Identificação de Pragas e Doenças na Cultura da Soja por meio de um Sistema Computacional em Linguagem Natural

Carolinne Roque e Faria
Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Londrina
Londrina, Paraná - Brasil
carolinne.rf@outlook.com

Cinthyan Renata Sachs Camerlengo de Barbosa
Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Londrina
Londrina, Paraná - Brasil
cinthyan@uel.com

ABSTRACT

Technology is becoming expressively popular among agribusiness producers and is progressing in all agricultural area. One of the difficulties in this context is to handle data in natural language to solve problems in the field of agriculture. In order to build up dialogs and provide rich researchers, the present work uses Natural Language Processing (NLP) techniques to develop an automatic and effective computer system to interact with the user and assist in the identification of pests and diseases in the soybean farming, stored in a database repository to provide accurate diagnoses to simplify the work of the agricultural professional and also for those who deal with a lot of information in this area. Information on 108 pests and 19 diseases that damage Brazilian soybean was collected from Brazilian bibliographic manuals with the purpose to optimize the data and improve production, using the spaCy library for syntactic analysis of NLP, which allowed the pre-process the texts, recognize the named entities, calculate the similarity between the words, verify dependency parsing and also provided the support for the development requirements of the CAROLINA tool (Robotized Agronomic Conversation in Natural Language) using the language belonging to the agricultural area.

KEYWORDS

Agriculture 4.0, Human-Computer Interaction, Natural Language Processing.

1 INTRODUÇÃO

Os recursos computacionais aliados aos meios de comunicação [1] oferecem a seus usuários cada vez mais funcionalidades e alternativas de usabilidade, que foram desenvolvidas para melhorar a interação com o usuário.

Atualmente, existem inúmeros invasores que prejudicam a cultura da soja e que estão distribuídos em regiões produtoras, em decorrência da falta de preparo em fazer uma análise do solo antes de realizar o plantio. Existem vários produtos registrados para o controle das pragas. Segundo a EMBRAPA *(Empresa Brasileira de Pesquisa Agropecuária)*, o Manejo Integrado de Pragas da Soja (MIP-Soja) [2] é uma tecnologia que utiliza um conjunto de técnicas econômicas e ambientalmente sustentáveis para o manejo eficiente de pragas que atacam as lavouras de soja. No entanto, nos últimos anos, o agricultor tem feito o uso de agrotóxicos de maneira desequilibrada [3], não rotaciona e não usa de forma racional os produtos para o manejo das pragas e insetos.

A identificação das pragas ou doenças na lavoura é feita a partir da observação da planta, seja a haste seca ou a coloração das folhas ou um grão murcho etc. De acordo com Chaudhury et al. [4], fazer uma análise visual da planta para detectar o vilão é ineficiente e difícil para plantações de grandes proporções, visto que o ataque acontece de forma inesperada e agressiva que requer um profissional qualificado e bem treinado para a realização de tal função [5].

Devido à grande participação do Brasil na produção de soja e ao aumento de produtividade, ter uma ferramenta que auxilie o produtor sobre as principais pragas e doenças na cultura, garante a sanidade das plantas, significa melhor produção e, desse modo, o agricultor pode lidar com as necessidades da lavoura. Sendo assim, 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) aplicadas em análises de grandes volumes de dados.

Considerando a importância da ciência e sistemas inteligentes associados à produção de conhecimento, as técnicas de PLN são fundamentais nas tomadas de decisão em práticas agrícolas, as quais pretendem intensificar a análise dos dados de uma determinada cultura e potencializar a produção do agricultor.

Este trabalho tem o intuito de facilitar o desenvolvimento de sistemas inteligentes e significativos aos agricultores por meio do PLN para identificar as ameaças que afetam a cultura analisada e melhorar a produtividade. Assim, uma Interface em Linguagem Natural para Banco de Dados (ILNBD) foi desenvolvida, a qual usa possíveis perguntas dos agricultores e estudantes de agronomia para pesquisas quanto às pragas e doenças na cultura da soja. Problemas linguísticos dessas perguntas em LN foram levantados e estão sendo estudados soluções para a implementação de um *chatbot*.

Este artigo está organizado da seguinte forma: a seção 2 apresenta os trabalhos relacionados; a seção 3 apresenta a fundamentação teórica sobre Processamento em Linguagem Natural; a seção 4 traz questões sobre estudos em ILNBDs e problemas linguísticos encontrados para consultas sobre pragas e doenças na cultura da soja; a seção 5 aborda a arquitetura geral desse sistema e as etapas desenvolvidas; a metodologia do desenvolvimento deste trabalho será apresentada na seção 6; o levantamento de algumas perguntas da ILNBD sobre pragas e

doenças da soja é apresentado na seção 7; por fim a seção 8 aponta as considerações finais.

