XVII Encontro Nacional de Pesquisa em Ciência da Informação **(XVII ENANCIB)**

GT 8: Informação e Tecnologia

A INTERPRETAÇÃO SEMÂNTICA DE TEXTOS CIENTÍFICOS EM PORTUGUÊS: UMA PERSPECTIVA EM CIÊNCIA DA INFORMAÇÃO

THE SEMANTIC INTERPRETATION OF SCIENTIFIC TEXTS IN PORTUGUESE: AN INFORMATION SCIENCE PERSPECTIVE

Dominique Lira Vieira Corrêa[1], Piotr Trzesniak[2], Raimundo Nonato Macedo Santos[3]

Modalidade da apresentação: Comunicação Oral

Resumo: Investiga os requisitos da busca semântica para a aplicação em textos científicos, através da análise da extração de relacionamentos semânticos do tipo "causa e efeito" em 60 resumos, em português, de artigos científicos da área de Ciências Agrárias. O estudo apresentou, por meio de considerações de ordem qualitativa e quantitativa, uma comparação entre o processo manual e automático de extração de sentenças de causa e efeito. Esses documentos foram previamente analisados de forma manual, e as sentenças de causa e efeito foram extraídas através da leitura dos resumos. Objetiva comparar as sentenças identificadas diretamente pelo pesquisador e as sentenças reconstruídas automaticamente a partir de programa implementado em planilha eletrônica. Conclui enfatizando que o uso de técnicas automáticas acelera o processo de extração de relações de causa e efeito e pode ser usada como alternativa ao processo custoso de identificação manual de informações semânticas. O resultado mais expressivo da presente pesquisa é o estabelecimento preliminar de rotinas para a versão automatizada.

Palavras-chave: Recuperação da Informação. Busca semântica. Sentenças de causa e efeito.

Abstract: *It investigates the semantic search technology for the application in scientific texts, by analyzing the extraction of semantic relationships such as "cause and effect" in 60 abstracts, in Portuguese, of scientific articles in the area of Agricultural Sciences. The study shows, through*

[1] Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Pernambuco.
[2] Professor Visitante no Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Pernambuco.
[3] Professor do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Pernambuco

qualitative and quantitative considerations, a comparison between manual and automatic extraction process of cause and effect sentences. These documents were previously analyzed manually, and the sentences of cause and effect were extracted by reading the summaries. The goal is to compare the sentences identified directly by the researcher and sentences automatically reconstructed from the set of programmed cells. The research concludes emphasizing that the possibility of using automatic techniques accelerates the process of creating and extracting of cause and effect relationship and may be used as an alternative to costly manual process of identifying semantic information. We can point out as the most significant result of this research the preliminary establishment of routines for automated version.

Keywords: *Information Retrieval. Semantic Search. Cause and effect sentences.*

1 INTRODUÇÃO

O enorme volume de informações disponíveis torna difícil não apenas recuperá-las, mas, também, gerenciar a recuperação e seus resultados de modo inteligente. Por essa razão, se faz necessário buscar respaldo teórico e prático no intuito de conhecer os recursos eficazes para organizar e dar acesso ao conhecimento. Nesse sentido, o Observatório Temático e Laboratório – Ensino, Tecnologia, Ciência e Informação (OtletCI) funciona como facilitador do processo de desenvolvimento de competências, métodos e técnicas para a realização de monitoramento e elaboração de diagnósticos. Vem sendo instituído desde 2010 no ambiente do Programa de Pós-Graduação em Ciência da Informação (PPGCI) da Universidade Federal de Pernambuco (UFPE).

Muitas pesquisas desenvolvidas nos últimos anos procuram encontrar padrões que possibilitem agregar valor para superar os obstáculos da linguagem natural[4]. Inúmeros pesquisadores vêm trabalhando na concretização de soluções para a comunicação homem-máquina. Nessa perspectiva, muito se tem discutido acerca de inovações e possibilidades das ferramentas de cunho semântico nos sistemas de recuperação da informação, em geral.

Tecnologias vêm sendo exploradas no contexto da Web Semântica (BERNERS-LEE et al., 2001), com base no projeto e na implementação de padrões de metadados que adicionam aos dados informações relevantes sobre seus contextos, marcando-os semanticamente; e com base em mecanismos de busca que levem em conta esses dados marcados. A ideia de representar documentos em meio digital, através da estruturação de seu conteúdo, ganha ênfase, a partir do momento que propõe novas técnicas para a representação da informação e do conhecimento. Esse enfoque pretende enriquecer a estrutura da informação e agregar componentes semânticos, que podem ser processados de forma

[4]Linguagem natural é aquela falada espontaneamente por um grupo humano ou aquela escrita na obra por seu autor.

automática. Para Martins (2014), os sistemas baseados em conhecimento podem ser vistos como conectores semânticos, recebendo informações de diversas origens e sendo capazes de analisá-las, interpretá-las, identificando a sua relevância e estando aptos a direcionar soluções de acordo com interesses dinâmicos.

