilitando a interação do usuário com o sistema. Exemplo: *Conferiu as pragas na lavoura.*

• **Expressões extragramaticais:** a ILNBD deve auxiliar o usuário mesmo se as perguntas forem mal formadas, como no caso de erros de digitação.

Os problemas linguísticos abordados acima não surgem somente em ILNBDs.

É necessário estabelecer um conjunto de regras que reco-

nhece os vários tipos de diagnósticos. Por meio da gramática elaborada, foi viável trabalhar o domínio proposto [19] para os tipos de construção:

a) Sentença;

b) Sintagma nominal: trecho da oração que define completamente uma entidade ou conjunto de entidades do mundo do falante [33];

c) Sintagma verbal: determina uma atividade ou estado no tempo, como os verbos. O verbo é o "centro" do sintagma verbal. Caso o verbo seja intransitivo, forma um sintagma verbal. Se o verbo for transitivo, deve haver um sintagma nominal que é o objeto da atividade, para completar o sintagma verbal;

d) Sintagma adjetival: para Savadovsky [33], um verbo de ligação atribui qualidades a um sintagma nominal ou qualifica um verbo intransitivo ou sintagma nominal. Inclui um adjetivo, que pode ser seguido de um sintagma preposicional, precedido de advérbios;

e) Sintagma preposicional: formado por uma preposição seguida de um sintagma nominal;

f) Sintagma adverbial: composto de um ou mais advérbios seguidos ocasionalmente por uma preposição. "Função" como a do advérbio e nas situações mencionadas [33];

g) Oração subordinada adjetiva restritiva: são as que limitam a extensão do nome a que se referem, ou seja, determinam um subconjunto dentro de um conjunto. Este tipo de oração inicia-se por um pronome relativo como que, quem, o(a) qual, os(as) quais etc.;

h) Sentenças Sim/Não: são aquelas que procuram o valor verdade de uma fórmula (*True*/*False*). Por exemplo em *"A folha está manchada?"*;

i) Sentenças -*wh:* são as que procuram valor de instanciação de funções temáticas. Exemplo: *"A praga atacou a semente!"*;

j) Sentenças alternativas na forma normal (não- clivada): perguntas sobre alternativas exclusivas no predicado como em *"Apresenta coloração amarela ou esverdeada?"* [34];

k) Sentenças de solicitação de explicação: *"Por que?"*;

l) *S* existencial: *"Alguma região foi examinada?"*;

m) *S* na voz ativa: *"O ataque compromete alguma região?"*;

n) *S* na voz passiva: *"A região é comprometida pelo ataque?"*;

o) *S* clivada: extraposição do verbo ser, por exemplo, em *"É a análise que confirmou o diagnóstico?"*.

A gramática é a representação da linguagem verbal interpretável/gerável. Contém regras de estruturação sintática e de morfossintaxe (gênero e número) [19]. Contém valores possíveis das categorias sintáticas de nível mais baixo, como de complementos "tempo", "direção", "lugar" e outros [34].

5. Arquitetura de um Sistema em Linguagem Natural para Banco de Dados

A arquitetura de um sistema computacional [35] que executa a língua natural pode variar de acordo com as especificidades da aplicação. O tradutor automático é um possível exemplo em um sistema mais completo que deverá ser capaz de:

a) identificar, ou seja, extrair cada uma das palavras da sentença;

b) analisar sintaticamente a sentença, isto é, relacionar a cada palavra suas propriedades e funções sintáticas;

c) construir uma nova sentença para retomar o sentido das informações levantadas anteriormente, ou seja, extrair um significado absoluto da mesma, a partir dos significados das palavras e das relações entre elas;

d) associar o significado extraído em uma representação adequada. Essa representação pode ser independente da língua destino (interlíngua);

e) transformar a representação anterior em uma sentença na língua destino, isto é, traduzir ou associar as palavras da língua origem para a língua destino, desde que sejam equivalentes.

A partir disso, é possível determinar as inúmeras maneiras de escrever ou expressar certo material informacional, fazendo com que o sistema controle automaticamente a geração da tarefa e que o processo dessa seja mais simples do que a de interpretação, como sistemas que têm a função específica de transmitir informações constantes em uma base de dados em PLN.

A Compreensão da Língua Natural (CLN) e Geração de Língua Natural (GLN) são subcampos da inteligência artificial e linguística computacional [36] e atuam na interação entre humanos e computadores, não apenas no processamento da linguagem e suas etapas para análise do discurso, como a geração efetiva de linguagem natural por um computador [37], sob a forma e estrutura da linguagem natural humana.

A arquitetura geral de um sistema de interpretação, baseada em Oliveira [35], é exibida na Figura 3, a qual está sendo pautada este artigo.

Os indicadores de processamento são representados por retângulos, enquanto os recursos necessários ao processamento, de ordem linguística (gramática, léxico) ou não (modelos do domínio e do usuário), aparecem representados por círculos. Esses recursos são essenciais durante a fase de interpretação, bem como na de geração. Destaca-se as aplica ções que não envolvem a interpretação de uma sentença que têm sua arquitetura simplificada, eliminando-se alguns dos módulos e/ou bases de conhecimento que aparecem na Figura 3.

Os recursos linguísticos presentes na arquitetura de interpretação e geração de linguagem natural, baseada em Silva et al. [35], são expostos a seguir.