2 TRABALHOS RELACIONADOS

A automatização de ferramentas para melhorar o desempenho de suas tarefas aumenta a cada dia em todo mundo e no caso da agricultura não é diferente. O AgroPortal [6] reutiliza as ferramentas e *insights* semânticos do domínio biomédico para atender não apenas a agronomia, mas também às ciências da alimentação, das plantas e da biodiversidade. Foi oferecido um portal que apresenta hospedagem, pesquisa, versionamento, visualização, comentário e recomendação de ontologias. Nesse referido projeto foi apresentado o conteúdo e os recursos da plataforma, incluindo os acréscimos à tecnologia original, bem como os resultados preliminares de cinco casos de uso agronômico que eram as principais fontes de ontologias e vocabulários. Com base na experiência e na tecnologia existente adquirida no domínio biomédico, o AgroPortal apresentou um repositório robusto e rico em recursos e de grande valor para o domínio agronômico.

A detecção precoce de doenças foliares no tomateiro [7] no Brasil é feita pela utilização de uma chave de classificação manual baseada em imagens da cultura que quantificam o grau de infestação e contaminação por *P. infestans.*

O trabalho Manejo Tecnológico de Lavouras por meio de Dispositivos Móveis e Agricultura de Precisão [8] apresenta uma abordagem computacional focada na melhoria da qualidade da cultura do tomate, com baixo custo, baseada em técnica de agricultura de precisão para apoiar os agricultores na inspeção das lavouras e na detecção precoce da principal doença que afeta a cultura do tomate.

Foi feito um estudo sobre a análise de textura e cor das folhas de arroz *(Oryza sp.)* por uma simulação que contém 400 amostras para identificar [9] as doenças *brown spots* e *blast diseases* e a sua classificação foi realizada por uma Rede Neural *Perceptron* Multicamadas. As imagens de folhas doentes e sadias atingiram uma precisão de 89,26% de acertos na classificação dos *pixels*.

Um *chatbot* para a divulgação do Atlas Linguístico do Brasil [10] apresenta características de interdisciplinaridade na área de Linguística e Ciências da Computação, com o foco principal em Processamento de Linguagem Natural. A ideia com esse *chatbot*, além de divulgar o Atlas do Brasil, é despertar em outros pesquisadores o desejo de desenvolver outros assuntos que sejam peculiares ao ensino. Nesse trabalho foi utilizada a ferramenta *WhatsApp* para fazer as pesquisas. Pôde-se perceber também nesse chatbot que há casos de uso de sinônimos da agricultura como variantes para palavra *mandioca* dependendo da região do Brasil, como *macaxeira*, *aipim*, etc.

Por outro lado, o trabalho de Rocha e Sartin [11] alia a *visão computacional* à análise de folhas de soja, com o objetivo de encontrar e extrair características que permitam a detecção de doenças foliares. Tal trabalho faz uso de pré-processamento da imagem de folhas de soja, utilizando-se de filtros de média, mediana e métodos de detecção de bordas e linhas para identificar a folha da soja. O trabalho também faz uso de segmentação de imagens, porém com uma dificuldade na obtenção das veias nas folhas de soja mais escuras.

Ferreira [12] realizou estudos de cultura de soja, construindo um banco de imagens com mais de 15 mil imagens do solo, soja e ervas daninhas de folhas largas e gramíneas. A partir dessas fotos uma *Rede Neural Convolucional* foi treinada para detectar as ervas daninhas, sendo os resultados comparados com os Algoritmos de Máquina de Vetor Suporte, sendo que as Redes Neurais Convolucionais obtiveram uma precisão de mais de 98% na detecção das ervas daninhas de folhas largas e gramíneas.

Alguns trabalhos, como [13], têm como enfoque identificar dados, informações e conhecimento no processo de *tomada de decisão*. Múltiplos critérios podem ser usados. Para atingir esse objetivo foi preciso fornecer definições adequadas de dados, informações e conhecimentos, bem como algoritmos específicos dos modelos de tomada de decisão. Uma abordagem sobre o problema de seleção de secadores de grãos sob condições de uma empresa agrícola específica foi abordada no referido trabalho.

Por fim cabe ressaltar que não foi possível encontrar trabalhos que utilizam língua natural para identificação de pragas e doenças na cultura da soja. Assim, este trabalho será uma grande contribuição aos profissionais da agricultura.

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 [14].

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

Para o desenvolvimento de uma interface voltada para a área do Agronegócio, faz-se uso das ferramentas de PLN capazes de disponibilizar módulos autômatos para realizar tarefas específicas e especializadas e outros módulos que armazenam um modelo de conhecimento proposicional.

As ferramentas de PLN baseiam-se na aplicação de técnicas para analisar dados textuais e têm 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 [15].

A utilização de uma ILNBD para identificação de pragas e doenças na sojicultura para profissionais da área que lidam com um amplo volume de informações por meio de um sistema de geração de frases em língua natural é vantajosa, pois o usuário não precisa aprender a se comunicar com o sistema, suportam anáforas e elipses e as perguntas são facilmente expressas em linguagem natural e permitem frases curtas.

Para Santos [16], os processos de Compreensão de Linguagem Natural (CLN) extraem informações em linguagem natural, tratam palavras ambíguas, sinônimos e polissemia para permitirem a sua manipulação por parte dos computadores e a Geração de Linguagem Natural (GLN) é o processo de produção automática de texto a partir de dados estruturados em um formato legível com frases significativas.

O léxico é uma estrutura de dados onde são armazenadas as palavras e associadas a elas algumas de suas informações. Podem ser usadas duas categorias de dicionários: o Dicionário de Formas, que apresenta a estrutura completa das palavras e auxilia no reconhecimento dessas e o Dicionário de Bases que é formado pelos radicais e terminações das palavras de origem [19].