Dentre as tecnologias da Web Semântica, encontra-se a busca semântica, permitindo que o conhecimento seja organizado em universos conceituais de acordo com seu significado. Nesse contexto, as palavras deixam de ser simples palavras para converterem-se em conceitos, e os buscadores evoluem de motores de buscas à maquinas de aprendizagem. Esse tipo de organização permite que ferramentas de busca por informação sejam capazes de fazer a seleção e a filtragem dessa informação baseadas na semântica dos termos de busca e dos itens pesquisados.

Quando um mecanismo qualquer atende uma consulta, devolvendo centenas de milhares de páginas e documentos de dez em dez, e hierarquiza essa devolução conforme critérios de popularidade e buscas anteriores do consulente, pode até atender demandas comerciais e de interesse geral, porém, dificilmente, priorizará uma informação científico-tecnológica: a resposta procurada, por ser especializada, aparecerá na 98756ª posição. Recursos como o *Google Scholar* tentam contornar esse problema, mas acabam caindo no outro extremo, de devolver documentos a menos, por não cobrir integralmente todas as possíveis fontes de interesse.

Os "mecanismos quaisquer", a que nos referimos, baseiam-se preferencialmente em presença, frequência e proximidade das palavras inseridas como critério de busca. É um método de força bruta, sem muita inteligência semântica, desconsiderando o significado das palavras e o sentido de sua sequência, ou seja, vê as palavras como figuras vazias e não atribui qualquer significado à sua combinação (por exemplo, não distingue "aumento de juvenis em crimes" de "aumento de crimes em juvenis*")*. Isso o impede de ser científica e tecnologicamente mais eficaz: esse método apenas retorna o que recebe, não infere, não deduze, não relaciona.

Uma primeira alternativa de melhorar esse contexto é o uso de metadados, mas esses são em número limitado; dependem do discernimento do autor, que pode não destacar algum conteúdo do documento que seja exatamente o interesse do consulente; e não estão disponíveis para todos os tipos de informação.

A necessidade de explicitação sobre as técnicas para aumentar a eficiência dos mecanismos de busca, por meio da utilização de linguagens que permitam definir dados e

regras para o raciocínio sobre esses dados, motivou a proposta dessa pesquisa que é a de estabelecer uma rotina computacional capaz de ler, interpretar e marcar textos científicos em português, de modo a possibilitar sua inclusão em buscas semânticas inteligentes.

Para tanto, partiu-se do estudo sobre a busca semântica para ampliar a compreensão da natureza da mesma e para colaborar metodologicamente (na fase experimental) para o desenvolvimento de mecanismos de busca de informação científica e tecnológica, cuja temática é de interesse para a CI.

Esta pesquisa se desenvolveu no âmbito do OtletCI. Para avançar na questão de como extrair informação relevante e de como representá-la para fins de recuperação semântica da informação, a pesquisa teve como objetivo comparar as sentenças identificadas diretamente pelo pesquisador e as sentenças reconstruídas automaticamente a partir de programa implementado em planilha eletrônica. Trabalhando diretamente com resumos de artigos científicos da área de Ciências Agrárias, localizam-se sentenças que envolvam relações de causa-efeito. Por outro lado, empregando recursos computacionais de identificação morfológica e sintática, decompõem-se e se recompõem os textos, igualmente destacando-se sentenças que se presume atendam as mesmas condições (relações de causa-efeito). A relativa convergência dos resultados obtidos por uma e outra via, e o estabelecimento preliminar de rotinas para a versão automatizada são o resultado aqui alcançado, que se encontra ilustrado para alguns dos resumos estudados.

A escolha pelo estudo dos elementos causais presentes nos textos de artigos científicos da área de Ciências Agrárias em detrimento de outros recursos da linguagem se deu pela percepção de que formulações de relações causais marcadas nos textos poderia tornar-se mecanismo central no processo de construção e recuperação do conhecimento científico. Assim, tratou-se de privilegiar a observação da causalidade expressa por meio de verbos. Nesse sentido, partiu-se do pressuposto de que a extração e representação computacional de relações de causa e efeito, pelo maior grau de informação semântica embutida, podem vir a se tornar mais eficazes do que as palavras-chave usualmente extraídas e utilizadas como descritores em outros processos automatizados de representação de documentos.