Léxico: conjunto de palavras ou expressões da língua associadas a atributos, ou traços morfossintáticos e traços semânticos (opcionais). Durante a interpretação, o léxico é

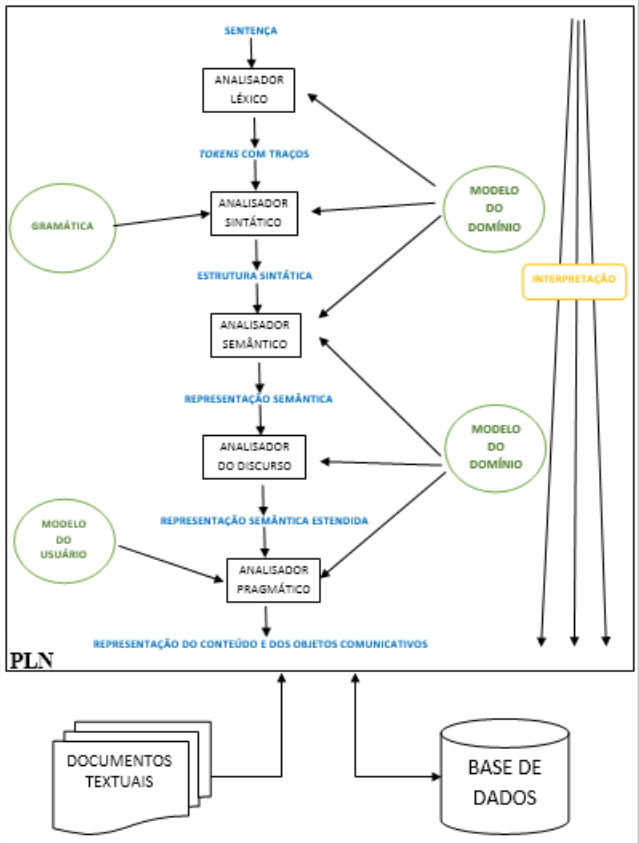

Figura 3. Arquitetura de um Sistema de Interpretação de Linguagem Natural

acessado pelos analisadores léxico, sintático e semântico, cada um deles visam funções específicas, sendo suas principais tarefas: reconhecer *tokens* da sentença de entrada e recuperar seus principais traços; reconhecer ou atribuir categorias sintáticas a esses, para a obtenção da estrutura profunda da sentença e verificar a validade do relacionamento semântico da *token* sob análise em função do contexto em que ela ocorre na sentença em relação aos demais durante a análise dos componentes sentenciais.

Gramática: conjunto de regras gramaticais que definem as cadeias de palavras válidas em uma sentença em linguagem natural.

Modelo de Domínio: fornece informações sobre o domínio específico da aplicação.

Modelo de Usuário: permite reconhecer características do significado textual a partir do contexto do discurso.

Após isso, é possível estabelecer os inúmeros modos de escrever, fazendo com que o sistema manuseie automaticamente a geração da tarefa e que o processo dessa transmita informações continuamente em uma base de dados em PLN.

6. Metodologia e Desenvolvimento do Sistema da Identificação das Pragas da Soja

Esta seção trata da proposta sugerida para este trabalho, que tem como objetivo principal criar um software para contribuir nas tomadas de decisões dos profissionais da área, como uma ferramenta de diálogo em LN entre o sistema e o usuário para otimizar as atividades guiadas à agricultura.

As etapas para o desenvolvimento do Sistema da Soja são baseadas em Sampaio [38], compostas por:

• definições das funcionalidades do sistema para permitir que o usuário controle as funções desse;

• levantamentos das informações e técnicas necessárias para a construção das partes mais importantes dos sistemas para determinar a melhor maneira de realizar uma tarefa;

• definições das ferramentas necessárias para a construção do sistema para o desenvolvimento de aprendizado de máquinas. Lembrando que o sistema precisa lidar com informações e problemas do mundo real;

• desenvolvimento do sistema seguindo uma metodologia de desenvolvimento de software para nortear o desenvolvimento de sistemas;

• análises dos resultados obtidas com os testes para chegar às conclusões e verificar se o objetivo do projeto foi atingindo.

A elaboração deste trabalho permite contribuir com a área agrícola, principalmente no que diz respeito à viabilidade de um assistente virtual, utilizando técnicas de PLN na identificação de pragas e doenças nas plantas da soja.

A partir das classificações das características das pragas e doenças da soja, um modelo foi construído por meio da biblioteca spaCy [39], utilizando técnicas de PLN para aplicar regras gramaticais às sentenças, reconhecendo suas estruturas e extraindo seus significados, conhecida por fazer o pré-processamento, envolvendo:

• *Tokenization* (tokenização): subdivide a base de dados em *tokens*, como é vista na Figura 4 para a pergunta *"As folhas atacadas ficam com grandes áreas recortadas ou são completamente consumidas?"*.

```
1 [token.orth_ for token in doc]
```

```
['As',
 'folhas',
 'atacadas',
 'ficam',
 'com',
 'grandes',
 'áreas',
 'recortadas',
 'ou',
 'são',
 'completamente',
 'consumidas',
 '?']
```

Figura 4. Tokenização

• *Lemmatization and Stemming* (Lematização e Stemização): a lematização é o processo de redução de palavras à sua base (raiz), enquanto a stemização permite checar se uma palavra é raiz de outra. São etapas importantes para análise de grandes volumes de textos, como é exemplificada na Figura 5 a seguir, para a seguinte frase: *"As folhas atacadas ficam com grandes áreas recortadas ou são completamente consumidas?"* .

```
for token in documento:
  print(token.text, token.lemma_, stemmer.stem(token.text))
```

```
As As as
folhas folhar folh
atacadas atacar atac
ficam ficar fic
com com com
grandes grande grand
áreas área áre
recortadas recortar recort
ou ou ou
são ser são
completamente completamente complet
consumidas consumir consum
? ? ?
```