A arquitetura de um sistema de PLN genérico ajusta segundo as suas características de aplicação, como o tradutor automático, que é capaz de [17]: extrair cada uma das palavras da sentença e identificá-las; analisa sintaticamente a sentença; elabora uma nova sentença para retomar o sentido das informações levantadas anteriormente, ou seja, sintetizar um significado absoluto da mesma, a partir dos significados das palavras e das relações entre elas; associa o significado adquirido de uma representação adequada, que pode ser independente da língua destino; associa as palavras equivalentes da língua origem para a língua destino.

Ao executar essas aplicações é possível escrever um conteúdo informacional de inúmeras maneiras, manipulando o sistema para que comande automaticamente a geração de tarefa de acordo com sistemas que têm a função específica de declarar ou informar em uma base de dados em PLN.

4 INTERFACE EM LINGUAGEM NATURAL PARA BANCO DE DADOS

A interface consiste na interação entre o usuário e o sistema e a ligação de ambos permite a execução de tarefas e pode ser descrita por três elementos: dispositivos de entrada e saída, que são meios físicos da interação; modelo conceitual refere-se à visão do usuário perante as funcionalidades do sistema e a interação do sistema, que permite que os recursos utilizados realizem as tarefas desejadas do usuário.

De acordo com Jain et al. [18], 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 que esse entenda como manuseá-la, seja 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 (no caso deste trabalho, o português) para que os usuários apenas pensem nas ações que querem expressar, sem investir tempo para aprender uma linguagem de programação.

As vantagens de uma ILNBD [19] 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.

Ainda em [19] 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 as 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: *Toda praga prejudica alguma cultura?*

A pergunta tem duas leituras:

a) ∀ praga ∃ cultura prejudica (praga, cultura);
b) ∃ cultura ∀ praga prejudica (praga, cultura).

A leitura (a) permite a cada praga prejudicar uma cultura diferente enquanto em (b) toda praga prejudica a mesma cultura. A intenção é manter a ordem dos quantificadores da esquerda para a direita. Métodos heurísticos apresentados em Barbosa [19] podem resolver ambiguidade de escopo.

● **Conjunção e disjunção**: a palavra "*e*" pode denotar disjunção ao invés de conjunção. Exemplo: *Teve queimada e granizo hoje?* Deve ser excluída a possibilidade do "e" denotar uma conjunção, visto que no dia normalmente não pode ter acontecido as duas possibilidades (queimada e granizo) e transformaria os *e*'s em *ou*'s.

● **Nominais compostos:** são "substantivo-substantivo" e "adjetivo-substantivo". Esses nominais dificultam a determinação de seus significados. Exemplo de substantivo-substantivo é

encontrado na frase: *O professor agricultor diagnosticou a doença.* Podem-se encontrar dificuldades no significado da frase apresentada, pois seria um professor que também exerce a função de agricultor ou seria um agricultor que também exerce a função de professor? Como exemplo de adjetivo-substantivo tem-se: *grande pesquisador*. Aqui poderia denotar um pesquisador de estatura alta ou pesquisador altamente qualificado. Para evitar dúvidas na compreensão, sugere-se [19] definir os significados de cada composto durante a fase de configuração dos SGBDs (Sistema Gerenciamento de Banco de Dados). Assim quando um substantivo é ensinado como, por exemplo, *professor*, os adjetivos desse são perguntados quais são: *doutor*, *mestre*, *concursado*, *temporário*, etc. O mesmo para outros substantivos vinculados ao *professor* como: *chefe*, *orientador*, *tutor* etc.

● **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*. "Ela" é uma referência anafórica ou anáfora. Aqui poderia ter ambiguidade se "ela" refere-se à praga ou ameaça. Em Barbosa [19] mostra também possíveis soluções que as ILNBDs dão para o tratamento dessas referências anafóricas.

● **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.* Nessa frase sabe-se que "conferiu" refere-se à 3ª pessoa do singular na Língua Portuguesa.

● **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 reconhece os vários tipos de diagnósticos.

5 ARQUITETURA DE UMA INTERFACE EM LINGUAGEM NATURAL

O desenvolvimento de uma arquitetura de uma Interface em Linguagem Natural (ILN) depende do domínio de aplicação [20] e é necessário tentar abstrair essa dependência para obter uma arquitetura mais genérica e independente do contexto, com estruturas básicas para o PLN, o que adequa para uma melhor compreensão do desenvolvimento de interfaces em LN.

É preciso compreender o "público" que será atingindo para elaborar a interface em LN, considerar os objetivos da comunicação para definir as regras gramaticais e como esses fatores influenciam os sistemas computacionais. A Fig. 1 apresenta uma arquitetura genérica para uma ILN.