Dada a importância do conhecimento dentro da economia atual e, por outro lado, a incapacidade humana de trabalhar com tamanha quantidade de informações, infere-se que o desenvolvimento de ferramentas que “amplifique” a capacidade humana de procurar por informações relevantes para o seu negócio, dentro da enorme quantidade e variedade de publicações existentes atualmente, é estratégica para um novo modo de criação de tecnologia e inovação (MARCONDES, 2011). Aprofundar no conhecimento da tecnologia de busca

semântica pode ajudar pesquisadores e gestores de desenvolvimento e inovação a realizar melhores planos, projetos e decisões, permitindo, deste modo, ganhos significativos em vantagem competitiva. Daí a importância de projetos que apliquem essa tecnologia ao problema de busca e demonstrem seus resultados através de experimentos.

2 FUNDAMENTOS TEÓRICOS SOBRE BUSCA SEMÂNTICA

A semântica (derivada do grego *sêmainô*, que quer dizer "significar") é definida, estritamente, *como o estudo do sentido das palavras* (GUIRAUD, 1972). Sentido, na definição do autor, quer dizer *significado* ou *emprego*. Esse campo de estudo envolve níveis diferenciados de tratamento dos problemas do significado, nos quais especificam-se o caráter e o vínculo disciplinar. Para Almeida (2011), significado é um elemento analisável pela estrutura contextual que o circunda.

Na intenção de melhorar métodos e técnicas de organização do acesso aos recursos de conhecimento digital no ambiente de rede, a tentativa deve ser feita para substituir sistemas de organização do conhecimento (SOC), semanticamente fracos, por sistemas com uma semântica mais forte. Isso pode ser alcançado, por meio da criação de novos modelos de SOC ou com a modernização e o desenvolvimento de SOC tradicional.

De acordo com Sosinska-Kalata (2014), a criação de SOC semanticamente fortes exige uma identificação precisa de estruturas conceituais de vários domínios do conhecimento, bem como uma análise precisa de contextos epistemológicos e culturais da criação de conhecimento e de aspectos pragmáticos da aplicação do conhecimento.

A busca semântica tenta compreender a intenção do usuário e o significado contextual dos termos usados na busca, como eles aparecem no espaço de dados pesquisáveis, seja na *web* ou dentro de um sistema fechado, para gerar resultados mais relevantes e precisos (AMANQUI, 2014).

Usando semântica, os sistemas podem compreender "se" palavras ou frases são equivalentes. Ao procurar por referências à palavra "Jaguar", no contexto da indústria automobilística, o sistema pode ignorar as referências aos felinos. Usando semântica, pode-se melhorar a forma pela qual a informação é apresentada, na sua forma mais simples, ao invés de uma pesquisa em uma lista linear de resultados, e esses podem ser agrupados por significados. Assim, uma busca por "Jaguar" pode fornecer documentos agrupados, de acordo com o que eles são (carros, animais ou temas diferentes), todos juntos. No entanto, pode-se ir

mais longe ~~do~~ que isso, usando da semântica, para mesclar informações de todos os documentos relevantes, e fazendo inferências, a partir do conhecimento existente, para criar novos conhecimentos.

A semântica e a sintaxe têm papéis importantes na recuperação da informação, na medida em que permitem ao software identificar a estrutura lexical das frases e o significado dos termos que representam o conteúdo do documento. Na busca semântica, a geração de resultados relevantes envolve, por exemplo, compreensão da intencionalidade do pesquisador e o contexto do termo pesquisado.

A busca semântica tem mostrado um potencial significativo na função de melhorar o desempenho da recuperação da informação. Comparados aos motores de busca tradicionais, que se concentram na frequência de aparecimento das palavras no texto, os motores de busca semântica são mais propensos a compreenderem os significados, escondidos por meio da adição de tags semânticas em textos, a fim de estruturarem e conceituarem os objetos dentro dos documentos. Mangold (2007) define busca semântica como um processo de recuperação de documentos, que aproveita o conhecimento de domínio do contexto semântico dos termos de consulta, aumentando sua precisão e revocação.

A semântica e a sintaxe têm papéis importantes na recuperação da informação, na medida em que permitem ao software identificar a estrutura lexical das frases e o significado dos termos que representam o conteúdo do documento. Na busca semântica, a geração de resultados relevantes envolve, por exemplo, compreensão da intencionalidade do pesquisador e o contexto do termo pesquisado. A análise semântica profunda de documentos possibilita a extração de informação de domínio especializado, em uma área rica em conhecimento, expresso em língua natural (LIMA; NUNES; VIEIRA, 2007).

Segundo Oliveira Neto, Tonin e Pietrich (2010), à medida que se avança no processamento da linguagem natural, é necessário fazer uma interpretação do todo e cessar a análise do significado de suas partes, como ocorre na análise das informações morfológicas (léxicas), sintáticas (regras gramaticais) e semânticas (significados). Em nível pragmático, baseia-se na relação da linguagem com o contexto na qual é utilizada. Em muitos casos, não se pode realizar uma interpretação literal e automatizada dos termos utilizados. Em determinadas circunstâncias, o sentido das palavras que formam uma frase tem que ser interpretado num nível superior, recorrendo ao contexto em que a frase é formulada.