Figura 5. Lematização e Stemização

• POS *tagging (part-of-speech tagging)*: classifica corretamente todas as palavras de cada sentença do texto por categorias gramaticais, como é mostrada na Figura 6.

```
for token in documento:
  print(token.text, token.pos_)
```

```
As DET
folhas NOUN
atacadas VERB
ficam VERB
com ADP
grandes ADJ
áreas NOUN
recortadas VERB
ou CCONJ
são AUX
completamente ADV
consumidas ADJ
? PUNCT
```

Figura 6. POS *tagging*

• Análise Léxico-Morfológica: responsável por manipular o léxico, que é composto por palavras que armazena seu significado. Identificar as partes da frase é essencial porque ajuda a entender as frases de entrada e constrói com mais exatidão as frases de saída (resposta). Assim, na Figura 7 aplica-se o *token* nos textos e extrai as etiquetas morfológicas para classificação de cada palavra, a partir da seguinte frase: *"As folhas atacadas ficam com grandes áreas recortadas ou são completamente consumidas?"*

```
[(token.orth_, token.pos_) for token in doc]
```

```
[('As', 'DET'),
 ('folhas', 'NOUN'),
 ('atacadas', 'VERB'),
 ('ficam', 'VERB'),
 ('com', 'ADP'),
 ('grandes', 'ADJ'),
 ('áreas', 'NOUN'),
 ('recortadas', 'VERB'),
 ('ou', 'CCONJ'),
 ('são', 'VERB'),
 ('completamente', 'ADV'),
 ('consumidas', 'VERB'),
 ('?', 'PUNCT')]
```

Figura 7. Análise Léxico-morfológica

• Análise sintática: responsável por organizar o conjunto das palavras e então, aplica-se regras gramaticais à sentença para reconhecer a estrutura e extrair seus significados;

Posteriormente, essa ferramenta rotula todos os *tokens* a partir da análise para prever qual tag provavelmente se aplica nesse contexto. Para isso, conta-se com as seguintes tarefas: **Text** (texto puro), **Lemma** (reduz as palavras em seu formato base/raiz), **POS** (*tags* simples que determinam as categorias gramaticais de um *token*), **Dep** (Dependência sintática é a relação entre os tokens presentes dentro de uma sentença para entender o seu significado), **Shape** (classificação da palavra em maiúscula ou minúscula), **Alpha** (especificação das palavras em alfanuméricas ou não), **Stop** (indica se as palavras são consideradas *stopwords*), como é detalhada na Figura 8 para a frase

"As folhas atacadas ficam com grandes áreas recortadas ou são completamente consumidas?" .

```
for token in doc:
    print(token.text, token.lemma_, token.pos_, token.tag_, token.dep_,
          token.shape_, token.is_alpha, token.is_stop)
```

```
As As DET <artd>|ART|F|P|@>N det Xx True True
folhas folhar NOUN <np-def>|N|F|P|@SUBJ> nsubj xxxx True False
atacadas atacar VERB <mv>|V|PCP|F|P|@ICL-N< acl xxxx True False
ficam ficar VERB <mv>|V|PR|3P|IND|@FS-STA ROOT xxxx True False
com com ADP PRP|@<ADVL case xxx True True
grandes grande ADJ ADJ|F|P|@>N amod xxxx True True
áreas área NOUN <np-idf>|N|F|P|@P< obl xxxx True False
recortadas recortar VERB <mv>|V|PCP|F|P|@ICL-N< acl xxxx True False
ou ou CCONJ <co-fcl>|<co-fmc>|<co-vfin>|KC|@CO cc xx True True
são ser VERB <cjt>|<mv>|V|PR|3P|IND|@FS-STA aux:pass xxx True True
completamente completamente ADV ADV|@ADVL> advmod xxxx True False
consumidas consumir VERB <pass>|<mv>|V|PCP|F|P|@ICL-AUX< conj xxxx True False
. . PUNCT PU|@PU punct . False False
```

Figura 8. Análise sintática

O spaCy [39] possibilita descrever a relação sintática das palavras que se conectam na formação da árvore. Isso permite percorrer toda a árvore e retornar uma sequência ordenada de *tokens* e verificar os atributos e domínios das palavras. Nessa fase, além do que foi descrito anteriormente, conta-se com o

Head Text (relação entre as palavras nos tokens), **Head Pos** (rotula as palavras em categorias) e **Children** (dependentes sintáticos do token) e são apresentados na Figura 9 para a frase *"As folhas atacadas ficam com grandes áreas recortadas ou são completamente consumidas?"* .

```
for token in doc:
    print(token.text, token.dep_, token.head.text, token.head.pos_,
          [child for child in token.children])
```

```
As det folhas NOUN []
folhas nsubj ficam VERB [As, atacadas]
atacadas acl folhas NOUN []
ficam ROOT ficam VERB [folhas, áreas, consumidas, .]
com case áreas NOUN []
grandes amod áreas NOUN []
áreas obl ficam VERB [com, grandes, recortadas]
recortadas acl áreas NOUN []
ou cc consumidas VERB []
são aux:pass consumidas VERB []
completamente advmod consumidas VERB []
consumidas conj ficam VERB [ou, são, completamente]
. punct ficam VERB []
```

Figura 9. Navegando pela análise sintática

• Reconhecimento de Entidades Nomeadas (*Named-entity recognition - NER*): Essa ferramenta também permite fazer o Reconhecimento de Entidade Nomeada, como é conhecida em português, e tem a função de identificar e classificar palavras ou frases em um texto de acordo com classes definidas para o modelo [40].