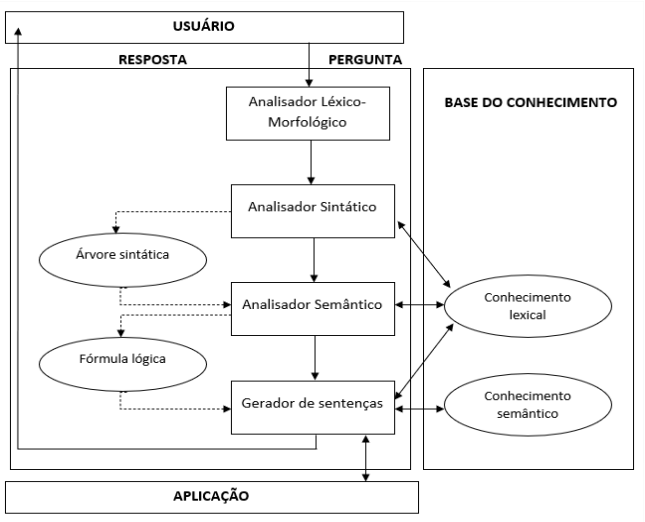

Figura 1: Arquitetura genérica de uma ILN

A eficiência da interface deve-se à base do conhecimento que contém informações necessárias ao processamento da sentença do usuário junto à aprendizagem interativa do utilizador. Essa base é constituída por:

● **Conhecimento lexical**: dicionário que armazena léxicos com seus significados e classe gramatical. Para o processamento de qualquer linguagem por meio de uma gramática, informações relativas aos tokens da linguagem são relevantes [20];

● **Conhecimento semântico:** termos dependentes do domínio da aplicação e das relações entre si sobre as categorias dos termos;

● **Analisador léxico-morfológico:** classifica gramaticalmente cada palavra de uma sentença de acordo com suas características morfológicas, pois é necessário compreender o significado de cada palavra para compreensão de uma frase;

● **Analisador sintático:** por meio da relação de palavras, obtêm-se as informações e para explicitar essa relação nessa fase, permite-se construir árvores de derivação;

● **Analisador semântico:** depois de relacionada as palavras de uma frase na árvore de derivação, é feita uma análise das palavras de uma sentença para ser compreendidas pelo sistema;

● **Gerador de sentenças/ações:** a partir da análise semântica é possível construir sentenças em LN baseado nas ações para serem executadas pelas aplicações.

6 METODOLOGIA E DESENVOLVIMENTO DO SISTEMA

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.

As etapas para o desenvolvimento do Sistema da Soja são baseadas em Sampaio [21], compostas por: • definições das funcionalidades do sistema para permitir que o usuário controle as funções do sistema; • 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.

7 TIPOS DE CONSULTAS NA ILNBD SOBRE PRAGAS E DOENÇAS DA SOJA

Este software pretende contribuir nas tomadas de decisões dos profissionais da área, com o uso de técnicas de PLN para otimizar as atividades guiadas à agricultura.

O dicionário utilizado é uma estrutura de dados, no qual as palavras são armazenadas e associadas a elas algumas de suas informações [14]. O repertório das palavras registradas no dicionário conta com 108 pragas e 19 doenças no cultivar da soja, selecionadas por manuais e indicações bibliográficas [22], [23], [24], [25].

Estão sendo validados por meio de um analisador léxico os principais aspectos que identificam as ameaças da sojicultura, considerando a categoria morfológica, nome científico, características biológicas, comportamento, danos, controle, localização na planta e localização geográfica e ciclo.

Posto que este projeto tem o objetivo de colaborar com o profissional da área da agricultura, foi possível desenvolver a ferramenta CAROLINA (acrônimo para **C**onversação **A**gronômica **RO**botizada em **LI**nguagem **NA**tural) e foi adotado o modelo em cascata por meio das fases de análises de requisitos, projeto, implementação, testes (validação), integração e manutenção de software, o que permitiu 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 que eu devo fazer?"; "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?".

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 foi desenvolvido um sistema para dialogar com agricultores sobre as principais características das principais pragas e doenças na cultura da soja. Este sistema recebe o texto e analisa as palavras de uma sentença isoladamente, como é visualizado na Fig. 2.

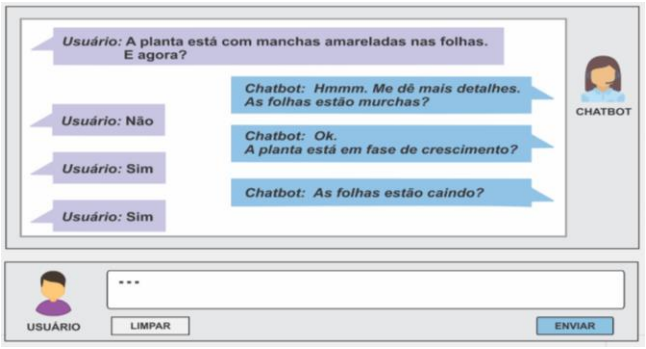

Figura 2: Interface de Consulta à Base de Dados das Pragas da Soja

A seguir, um exemplo de como o sistema analisa/processa as perguntas, extrai as informações armazenadas no banco de dados e fornece a resposta.

I) Pergunta: "Qual a localização da Lagarta-da-maçã-do-algodoeiro na soja?"

Análise: Pronome + artigo + substantivo + preposição + substantivo composto + preposição + substantivo

II) Resposta: "As Lagartas-da-maçã-do-algodoeiro atacam as vagens."