Um motor de busca semântico tenta fazer sentido aos resultados da pesquisa com base no contexto, identificando automaticamente os conceitos que estruturam os textos. A estrutura lexical compreende o conjunto de vocábulos de uma língua e abrange o conhecimento

linguístico, partilhado pela sociedade na qual é falada, possuindo valor diferente de língua para língua. Para Borges (2009), toda língua tem seu próprio recorte e sua própria semântica e essa língua pode ser repleta de regionalismos, metáforas, gírias, linguagem figurada, denotação e conotação. Tudo aquilo que está presente na vida das pessoas possui um nome, que é parte do léxico.

É notório que a informação e o conhecimento são fatores determinantes na criação de riqueza, transformação social e desenvolvimento humano. Com isso, as barreiras geográficas foram transpostas mais facilmente. Entretanto, a barreira linguística é um ponto relevante a ser discutido. Essa barreira imposta pela língua tem se tornado um ponto crítico na transferência de informações e, principalmente, na análise e representação de conteúdos informacionais (HUDON, 1997).

Para tanto, Branco et al. (2012) afirmam que muitos idiomas não estão ainda equipados com a tecnologia básica para a análise de texto, nem com os recursos linguísticos, essenciais para o desenvolvimento dessa tecnologia. Desse modo, é preciso realizar um esforço, em grande escala, para que seja alcançado o objetivo ambicioso de se assegurar tecnologia da linguagem de alta qualidade, para todas as línguas.

A busca semântica tem mostrado um potencial significativo na função de melhorar o desempenho da recuperação da informação. Comparados aos motores de busca tradicionais, que se concentram na frequência de aparecimento das palavras no texto, os motores de busca semântica são mais propensos a compreenderem os significados, escondidos por meio da adição de tags semânticas em textos, a fim de estruturarem e conceituarem os objetos dentro dos documentos.

3 METODOLOGIA

Quanto aos objetivos a serem alcançados neste estudo, realizou-se uma pesquisa empírico-descritiva, uma vez que o estudo pretende descrever os fatos e fenômenos de determinada realidade. Em relação aos procedimentos, a pesquisa pode ser classificada como experimental, já que tem o objetivo de comprovar experimentalmente hipóteses, através da experiência em compreensão computadorizada de linguagem natural.

O experimento foi realizado por meio da leitura dos resumos de um conjunto de artigos científicos relacionados à agronomia, todos escritos na língua portuguesa. A coleta dos artigos científicos foi realizada em outubro de 2015, por meio da *home page* da *Scientific*

Electronic Library Online – SciELO. A seleção dos artigos se deu utilizando-se do seguinte critério de busca: na opção "Assunto: Ciências Agrárias > Títulos correntes> Pesquisa agropecuária Brasileira." Nessas opções, foram feitas duas buscas, uma inserindo a palavra *fertilizante* e outra inserindo a palavra *aminoácidos*. Foram selecionados 30 artigos para o termo *fertilizante* e 30 para o termo *aminoácidos*, publicados nos últimos anos, totalizando 60 artigos.

A importância de delimitar o assunto dos textos em uma área específica – no caso, a área de Ciências Agrárias – foi devido à necessidade de escopo e contextualização. Não há, entretanto, restrições de aplicabilidade da metodologia para documentos textuais oriundos de outras áreas do conhecimento.

As ferramentas tecnológicas utilizadas nesta metodologia necessitavam de documentos submetidos em formato de arquivos de texto simples. Como os documentos coletados na *web* se encontravam em formato PDF, esses documentos precisaram ser convertidos para o formato texto simples. O processo foi feito manualmente, o resumo em português de cada artigo foi copiado e transferido para arquivos em formato TXT (bloco de notas).

Em seguida foi realizado uma análise morfossintática dos 60 resumos. Utilizou-se o *software Palavras***,** uma ferramenta de processamento morfossintático de textos em português, desenvolvida por Bick (1996) em sua tese de doutorado na *Southern University of Denmark,* e que faz parte de um conjunto de ferramentas multilingues chamado VISL[5] (*Virtual Interactive Syntax Learning*). Sua função básica é identificar as classes gramaticais e os elementos sintáticos e semânticos que compõem uma sentença ou texto. O princípio de análise da ferramenta é a gramática de restrições (*constraint grammar* – CG), que faz a análise do texto morfologicamente (lexemas), de grupos de palavras e da composição da oração. Com isso, o programa obtém uma análise nos níveis ortográfico, semântico e sintático. Após a aplicação da identificação do léxico, o programa elimina as ambiguidades encontradas em cada palavra, por meio da aplicação de um conjunto de regras na sentença, identificando e eliminando possibilidades de formas sintáticas inexistentes (MAIA, 2008).