Conforme abordado em Speck e Ngomo [41], as principais atividades feitas pelo NER são identificar os *tokens* em um texto não estruturado e classificá-los em tipos de entidades definidas de acordo com a peculiaridade do domínio. Essa tarefa permite capturar termos nas bases de dados textuais, identificá-la e classificá-la, conforme mostra a Figura 10 para a frase:

"No Brasil, há registros da ocorrência de Percevejo-castanho em várias regiões, embora os danos dessa praga tenham sido mais frequentes nos estados de Mato Grosso, Goiás e Mato Grosso do Sul."

```
for entidade in documento.ents:
  print(entidade.text, entidade.label_)
```

```
Brasil LOC
Percevejo-castanho PER
Mato Grosso LOC
Goiás LOC
Mato Grosso do Sul LOC
```

Figura 10. Reconhecimento de Entidades Nomeadas

Ao utilizar a função displaCy por meio da ferramenta spaCy, é possível destacar visualmente as entidades e os tipos, como é apresentada na Figura 11 abaixo.

• *Parsing* de dependências: depois de classificar corretamente todas as palavras de cada sentença do texto por categorias gramaticais, é feita a relação de dependência entre palavras, como exibe a Figura 12.

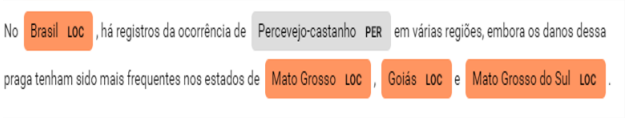

Figura 11. Visualização do NER por meio do displaCy

"O Coró-da-soja é uma praga que ataca a raiz da soja."

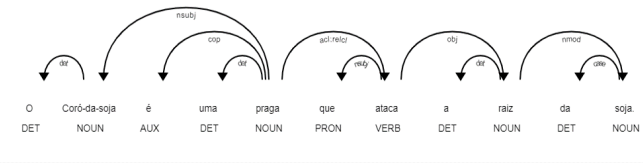

Figura 12. *Parsing* de dependência

7. Banco de Dados Não Relacional para Identificação de Pragas e Doenças na Cultura da Soja

Para o presente trabalho optou-se por um Banco de Dados Não Relacional, também conhecido como Banco de Dados NoSQL (*Not Only Structured Query Language*) do tipo grafo, o Neo4j, que possui uma interface robusta e flexível perante outros Banco de Dados (BDs)da mesma categoria. Além disso, permite consultas com um alto nível de abstração e segundo Fernandes [42], os dados são armazenados em uma estrutura de nós (entidades), arestas (relacionamento entre os nós) e propriedades (atributos das entidades), podendo essas serem representadas tanto nos nós quanto nas arestas do grafo.

Os conceitos do modelo de dados do Neo4j [43] são descritos abaixo:

Vértices: consiste no elemento principal do modelo de dados do Neo4j. Eles podem ser conectados por meio de relacionamentos, ter uma ou mais propriedades (atributos guardados como um par chave-valor) e ter um ou múltiplos rótulos (identificador do tipo do vértice);

Relacionamentos: conectam dois vértices e são orientados, ou seja, possuem direção. Eles podem apresentar uma ou mais propriedades e dispõem um tipo, que tem a mesma função dos rótulos dos vértices. Porém, no caso de relacionamentos são limitados a um único tipo;

Propriedades: são valores nomeados, em outras palavras, pares chave-valor. A chave sempre é uma String e seu valor pode ser um número (*Integer* ou *Float*), uma *String*, um *Boolean*, um tipo *Spacial Point*, uma gama de tipos temporais (*Date, Time, LocalTime, DateTime, LocalDateTime* e *Duration*). Listas e mapas de tipos simples também são permitidos.

Ainda, o Neo4j utiliza um armazenamento nativo em grafo. É também chamado de *index-free adjacency*, onde

cada vértice aponta fisicamente para sua localização na memória, melhorando o desempenho de acesso [43] A linguagem de consulta do Neo4j é a Cypher [44], a qual é inspirada na linguagem SQL e adota o conceito de combinação de padrões da linguagem SPARQL. Cypher descreve os vértices, os relacionamentos e as propriedades como se formassem um desenho utilizando caracteres ASCII, tornando, assim, as consultas mais fáceis de ler e entender [45]. Como é mostrada na Figura 13, tem-se a estruturação dos dados armazenados em nós dos Ácaros Fitófagos e associação às informações por meio das arestas do Ácaro-Vermelho, Ácaro-Branco e Ácaro-Verde.

Figura 13. Neo4j

O objetivo é criar uma interface computacional em linguagem natural para garantir um bom desempenho das tarefas na área da agricultura, como uma ferramenta de diálogo entre o sistema e o usuário que possibilita um fácil acesso às informações em um repositório da base de dados e visa aumentar o repositório conforme ocorre o diálogo, e então, os dados crescem exponencialmente. Assim, o banco de dados armazena não só os dados, mas as suas relações de maneira eficiente.

Os principais aspectos que identificam as ameaças da sojicultura são mostradas na Tabela 1.

8. Sistema Conversacional para Identificação de Pragas na Cultura da Soja

A ferramenta CAROLINA (**Co**nversação **A**gronômica **RO**botizada em **LI**nguagem **NA**tural) é uma plataforma desenvolvida para alunos/professores de cursos voltados à agricultura e/ou produtores rurais. Essa pretende proporcionar dados sobre as pragas e doenças na cultura da soja, que são armazenados em um repositório de banco de dados para facilitar o acesso à informações e o diagnóstico do profissional.