Para reconhecer as referências dos diagnósticos são estabelecidas algumas regras necessárias. Baseado no trabalho de Barbosa [14] foi possível trabalhar o domínio proposto para os tipos de construção:

a) Grupos gramaticais, como: - **Sentenças;** - **Sintagma nominal**, em que o trecho da oração é que define completamente uma entidade ou conjunto de entidades do mundo do falante [26]. Exemplos: "*A praga*"; "*Algum dano*"; "*Toda planta"*; - **Sintagma verbal** determina uma atividade ou estado no tempo, como os verbos. Se o verbo for intransitivo (exemplo: "*morreu*") forma-se um sintagma verbal. Se for transitivo (exemplo: "*os ácaros atacam as plantas*") deve haver um sintagma nominal que é o objeto da atividade para completar o sintagma verbal; - **Sintagma adjetival** que por meio de um verbo de ligação atribui qualidades a um sintagma nominal ou qualifica um verbo intransitivo ou sintagma nominal. Exemplo: "*A haste está podre*"; - **Sintagma preposicional** que composto por uma preposição seguida de um sintagma nominal. Exemplo: "*A planta está comprometida de*

ferrugens"; - **Sintagma adverbial** que é formado por um ou mais advérbios seguidos ocasionalmente por uma preposição. Exemplo: *"A praga certamente afetará a raiz da planta"*; - **Oração subordinada adjetiva restritiva** que são as que limitam a extensão do nome a que se referem. Esse tipo de oração inicia-se por um pronome relativo (que, quem, o(a) qual, os(as) quais, etc.). Exemplo: "*Os percevejos que atacaram a planta*";

b) Sentenças Sim/Não: são as que procuram o valor verdade de uma fórmula, seja ela "True/False", como: *"A haste está podre?"*;

c) Sentenças –*wh*: procuram valor de instanciação de funções exclamativas, como em *"A folha da soja está manchada!"*;

d) Sentenças alternativas na forma normal (não-clivada): sentenças alternativas feitas no predicado. Exemplo: *"Apresenta manchas amareladas ou púrpuras?"*;

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

f) S existencial: *"Alguma parte da planta foi analisada?"*;

g) S na voz ativa: *"A manifestação compromete alguma parte da planta?"*;

h) S na voz passiva: *"A planta está comprometida pela infestação?"*;

i) S clivada: extraposição do verbo ser em *"É a mancha que confirmou o diagnóstico?"*.

A representação da linguagem verbal interpretável/gerável é feita pela gramática. Essa contém regras de estruturação sintática e de morfossintaxe (gênero e número) [14]. Contém valores possíveis das categorias sintáticas de nível mais baixo, como de complementos "tempo", "direção", "lugar" e outros [27].

As palavras armazenadas no dicionário são associadas às informações gramaticais. As principais pragas e doenças foram cadastradas nesse dicionário totalizando 108 ameaças.

O registro de cada palavra foi feito de acordo com algumas categorias morfológicas, as quais são avaliadas isoladamente e separadas em classes gramaticais pelo processo de POS *tagging*, isto é, para cada token de um texto é feita a identificação da classe gramatical a que ele pertence baseado em sua definição e no contexto, visto que uma palavra pode pertencer a mais de uma classe ou possuir mais de um significado [28].

Miura [29] salienta que a análise de uma sentença em linguagem natural pode resultar em mais de uma possível interpretação. Essa ambiguidade pode se manifestar de diversas formas, sendo elas: morfológica (léxica, flexiva ou léxico-flexiva), sintática (identificação do constituinte ou coesão) ou semântica (léxica, unidade polilexicais, escopo ou papel temático).

O spaCy é uma biblioteca para análise sintática e, segundo Allen [30], o analisador sintático tem três particularidades: **a)** um analisador tem como entrada uma sentença e como resultado produz a análise; **b)** uma gramática é obtida a partir de um conjunto de regras que o analisador pode utilizar; **c)** um léxico é definido por um dicionário de palavras corretas e *parts of speech.*

Dada à importância do assunto, o presente trabalho conduz a análise sintática, na qual visou gerar uma estrutura de árvore que represente a estrutura sintática da sentença analisada. O sistema foi desenvolvido por meio da linguagem de programação Python e o analisador sintático escolhido foi o SpaCy. Assim, foi possível fazer o processo de *part of speech tagging (*classificar todas as sentenças do texto por categorias gramaticais e a relação de dependência entre palavras).

É destacado por Ferreira [31] que ao nível de palavras, as principais *parts of speech tags* fornecem informações significativas sobre uma palavra em seu contexto. Testes foram realizados [32] com diversos analisadores e a biblioteca *SpaCy* apresentou suportar todos os requisitos e condições para o desenvolvimento do sistema apresentado nesta dissertação, além de ser simples, suportar a Língua Portuguesa e ser código aberto.

Esta ferramenta também permite fazer o Reconhecimento de Entidade Nomeada, como é conhecida em português, tem a função de identificar e classificar palavras ou frases em um texto de acordo com classes definidas para o modelo [33]. Conforme abordado em Speck e Ngomo [34], 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 e identificar sobre o que se trata.