Em seguida os dados gerados pelo *software Palavras* foram transferidos para o *Excel* através do seguinte procedimento: copiando o trecho da página HTML, contendo a análise morfossintática e, com o uso de um programa que automatiza a substituição das marcas HTML de fonte por ";", foi possível construir, para cada resumo, um arquivo no formato .csv, que pode ser carregado no *Excel*, a fim de ser melhor visualizado e manipulado.

[5] Disponível no endereço da Internet: http://visl.sdu.dk/visl/

A próxima etapa foi a análise semântica com o processo manual e o processo automático. Para a análise manual tomamos o corpus construído com os 60 resumos selecionados. Os resumos foram analisados individualmente, localizando e extraindo as relações de causa e efeito presentes em sentenças distintas de cada resumo. No W*ord*, utilizando a ferramenta *cor da fonte*, foram marcados da cor azul a expressão de causa, da cor vermelha o verbo que faz a ligação entre a causa e o efeito, e da cor verde a expressão que designa o efeito. Para facilitar o trabalho de análise, todas as sentenças marcadas foram transferidas para o *Excel*, onde foram separadas por colunas (causa, verbo e efeito).

Para extrair as sentenças de causa e efeito automaticamente foram utilizados os seguintes procedimentos: com os dados transferidos do *software PALAVRAS* para a planilha do *E*xcel, foi possível realizar uma programação para localizar sentenças de causa e efeito automaticamente. A rotina para realizar esta tarefa, criada pelo Prof. Piotr Trzesniak em planilha do *Excel* contendo a saída formatada do Parser *PALAVRAS (* Lema, Classe Gramatical e Marcação Morfológica*)* será apresentada na próxima seção.

4 RESULTADOS

São apresentados a seguir os resultados desta pesquisa cujo o objetivo foi comparar as sentenças identificadas diretamente pelo pesquisador e as sentenças reconstruídas automaticamente a partir de um conjunto de células programadas no *Excel*.

4.1 Procedimentos para extração automática de sentenças de causa e efeito

A principal tarefa de um programa que recupera sentenças de causa e efeito é buscar os verbos que constituem a parte mais importante da sentença de causa e efeito. Uma vez que se encontra verbos na oração, se busca o sujeito e o objeto correspondente a cada verbo.

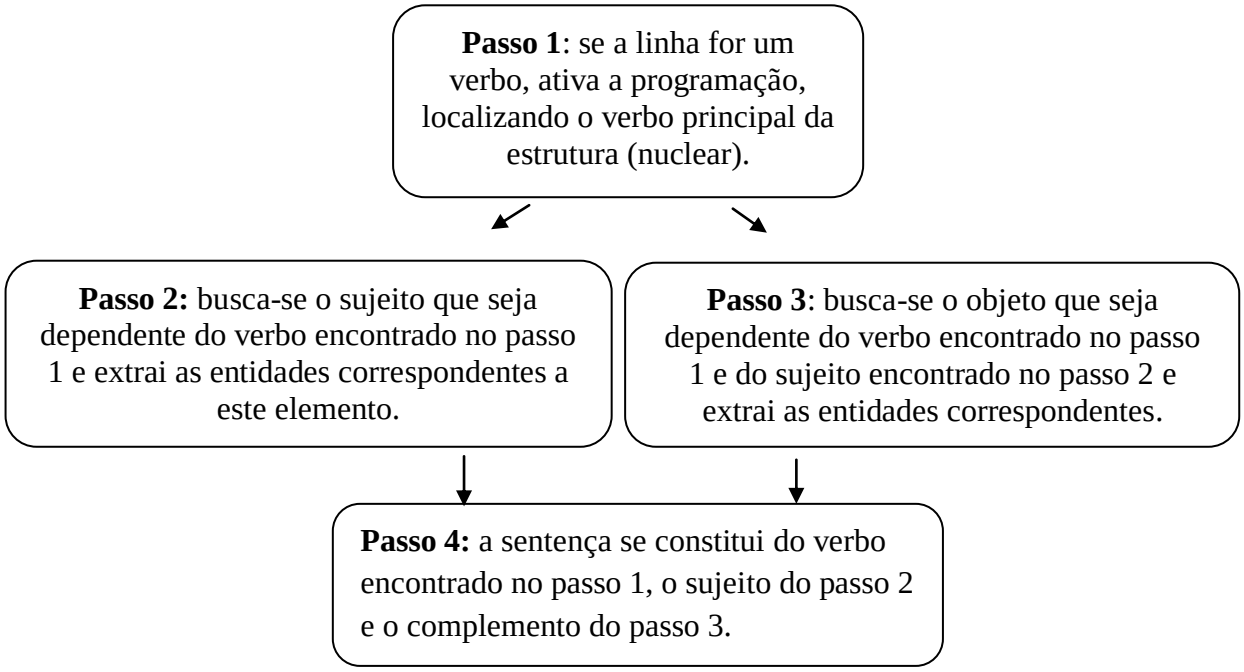
Exemplo de uma sentença de causa e efeito:

Maior dose de fertilizantes provoca redução na porcentagem de plantas quebradas na produtividade de grãos de milho.