Perante aos problemas levantados, a elaboração deste trabalho permite contribuir com a área agrícola, principalmente no que diz respeito à viabilidade de um assistente virtual, utilizando técnicas de PLN na identificação de pragas e doenças nas plantas da soja.

Tabela 1. Classificação e características para identificação das pragas

Praga	Nome comum e nome científico
Descrição	Descreve a morfologia e ciclo da praga
Biologia	Características biológicas
Danos	Como se comportam e os danos causados
Ocorrência na planta	Localização do ataque da planta
Distribuição geográfica	Localização geográfica de onde essas pragas são encontradas
Controle	Métodos de controle
Categoria	As pragas foram categorizadas em: pragas que atacam plântulas / pragas que atacam raízes / pragas que atacam caules / pragas que atacam folhas / pragas que atacam vagens / pragas subterrâneas / pragas da parte aérea / pragas de solo / pragas mastigadoras /pragas sugadoras / outros

O repertório das palavras registradas no dicionário conta com 108 pragas da cultivar da soja, selecionadas por manuais e indicações bibliográficas, dentre elas, Santos [46], Moreira e Aragão [47], Sosa-Gomez et al. [48], Ávila [49] e 19 doenças de Michalski [50] que estão disponíveis em *UCI Machine Repository.*

O processo descrito permite a extração das informações que são efetuadas com a finalidade de identificar a qual rótulo específico uma determinada praga ou relacionamento está atrelado. Posto que este projeto tem o objetivo de colaborar com o profissional da área da agricultura, foi possível realizar perguntas para realizar o diálogo e possibilitar consultas, como:

"Tem um verme provocando uma lesão na raiz da soja. E agora?";

"A minha semente está com uma mancha-púrpura? O que é?";

"A soja armazenada está infestada de insetos";

"O broto não está se desenvolvendo.";

"Tem uma praga esverdeada na plantação.";

"Como posso controlar os nematoides?";

"Tem uma praga atacando a haste da soja.";

"O que é um bichinho dourado na soja?";

"O que posso fazer para controlar a Antracnose?";

"Os percevejos atacam que parte da planta?";

"Qual o dano causado pela Formiga-cortadeira?";

"Os ácaros atacam as folhas da planta?";

"Como previnir a presença de pragas na minha lavoura?";

"Como controlar a Antracnose?";

"A haste da soja está apodrecendo.".

As perguntas vão surgindo conforme o diálogo entre o usuário e o sistema acontece (pergunta-resposta) e espera-se que o utilizador obtenha informações relevantes para tomada

de decisão. As palavras dessas foram catalogadas separadamente de acordo com a sua categoria morfológica para formar o dicionário utilizado na análise léxico-morfológica das palavras.

As ferramentas de ILNBD permitem ao usuário o acesso às informações armazenadas na base dados. A partir disso, está sendo desenvolvido um sistema para dialogar com agricultores sobre as principais características das principais pragas e doenças na cultura da soja. Esse sistema recebe o texto e analisa as palavras de uma sentença isoladamente, como são visualizadas na Figura 14 e Figura 15.

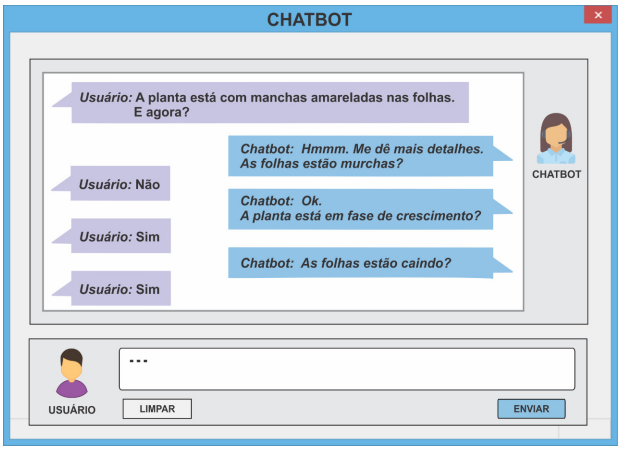

Figura 14. Diálogo da ferramenta do sistema conversacional CAROLINA

A partir do diálogo, o sistema deve fornecer o diagnóstico correto, como é mostrada na Figura 15.

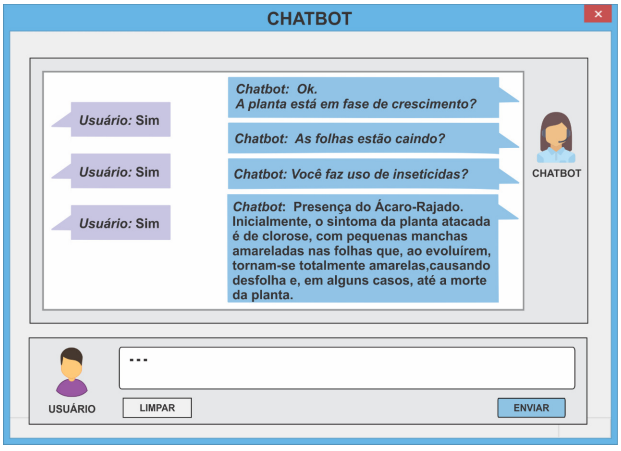

Figura 15. Diagnóstico da ferramenta do sistema conversacional CAROLINA

Ainda em Barbosa [19], são abordadas as estruturas básicas das famílias de árvores utilizadas, as quais são utilizadas no domínio da soja, como em:

- Verbos intransitivos: é selecionada por verbos que não precisam de complemento, possui sentido completo, como em: *A planta apodreceu.*

- Verbos transitivos diretos: é selecionada por verbos que pedem um complemento direto NP (têm estruturas adjetivais e preposicionais sofisticadas), como em: *A mancha-púrpura implica a planta.*

- Verbos transitivos indiretos: é selecionada por verbos que pedem um complemento regido por preposição, como em: *O agricultor se preocupa com a lavoura.*

- Verbos bitransitivos: é selecionada por verbos transitivo direto e indireto seguidos por um sintagma nominal e por um sintagma preposicional, como em: *O agrônomo recomendou pesticidas a todas as lavouras da região.*

- Verbos copulativos para o tratamento de predicativo do tipo adjetival: contém verbo copulativo seguido por sintagma adjetival, como o exemplo: *A haste está completamente nociva.*

- Verbos copulativos para o tratamento de predicativo do tipo preposicional: contém verbo copulativo seguido de sintagma preposicional, como em: *A raiz está comprometida.*

- Verbos copulativos para o tratamento de predicativo do tipo nominal substantivo: contém verbo copulativo seguido de sintagma nominal, como em: *A análise indica nematoides.*

- Locuções verbais com gerúndio: essa família é selecionada pelo gerúndio combinado com o verbo estar, no tratamento de locução verbal. Por exemplo: *A raiz prejudicada está comprometendo toda a planta.*

- Complemento sentencial: árvores passivas (contém preposição por) são manipuladas por ter árvores separadas dentro da família de árvores, por exemplo: *A lavoura está comprometida pelo ataque de pragas.*

9. Aplicação de métricas de classificação para o *software*

A principal dificuldade no desenvolvimento é dialogar com o usuário e reconhecer as suas intenções a partir de uma frase e respondê-lo automaticamente. Visto isso, o sistema inteligente de pré-atendimento aos agricultores, engenheiros agrônomos, especialistas e alunos tem o intuito de ser um canal alternativo de comunicação, para facilitar o acesso às informações e auxiliar no ensino para identificar o patógeno.

Para utilizar técnicas de processamento de linguagem natural para pré- processamento, foram aplicados algoritmos de aprendizado supervisionado de Aprendizado de Máquina à base de dados para a construção de classificadores que pudessem predizer a causa dos sintomas das plantas da soja a partir dos resultados (bom desempenho para auxiliar na tomada de decisão), com o uso de técnicas de PLN para otimizar as atividades guiadas à agricultura.

O algoritmo de aprendizado supervisionado é responsável por receber dados e encontrar uma função para ajustar o modelo de forma iterativa e que o mesmo se adapte às condições

apresentadas no conjunto de dados de treino com o objetivo de prever rótulos desconhecidos.

Assim, para treinar as métricas, foram escolhidos os algoritmos de classificação: *Random Forest* (RF) [51], Vetores de Suporte de Máquina (*Support Vector Machine*) [52] e K-vizinhos mais próximos (*K-Nearest Neighbors*) [53] para comparar e e então predizer quais são as doenças da soja, a partir das características descritivas. Os testes foram realizados uzando a Linguagem de Programação Python com a biblioteca Sklearn [1]. Também foi utilizada a técnica de *cross validation* com 10 *folds*, para evitar que a ordem dos dados na base pudessem induzir o modelo de aprendizado.

Os resultados do treino de classificação são apresentados na Figura 16, composta pela média das métricas de precisão, revocação e f1-*score*, as quais indicam que o modelo obteve uma performance alta.

Algoritmos	Precision	Recall	F1-score	Accuracy
KNN	0.87	0.92	0.88	0.96
Random Forest	0.97	0.93	0.95	0.96
SVM	0.94	0.94	0.94	0.97

Algoritmos	Precision	Recall	F1-score	Accuracy
SVM - Kernel	1.00	0.98	0.99	0.99

Figura 16. Comparação entre os algoritmos

Os dados incluem 35 atributos de 19 doenças da soja, como pode ser visto na Tabela 2.

Tabela 2. Classificação e características para identificação das pragas

Classificação	Características
Doenças	Fungo no cancro do caule / podridão escura na planta / fungo que causa podridão radicular / podridão da haste / podridão da semente / oídio / míldio / mancha marrom / praga bacteriana / folha enferrujada / mancha roxa na semente / infecção das plantas / mancha marrom na folha / queima das folhas / manchas circulares nas folhas / fungo na vagem e praga no caule / nematoide / branqueamento das folhas / ferimentos causados por herbicidas
Atributos	Mês; Condição da planta; Precipitação; Temperatura; Granizo; Histórico de colheita; Área danificada; Gravidade; Severidade; Uso de herbicida; Germinação; Crescimento da planta; Folhas; Manchas foliares; Cancro do caule; Lesão no cancro; Lesão na frutificação; Micélio; Coloração; Esclerócio; Vagens de frutas; Manchas nas frutas; Semente; Crescimento de bolores; Tamanho da semente; Enrugamento; Raízes.

[1] urlhttps://scikit-learn.org/

10. Considerações finais

A proposta deste trabalho foi implementar uma Interface em Linguagem Natural para Banco de Dados para interagir com o usuário por meio de frases referentes à soja, uma vez que há um aumento de interesse em utilizar os sistemas de computador como auxílio na tomada de decisão do agricultor.

O objetivo desta ferramenta é concentrar as informações sobre as principais pragas e doenças na cultura da soja em um repositório para auxiliar na tomada de decisão do agricultor, a partir de consultas e para isso um modelo foi construído por meio da biblioteca spaCy para aplicar regras gramaticais às sentenças, reconhecendo suas estruturas e extraindo seus significados, o que permitiu treinar o modelo para refinar seus dados.