O BD NoSQL do tipo grafo escolhido foi o Neo4j, que possui uma interface robusta e flexível perante outros BDs da mesma categoria. A vantagem de utilização do modelo baseado em grafos fica bastante clara quando consultas complexas são exigidas pelo usuário por ter um ganho de performance, permitindo um melhor desempenho das aplicações [35], como é mostrada na Fig. 3.

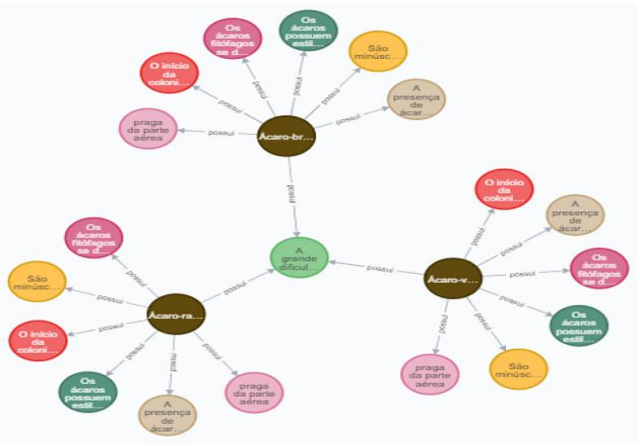

Figura 3: Dados dos Ácaros Fitófagos no Banco de Dados NoSQL Neo4j

A Fig. 3 tem 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. O

objetivo foi 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 aumentá-lo 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.

Na Tabela 1 é apresentada a extração das etiquetas morfológicas da sentença e a atribuição das mesmas no formato adequado para classificação de cada palavra pelo framework spaCy.

Tabela 1 - Aplicação de etiquetas morfológicas na sentença da base de dados

Texto puro	**'Coró-da-soja: Essas pragas atacam as raízes da soja.'**
Texto marcado	Coró-da-soja NOUN : PUNCT Essas DET pragas NOUN atacam VERB as DET raízes VERB da DET soja NOUN . PUNCT

Identificar as partes da frase é essencial porque ajuda a entender com mais precisão as frases de entrada e constrói com mais precisão as frases de saída (resposta).

A Fig. 4 apresenta o processo de *POS tagging* dos textos por meio do spaCy. Classifica todas as sentenças por categorias gramaticais e a relação de dependência entre as palavras, como em *O Coró-da-soja é uma praga que ataca a raiz da soja*.

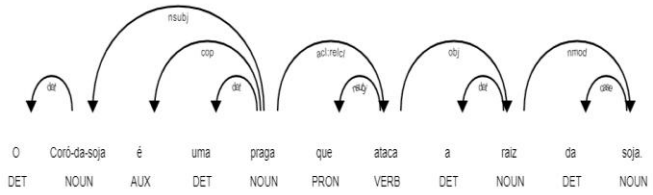

Figura 4: Relação de dependência

Assim, foi possível identificar e reconhecer elementos importantes contidos no texto com precisão (entidades nomeadas) em categorias pré-definidas no campo agronômico por meio do modelo NER, como é apresentado na Tabela 2, aplicando em uma sentença.

Tabela 2 - Reconhecimento de Entidades Nomeadas

Texto puro	**'As lagartas de S. Frugiperda estão normalmente presentes na cultura a ser dessecada para o plantio de soja. Ocorrem nos cotilédones ou em plântulas durante os estádios iniciais de desenvolvimento da cultura.'**
Reconhecimento de Entidades Nomeadas (NER)	[(s. frugiperda, 'PER'), (plântulas, 'LOC')]

A técnica utilizada permitiu a identificação específica das entidades na frase e classificou de acordo com o significado em questão, como a localização de danos/ataques nas plantas do tipo LOC e o indivíduo/ pragas do tipo PER. A entidade LOC significa localização e a entidade PER significa uma entidade de nome pessoal, nesse caso, as pragas.

Além da aplicação do NER, os principais aspectos que identificam as ameaças da sojicultora estão sendo validados por meio de um analisador léxico, considerando os atributos das pragas.

Estruturas básicas das famílias de árvores utilizadas são utilizadas no domínio da soja, como em [14]: **Verbos intransitivos:** é selecionada por verbos que não precisam de complemento e possui sentido completo, como em *A planta apodreceu*. **Verbos transitivos diretos:** é selecionada por verbos que pedem um complemento direto NP, 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: *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 são manipuladas possuir árvores separadas dentro da família de árvores, por exemplo: *A lavoura está comprometida pelo ataque de pragas.*

8 Considerações Finais

A proposta deste trabalho foi implementar um sistema inteligente para interagir com o usuário por meio de frases referentes às principais pragas e doenças da soja. Uma gama dessas frases foi analisada para que fosse possível realizar o Processamento em Linguagem Natural.

Esta ferramenta teve como propósito concentrar as informações sobre as principais pragas e doenças da sojicultura em um repositório e auxiliar o usuário na tomada de decisão do agricultor por meio de consultas.

Um modelo foi criado para extrair os textos do banco de dados Neo4j. Como objetivo principal teve a função sintática, responsável por organizar as estruturas gramaticais. Assim, foi possível a construção de um modelo por meio do framework spaCy para aplicar regras gramaticais à cada sentença, reconhecer a estrutura e extrair seus significados, além de permitir treinar o conjunto para refinar os dados.