Causa (sujeito): maior dose de fertilizantes;
Ação (verbo): provoca;
Objeto: redução na porcentagem de plantas quebradas na produtividade de grãos de Milho;

De posse da lista dos verbos que exprimem relação de causa e efeito localizados na análise manual, a rotina para realizar a extração das sentenças de causa e efeito, criada pelo Prof. Piotr Trzesniak em planilha do excel contendo a saída formatada do Parser *PALAVRAS,* está ilustrada na figura 1 conforme os passos a seguir:

Figura 1: procedimentos das células programadas no *Excel*



Fonte: elaborado pelo autor

A descrição da rotina é dividida em três grupamentos:

GRUPAMENTO SUJEITO + GRUPAMENTO AÇÃO + GRUPAMENTO OBJETO

1- Grupamento de ação

Para localizar uma ação de causa e efeito seis tipos de estrutura são possíveis:

a) Verbo simples (nuclear). Exemplo: causar, provocar
b) Verbos VIR, ESTAR, SER ou TER (VEST) + verbo nuclear VEST + adverbio + verbo nuclear
c) Verbos PODER ou DEVER + verbo nuclear
d) Verbos PODER ou DEVER + SER + verbo nuclear (ex. FAI e a QA podem ser utilizados nas rações de aves em substituição ao milho)
e) Verbos PODER ou DEVER+ adverbio+ SER+ verbo nuclear
f) Verbos PODER ou DEVER+SER+ adverbio + verbo nuclear

2- Grupamento do sujeito

Para localizar o sujeito se parte do princípio que todo substantivo é potencialmente um sujeito. As seguintes atividades são realizadas neste passo:

a) Localizar o substantivo, como todo substantivo é potencialmente um sujeito são destacados todos os substantivos;
b) Partindo de cada substantivo antes do verbo (elementos pré-verbais), se avança no texto agrupando as palavras até:
    (i) Encontrar um sinal de pontuação (vírgula, ponto final, interrogação, exclamação, ponto e vírgula), neste caso a busca é interrompida, e o sujeito foi encontrado contendo as palavras agrupadas;
    (ii) Ao encontrar um verbo, forma-se o sujeito. Complementando com no máximo 15 palavras após o primeiro substantivo até chegar ao verbo.

3- Grupamento objeto

O princípio deste passo é que o objeto é o que vem após o verbo nuclear. As seguintes atividades localizam o objeto da frase:

a) Localiza e reúne tudo que vem após o verbo nuclear até:
    (i) Encontrar um ponto sinal de pontuação (vírgula, ponto final, interrogação, exclamação, ponto e vírgula);
    (ii) Encontrar outro verbo
    (iii) Completar um total de 8 palavras

4- Sentença montada

Após os passos 1, 2 e 3 as sentenças são montadas. Se o núcleo verbal do agrupamento de ação for de causa-efeito, então o programa seleciona a sentença como de causa e efeito.

4.2 Comparação entre as sentenças identificadas manualmente e automaticamente

Os resultados, apresentados nesta seção, permitiram estabelecer algumas considerações sobre as extrações manual e automática em um *corpus* anteriormente processado de forma manual. Apresenta-se considerações de ordem qualitativa e quantitativa para tecer uma comparação entre os processos manual e automático de extração de sentenças

de causa e efeito da área de Ciências Agrárias. Esses documentos foram previamente analisados de forma manual e as sentenças de causa e efeito foram extraídas.

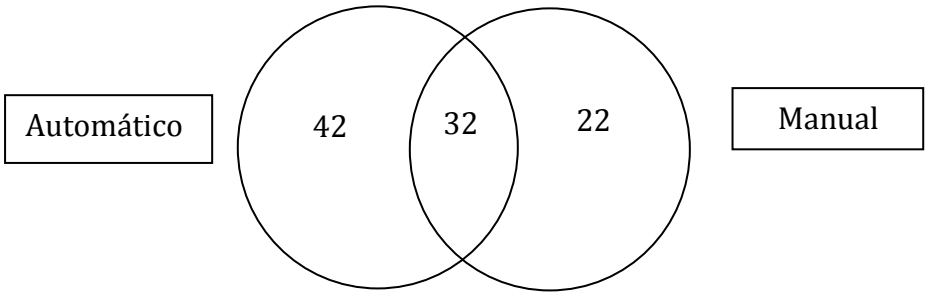
Quadro 1: quantidade de sentenças recuperadas