Pretende-se ainda descrever mais consultas na base de dados da agricultura, mas para isso faz-se necessário mais entrevistas com os agricultores para verificar se há outros dialetos nesse domínio para que sejam abarcadas no dicionário. É preciso validar se as regras gramaticais desse público são as mesmas da norma culta ou não para saber se há a necessidade de acrescentar regras à gramática.

Contribuição dos Autores

Todos os autores contribuíram igualmente para este trabalho.

Referências

[1] CHOWDHURY, G. G. Natural language processing. *Annual Review of Information Science and Technology*, Wiley, v. 37, n. 1, p. 51–89, 2005.

[2] PEREIRA, P. H. S. et al. *Análise de descritores de imagens na classificação de folhas de soja visando o diagnóstico de doenças*. *X Simpósio Nacional de Tecnologia em Agronegócio*, Presidente Prudente, v. 10, p. 89–100, 2018.

[3] JURAFSKY, D.; MARTIN, J. H. *Speech & language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, New Jersey, 2000.

[4] LACERDA, V. S. *Estimativa do Índice de Severidade de Ferrugem Asiática na Cultura da Soja por meio de Imagens Obtidas com Aeronave Remotamente Pilotada*. Dissertação (Mestrado) — Programa de Pós-Graduação em Ciência Aplicada da Universidade Estadual de Ponta Grossa, 2017.

[5] ROSA, R. P. *Método de classificação de pragas por meio de rede neural convolucional profunda*. Dissertação (Mestrado) — Programa de Pós-Graduação em Ciência Aplicada da Universidade Estadual de Ponta Grossa, 2018.

[6] FERREIRA, A. S. *Redes Neurais Convolucionais Profundas na Detecção de Plantas Daninhas em Lavoura de Soja*. Dissertação (Mestrado) — Universidade Federal do Mato Grosso do Sul, 2017.

[7] NAME, M. H. *Método Computacional para Avaliação do Crescimento Radicular da Cultura da Soja*. Dissertação (Mestrado) — Programa de Pós-Graduação em Ciência Aplicada da Universidade Estadual de Ponta Grossa, 2013.

[8] ROCHA, I. A. A.; SARTIN, M. A. Pré processamento e segmentação de imagens de folhas de soja com base na visão computacional. *Workshop de Tecnologias Emergentes em Computação 2.0*, v. 16, 2018.

[9] PETHYBRIDGE, S. J.; NELSON, S. C. Leaf doctor: A new portable application for quantifying plant disease severity. *Plant disease*, Am Phytopath Society, v. 99, n. 10, p. 1310–1316, 2015.

[10] FRANCO, J. R. *Método computacional para identificação do fungo cercospora kikuchii em sementes de soja*. Dissertação (Mestrado) — Programa de Pós-Graduação em Ciência Aplicada da Universidade Estadual de Ponta Grossa, 2017.

[11] BENTO, D. C. P. G. C. *Deteção e identificação de doenças em plantas utilizando Deep Learning*. Dissertação (Mestrado) — Programa de Pós-Graduação em Engenharia Informática do Instituto Superior de Engenharia de Porto, Portugal, 2019.

[12] MOSTACO, G. M. et al. Agronomobot: a smart answering chatbot applied to agricultural sensor networks. In: *14th international conference on precision agriculture*. Canada: International Society of Precision Agriculture, 2018. v. 24, p. 1–13.

[13] TEIXEIRA, D. S. et al. Manejosoja3d: Ambiente virtual para aprendizado de manejo da cultura da soja. In: *XXVIII Brazilian Symposium on Computers in Education*. Recife: VI Congresso Brasileiro de Informática na Educação, 2017. v. 28, n. 1, p. 776–786.

[14] JONQUET, C. et al. Agroportal: an ontology repository for agronomy, computers and electronics in agriculture. *IN PRESS, Elsevier*, v. 144, p. 126–143, 2017.

[15] KING, B.; WONG, K. *The 2017 CGIAR Inspire Challenge:* innovation strategies for digital agriculture. CGIAR Platform for Big Data in Agriculture, 2017.

[16] CARACCIOLO, C. et al. The agrovoc linked dataset. *Semantic Web*, IOS Press, v. 4, n. 3, p. 341–348, 2013.

[17] SILVA, J. A. *Ontologia na rastreabilidade de dados agrícolas*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2018.

[18] LIMA, C. E. T. *Um Chatterbot para criação e desenvolvimento de ontologias com lógica de descrição*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2017.

[19] BARBOSA, C. R. S. C. *Técnicas de parsing para gramática livre de contexto lexicalizada da língua Portuguesa*. Tese (Doutorado) — Instituto Tecnológico de Aeronáutica. São José dos Campos, 2004.

[20] RODRIGUES, E. L. F. *Geração de perguntas em linguagem natural a partir de bases de dados abertos e conectados: um estudo exploratório*. Dissertação (Mestrado) — Universidade do Vale do Rio dos Sinos, 2017.

[21] MIURA, N. K. *Geração incremental de parsers dependentes de contexto para o português brasileiro.* Tese (Doutorado) — Universidade de São Paulo, 2019.

[22] COSTA, P. C. da. *Aplicação de Ontologias e Processamento de Linguagem Natural à recuperação de informações para Revisão Sistemática*. 2019. Trabalho de Conclusão de Curso - Universidade Federal de Santa Catarina.