A criação deste sistema pretende otimizar a análise dos dados de uma cultura e aperfeiçoar a produção do agricultor a partir do diagnóstico, obtendo bom desempenho de análise de *parsing*, por meio da linguagem própria desse público.

Devido à pandemia do novo Coronavírus não foram feitos testes de usabilidade com os profissionais do campo agrícola para validar o nível de aceitação da ferramenta para verificar se esses conseguem compreendê-la, manipulá-la facilmente e se funciona da forma ideal para os mesmos. Assim, os autores pretendem dar continuidade a este trabalho por meio de uma próxima etapa de testes.

REFERÊNCIAS

[1] Luciana Ferreira Costa e Francisco Arruda Ramalho. 2010. A usabilidade nos estudos de uso da informação: em cena usuários e sistemas interativos de informação. *Perspectivas em ciência da informação*, 15, 1 (Jan./Abr 2010), 92-117.

[2] Embrapa. 1975. *Manejo Integrado de Pragas da Soja (MIP-Soja).* https://www.embrapa.br/soja/publicacoes, Abril.

[3] Silvia Cristina Vieira Gomes, João Luís Bazzo Florindo, Bruno Quiqueto Montezani, Yago Vieira, Fabrício Rimoldi. 2018. Convergências entre NR 31 e NR 06: mitigando efeitos nocivos do uso de agroquímicos no espaço geográfico do interior paulista. In *Anais do I Congresso sobre Ambiente, Tecnologia e Educação* (*CATE'18)*, IFSP, Tupã, 1-18.

[4] Gobinda Chowdhury. 2003. Natural Language Processing, *Annual Review of Information Science and Technology*, 37, (Jan. 2003), 51-89. DOI: 10.1002/aris.1440370103

[5] Pedro Henrique da Silva Pereira, Gabriel Oliveira; Rafael Shoiti S. Yokoo, Mariana M. S. Rodrigueiro e Luiz F. S. Coletta. 2018. Análise de descritores de imagens na classificação de folhas de soja visando o diagnóstico de doenças. In *Anais do X Simpósio Nacional de Tecnologia em Agronegócio* (*SINTAGRO'18)*, Fatec, Presidente Prudente, 89-100.

[6] Clement Jonquet, Anne Toulet, Elizabeth Arnaud, Sophie Aubin, Esther D. Yeumo, Vincent Emonet, John Graybeal, Marie Laporte, Mark A. Musen, Valeria Pesce, Pierre Larmande. 2018. AgroPortal: a vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture*, 144 (Jan. 2018), 126-143. DOI: 10.1016/j.compag.2017.10.012

[7] Fabio Mathias Correa, Julio Silva de Sousa Bueno Filho and Margarida Gorete Ferreira do Carmo. 2009. Comparison of Three Diagrammatic Keys for the Quantification of Late Blight in Tomato Leaves. *Plant Pathology*, 58, (Nov. 2009), 1128-1133. DOI: 10.1111/j.1365-3059.2009.02140.x

[8] Sérgio Manuel Serra da Cruz, Diogo. Nunes, Carlos Werly, Pedro Vieira da Cruz, Ana Claudia de Macedo Vieira e Marden Manuel Marques. 2015. Manejo Tecnológico de Lavouras através de Dispositivos Móveis e Agricultura de Precisão. In *Anais do XI Brazilian Symposium on Information System* (*SBSI'15)*, SBC, Goiânia, 379-386. DOI: 10.5753/sbsi.2015.5840.

[9] Pritimoy Sanyal and Sandip C. Patel. 2013. Pattern recognition method to detect two diseases in rice plants. *Imaging Science Journal*, 56, 6 (Julho 2013), 319-325. DOI: 10.1179/174313108X319397.

[10] Fábio Carlos Moreno, Edio Roberto Manfio, Cinthyan Renata Sachs Camerlengo de Barbosa e Jacques Duílio Brancher. 2015. Tical: Chatbot sobre o Atlas Linguístico do Brasil no WhatsApp. In *Anais do XXVI Simpósio Brasileiro de Informática na Educação (SBIE'15)*, SBC, Maceió. 279-288. DOI: http://dx.doi.org/10.5753/cbie.sbie.2015.279

[11] Ivan Augusto Alves da Rocha e Maicon Aparecido Sartin. 2018. Pré processamento e segmentação de imagens de folhas de soja com base na visão computacional. In *Anais do II Workshop de Tecnologias Emergentes em Computação (ETEC'18),* Even3, Sinop. DOI: 10.29327/17779

[12] Alessandro dos Santos Ferreira. 2017. *Redes Neurais Convolucionais Profundas na Detecção de Plantas Daninhas em Lavoura de Soja*. Campo Grande: FACOM da Universidade Federal do Mato Grosso do Sul. 80p. Dissertação de Mestrado.

[13] Martina Houskova Beránková and Milán Houska. 2011. *Data, information and knowledge in agricultural decision-making.* AGRIS on-line Papers in Economics and Informatics, v. 3 (Junho 2011), 74-82.

[14] Cinthyan Renata Sachs Camerlengo de Barbosa. 2004. *Técnicas de Parsing para Gramática Livre de Contexto Lexicalizada da Língua Portuguesa.* São José dos Campos: Departamento de Engenharia Eletrônica e Computação do Instituto Tecnológico de Aeronáutica. 171p. Tese de Doutorado.