	Extração manual	Extração automática
Total de sentenças de causa e efeito identificadas	54	74

Fonte: elaborado pelo autor

Como pode ser observado no quadro 1, a extração de sentenças de causa e efeito, por meio do processo automático, retornou um número maior de sentenças. A figura 2 ilustra através do diagrama a relação dos dois resultados (processo manual e processo automático):

Figura 2: diagrama dos resultados quantitativos



Fonte: elaborado pelo autor

Em relação ao conteúdo das sentenças, observou-se que dentro das 74 sentenças identificadas pelo processo automático, 42 não foram identificadas pelo processo manual. Dentre essas 42 sentenças, 8 são sentenças que incluem o advérbio de negação "não". A princípio, na busca manual, não foi levado em consideração esse tipo de sentença, por considerar que se está reportando ao "não acontecimento do efeito esperado", porém, ao se fazer a análise dessas sentenças recuperadas pelo processo automático, foi possível perceber que as sentenças são de causa e efeito, independente se estão acompanhadas do advérbio de negação ligado ao verbo. Alguns exemplos de sentenças de causa-efeito com advérbio de negação são:

- *O aminoácido não alterou as variáveis analisadas.*
- *Aminoácidos essenciais não melhorou com a utilização de a PBc como variável independente.*
- *A redução do teor de proteína bruta da ração não afetou o ganho de peso.*

- *O gesso não proporcionou alterações na produção de biomassa seca do milheto.*
- *A produtividade das culturas não foi alterada pela adição de qualquer dos fosfatos no primeiro ciclo de rotação.*

Em relação as outras sentenças que não foram encontradas no processo manual, é possível afirmar que houve uma desatenção humana do autor dessa Dissertação no momento da busca e identificação das sentenças de causa e efeito.

Ainda em relação ao conteúdo das sentenças, dentro das 54 sentenças localizadas no processo manual, 22 não foram encontradas pelo processo automático. Em relação às sentenças encontradas no processo manual que não foram encontradas no processo automático, possivelmente houve alguma falha que poderá ser revista em um esforço futuro de continuidade desta pesquisa. Entretanto, podemos identificar alguns problemas específicos do processo, que podem ter influenciado na não localização das sentenças:

- Falhas do processador *PALAVRAS*, na identificação errônea de palavras ou sinais especiais de formatação (ex. números seguidos por um ponto, números romanos, títulos de seções do texto sem pontuação final, abreviaturas, sinais gráficos como $, &, etc.);
- Sujeitos e objetos muito longos nas sentenças (mais de 8 palavras);
- A vírgula, às vezes, separa sentenças e, às vezes, separa elementos de uma sentença;

Vale ressaltar, também, que um olhar mais atento e minucioso permitiu verificar que a identificação manual oferece tratamento melhor para a exploração das sentenças de causa e efeito presentes nas estruturas dos resumos. Porém, mesmo considerando os problemas apontados e a eficácia qualitativa, se compararmos as performances, levando em conta a velocidade relativa dos processos de extração e o grande percentual de sentenças extraídas corretamente pelo programa, consideramos que o pressuposto apresentado na introdução **(**estabelecer uma rotina computacional capaz de ler, interpretar e marcar textos científicos em português, de modo a possibilitar sua inclusão em buscas semânticas inteligentes**)** se verificou.

O estudo apresentou, por meio de considerações de ordem qualitativa e quantitativa, uma comparação entre os processos manual e automático de extração de sentenças de causa e efeito. Para essa avaliação, tomamos o corpus construído com 60 resumos de artigos científicos da área de Ciências Agrárias. Esses documentos foram previamente analisados de forma manual e as sentenças de causa e efeito foram extraídas.

5 CONSIDERAÇÕES FINAIS

Inicialmente, as questões que instigaram a pesquisa indagavam sobre a possibilidade de enumerar e identificar vários tipos de relações semânticas no corpus analisado, evidenciando fatos que estão expressos nos textos. Considerando-se que existem vários tipos de relações semânticas, podemos inferir a impossibilidade de estudá-las a priori, devido ao curto período de tempo. Mesmo com esta limitação, neste trabalho, chegou-se à seguinte sistematização: trabalhando diretamente com resumos da área de Ciências Agrárias, localizam-se sentenças que envolvam relações de causa-efeito. Por outro lado, empregando recursos computacionais de identificação morfológica e sintática, decompõem-se e se recompõem os textos, igualmente destacando-se sentenças que se presume atendam as mesmas condições (relações de causa-efeito).