[23] DALE, R. Classical approaches to natural language processing. In: *Handbook of natural language processing*. New York: CRC Press, Taylor & Francis Group, 2010. p. 3–7.

[24] DOMINGUES, M. L. C. S. *Abordagem para o desenvolvimento de um etiquetador de alta acurácia para o Português do Brasil.* Tese (Doutorado) — Universidade Federal do Pará, 2011.

[25] GODDARD, C. *Semantic analysis: A practical introduction*. New York: Oxford University Press, 2011.

[26] MÜLLER, D. N. *Processamento de linguagem natural*. Porto Alegre, 2003. Disponível em: <https://www.inf.ufrgs.br/~danielnm/docs/pln.pdf>.

[27] CHOMSKY, N. *Reflexões sobre a Linguagem*. São Paulo: Cultrix, 1975.

[28] OLIVEIRA, F. A. *Processamento de linguagem natural: princípios básicos e a implementação de um analisador sintático de sentenças da língua portuguesa*. VI Congresso Brasileiro de Informática na Educação, Porto Alegre, v. 31, n. 3, 2002.

[29] MORAN, T. P. The command language grammar: A representation for the user interface of interactive computer systems. *International journal of man-machine studies*, Elsevier, v. 15, n. 1, p. 3–50, 1981.

[30] ESTEVES, G. O. *Avaliação de Interação Humano-Computador: um estudo de caso para Bioinformática*. 2016. Trabalho de Conclusão de Curso - Universidade de Brasília.

[31] JAIN, J.; LUND, A.; WIXON, D. The future of natural user interfaces. In: *CHI'11 Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery, 2011. p. 211–214.

[32] BARBOSA, C. R. S. C. *Gramática para consultas radiológicas em língua portuguesa*. Dissertação (Mestrado) — Instituto de Informática da Universidade Federal do Rio Grande do Sul, 1998.

[33] SAVADOVSKY, P. *A construção de interpretadores para linguagem natural*. Curitiba: Imprensa Oficial/Government Parana, 1988.

[34] GARCÍA, L. S. *Linx: Um Ambiente Integrado de Interface para Sistemas de Informação Baseados em Conhecimento*. Tese (Doutorado) — Departamento de Informática Pontifícia Universidade Católica do Rio de Janeiro, 1995.

[35] SILVA, B. C. D. et al. Introdução ao processamento das línguas naturais e algumas aplicações. *Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional*, v. 3, 2007.

[36] REITER, E.; DALE, R. *Building natural language generation systems*. Cambridge: Cambridge University Press, 2000.

[37] CAGAN, T. Opinionated natural language generation. *Herzliya: The Interdisciplinary Center*, p. 96, 2016.

[38] SAMPAIO, G. S. *Desenvolvimento de uma interface computacional natural para pessoas com deficiência motora baseada em visão computacional*. Dissertação (Mestrado), 2018.

[39] HONNIBAL, M. *spaCy: Industrial-strength Natural Language Processing (NLP) with Python and Cython*. 2015. Disponível em: <https://spacy.io/>.

[40] NADEAU, D. *Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision*. Tese (Doutorado) — University of Ottawa, 2007.

[41] SPECK, R.; NGOMO, A.-C. N. Ensemble learning for named entity recognition. In: SPRINGER. *International semantic web conference*. Leipzig, 2014. p. 519–534.

[42] FERNANDES, C. V. *Modelagem de banco de dados não relacional em Plataforma Big Data visando dados de internet das coisas*. 2017. Trabalho de Conclusão de Curso da Universidade Federal de Mato Grosso.

[43] RAJ, S. *Neo4j high performance*. Birmighan: Packt Publishing Ltd, 2015.

[44] PANZARINO, O. *Learning Cypher*. Birmighan: Packt Publishing Ltd, 2014.

[45] HOLZSCHUHER, F.; PEINL, R. Xiii performance of graph query languages: Comparison of cypher, gremlin and native access in neo4j. In: *Proceedings of the Joint EDBT/ICDT 2013 Workshops*. Genua: Workshop GraphQ, 2013. p. 195–204.

[46] SANTOS, O. S. *A Cultura da Soja 1: Rio Grande do Sul-Santa Catarina-Paraná*. 2. ed. São Paulo: Globo, 1995.

[47] MOREIRA, H. J. C.; ARAGÃO, F. D. *Manual de Pragas da Soja*. Campinas: FMC Agricultural Products, 2009. 144 p.

[48] SOSA-GÓMEZ, D. R. et al. Manual de identificação de insetos e outros invertebrados da cultura da soja. *Embrapa Soja-Documentos (INFOTECA-E)*, Londrina: Embrapa Soja, 2014.

[49] AVILA, C. J.; GRIGOLLI, J. F. J. Pragas de soja e seu controle. *Embrapa Agropecuária Oeste-Capítulo em livro científico (ALICE)*, Tecnologia e produção: Soja 2013/2014., 2014.

[50] MICHALSKI, R. S. Learning by being told and learning from examples: an experimental comparison of the two methods of knowledge acquisition in the context of development an expert system for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems*, v. 4, n. 2, p. 125–161, 1980.

[51] BREIMAN, L. Random forests. *Machine learning*, Springer, Boston, v. 45, n. 1, p. 5–32, 2001.

[52] SCHÖLKOPF, B. et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Canadá: MIT press, 2002.

[53] WEINBERGER, K. Q.; BLITZER, J.; SAUL, L. K. Distance metric learning for large margin nearest neighbor classification. In: *Advances in Neural Information Processing Systems*. Cambridge: MIT press, 2006. p. 1473–1480.