[15] Emílio Luiz Faria Rodrigues. 2017. *Geração de perguntas em linguagem natural a partir de bases de dados abertos e conectadas:* um estudo exploratório. São Leopoldo. PIPGCA da UNISINOS. 93p. Dissertação de Mestrado.

[16] Cedric Michael Santos. 2015. *Classificação de Documentos com Processamento de Linguagem Natural.* Coimbra. Departamento de Engenharia Informática e de Sistemas do Instituto Superior de Engenharia de Coimbra. 120p. Dissertação de Mestrado.

[17] André M. M. Neves. 2005. *iAIML:* Um mecanismo para o tratamento de intenção em chatterbots. Recife: PROPGCC do Centro de Informática da Universidade Federal de Pernambuco. 122p. Tese de Doutorado.

[18] Jhilmil Jain, Arnold Lund and Dennis Wixon. 2011. The future of natural user interfaces. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems* (*CHI'11*), ACM Press, Sinop. 211-214. DOI: 10.1145/1979742.1979527

[19] Cinthyan Renata Sachs Camerlengo de Barbosa. 1997. Interfaces em Linguagem Natural para Banco de Dados. Porto Alegre: CPGCC da Universidade Federal do Rio Grande do Sul. 165p. Trabalho Individual nº 640.

[20] João Luis Tavares da Silva. 1995. *Técnicas para o Desenvolvimento de Interfaces Homem-Máquina em Linguagem Natural* – Considerações. Porto Alegre: PG da Pontifícia Universidade Católica do Rio Grande do Sul. 39f. Trabalho Individual I.

[21] Gustavo Scalabrini Sampaio. 2018. *Desenvolvimento de uma Interface Computacional Natural para Pessoas com Deficiência Motora baseada em Visão Computacional*. São Paulo: PROPG do Mackenzie. 107p. Dissertação de Mestrado.

[22] Osmar Souza dos Santos. 1995. *A Cultura da Soja 1*. (2ª ed.). Rio Grande do Sul-Santa Catarina-Paraná. São Paulo: Globo.

[23] Henrique José da Costa Moreira e Flávio Damasceno Aragão. 2009. *Manual de Pragas da Soja.* Campinas: FMC Agricultural Products. 144p.

[24] Daniel Ricardo Sosa-Gómez, Beatriz Spalding Côrrea-Ferreira, Clara Beatriz Hoffmann-Campo, Ivan Carlos Corso, Lenita Jacob Oliveira, Flávio Moscardi. 2006. *Manual de identificação de insetos e outros invertebrados da cultura da soja.* Londrina, PR: Embrapa Soja. Documentos 269.

[25] Crébio José Ávila. 2017. *Pragas da soja e seu controle*. https://pragas.cpao.embrapa.br

[26] Pedro Savadovsky. 1988. *Construção de Interpretadores para Linguagem Natural*. Curitiba: EBAI, 108p.

[27] Laura Sanchez García. 1995. *LINX: Um Ambiente Integrado de Interface para Sistemas de Informação Baseado em Conhecimento*. Rio de Janeiro: CPG Informática da Pontifícia Universidade Católica do Rio de Janeiro. 191p. Tese de Doutorado.

[28] Marcos Paulo Moraes. 2019. *Mineração de Dados Aplicada à Identificação de Notícias Falsas*. Rio de Janeiro: Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro. 47p.

[29] Newton Kiyotaka Miura. 2019. *Geração incremental de parsers dependentes de contexto para o português brasileiro.* São Paulo: Departamento de Engenharia de Computação e Sistemas Digitais. Escola Politécnica da Universidade de São Paulo. 132p. Tese de Doutorado.

[30] Joseph Allen. 1995. *Natural Language Understanding* (1st ed). Redwood City, Califórnia: Benjamin-Cummings Publishing Cp.

[31] Renato Cesar Borges Ferreira. 2017. *Uma Abordagem Semiautomática para Identificação de Elementos de Processo de Negócio em Texto em Linguagem Natural*. Porto Aleegre: CPGCC da Universidade Federal do Rio Grande do Sul. 103p. Dissertação de Mestrado.

[32] Jinho Choi, Joel Tetreault and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In *Anais da VII Conferência Conjunta Internacional sobre Processamento de Linguagem Natural (JCNLP'15)*, ACL, Beijing, China. 387-396. DOI: 10.3115/v1/P15-1038.

[33] David Nadeayu. 2007. *Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision.* Canadá: Institute for Computer Science de Ottawa-Carleton. 150p. Tese de Doutorado.

[34] René Speck and Axel-Cyrille Ngonga Ngomo. 2014. *Ensemble learning for named entity recognition.* Department of Computer Science University of Leipzig, Alemanha, v. 8796, p. 519–534.

[35] Bernadette Farias Lóscio, Hélio Rodrigues de Oliveira e Jonas César de Sousa Pontes. 2011. NoSQL no desenvolvimento de aplicações Web colaborativas. In *Anais do VIII Simpósio Brasileiro de Sistemas Colaborativos (SBSC'11)*, v. 10, n. 1, UFF, Paraty.