Embora se tenha constituído a partir de algumas contribuições, o presente trabalho pode ser considerado útil, na medida em que abre caminho para aperfeiçoamento constante de metodologias de extração de relações semânticas de causa e efeito. A principal contribuição destaca-se com o próprio método proposto para extração de relacionamentos do tipo causa e efeito, tanto nos textos relacionados com a agronomia quanto em textos relacionados com outros assuntos. Foi gerado um modelo inicial de busca pelos relacionamentos de causa e efeito, que demonstra a grande utilidade deste artefato na geração de novas hipóteses de pesquisa sobre o assunto "fertilizantes e aminoácidos" por pesquisadores especialistas das Ciências Agrárias.

A possibilidade de usar técnicas automáticas acelera o processo de criação e extração de relações de causa e efeito e pode ser usada como alternativa ao processo custoso de identificação manual de informações semânticas. Dessa forma, busca-se superar o gargalo existente devido a grande demanda por dados semânticos, e a escassez de tal conhecimento e de mão de obra qualificada e disponível para gerá-lo em tempo hábil.

Afinal, conclui-se que mais importante que propor uma estrutura de relações de causa e efeito para a construção de sistemas de busca, o que podemos apontar como o resultado mais expressivo da presente pesquisa é o estabelecimento preliminar de rotinas para a versão automatizada. Espera-se, com esta pesquisa então, viabilizar a identificação de indícios de novas descobertas científicas no âmbito do Observatório Temático e Laboratório – Ensino, Tecnologia, Ciência e Informação (OtletCI).

REFERÊNCIAS

ALMEIDA, C. C. de. **Elementos de Linguística e Semiologia na Organização da Informação**. 1.ed. São Paulo: Cultura Acadêmica, 2011.

AMANQUI, F.K. M.. **Uma arquitetura para sistemas de busca semântica para recuperação de informações em repositórios de biodiversidade**.2014.82f. Dissertação (Mestrado- Programa de Pós-Graduação em Ciências da Computação e Matemática Computacional). Instituto deciências Matemáticas e de Cpmpitação, Universidade de São Paulo, 201. Disponível em:<http://www.teses.usp.br/teses/disponiveis/55/55134/tde-03072014-150009/en.php. Acesso em: jan 2016.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic *web***. Scientific American,** San Francisco, EUA, p. 28-37, 2001.

BORGES, G. S. B. **Indexação automática de documentos textuais:** critérios essenciais**.** 2009. 111 f. Dissertação (Mestrado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2009.

BRANCO, A. et al. **A Língua Portuguesa na era digital / The Portuguese Language in the Digital Age**. 1 ed. Berlin: Springer-Verlag, 2012. 85p.

HUDON, M. Multilingual thesaurus construction: integrating the views of different cultures in one gateway to knowledge and concepts. **Knowledge Organization,** Würzburg, v. 24, n. 2, p. 84-91, 1997.

LIMA, V. L. S.; NUNES, M. das G. V.; VIEIRA, R. Desafios do processamento de línguas naturais. In: SEMISH-SEMINÁRIO INTEGRADO DE SOFTWARE E HARDWARE, 34., 2007, Rio de Janeiro**. Anais...** Rio de Janeiro: SBC, 2007. p. 2202-2216.

MAIA, Luis Claudio Gomes. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. 2008. 158 f. Tese (Doutorado) – Curso de Doutorado em Ciência da Informação, Escola de Ciência da Informação, Universidade Federal de Minas Gerais - UFMG, Belo Horizonte, 2008

MANGOLD, C. A survey and classification of semantic search approaches. **International Journal of Metadata, Semantics and Ontologies**, v. 2, n. 1, p. 23-34, 2007. Disponível em: <ftp://inf.informatik.uni-stuttgart.de/pub/library/ncstrl.ustuttgart_fi/ART-2007-09/ART-2007-09.pdf.>. Acesso em: 12 nov. 2015.

MARCONDES, C. H. O papel das relações semânticas na Organização e Representação do Conhecimento em ambientes digitais. In: SILVA, F. C. C. da; SALES, R. de. (Org.). **Cenários da organização do conhecimento:** linguagens documentárias em cena. Brasília: Thesaurus, 2011, p. 129-168.

OLIVEIRA NETO, J. M.; TONIN, S. D.; PIETRICH, S. S. Processamento de linguagem natural e suas aplicações computacionais. In: ESCOLA REGIONAL DE INFORMÁTICA, 2., 2010, Manaus. **Anais...** Manaus: INPA, 2010. p. 1-10. Disponível em: < https://www.inpa.gov.br/erin2010/Artigo/Artigo9.pdf>. Acesso em: 05 out. 2015.

SOSISKA-KALATA, B. Semantization and standardization – cooperative or conflicting trends in knowledge organization? **Knowledge organization,** Würzburg, v.14, n. 2, p. 580, 2014.