Capítulo

09

On-Semse: Um *Framework* para Busca Semântica Baseado em Ontologia[1]

Daniel Albuquerque Carvalho
Graduado em Ciência da Computação (IFCE). E-mail: daniel.albuquerque@mail.com.

Cynthia Pinheiro Santiago
Mestra em Ciência da Computação (UFC). Professora EBTT IFCE, *campus* Tianguá. E-mail: cynthia.pinheiro@ifce.edu.br. ORCID: 0000-0003-4013-4751.

Francisca Raquel de Vasconcelos Silveira
Doutora em Informática Aplicada (UNIFOR). Professora EBTT IFCE, *campus* Tianguá. E-mail: raquel_silveira@ifce.edu.br. ORCID: 0000-0001-7445-605X.

INTRODUÇÃO

A quantidade de dados disponíveis na Internet está crescendo largamente devido a diversos fatores, como usuários, sistemas, sensores e aplicações que se utilizam da *web*. Milhões de transações ocorrem diariamente e grandes empresas - Meta, X e *Google*, por exemplo - adicionam e compartilham rotineiramente uma vasta quantidade de dados. Essa grande quantidade proporciona diversos desafios, os quais alavancam o desenvolvimento de técnicas que ajudem a extrair os dados de forma precisa e a alcançar uma busca semântica (Sayed; Muqrishi, 2017).

Extrair informações precisas da *web* é uma tarefa complexa, pois a maioria dos mecanismos de busca atuais adota sistemas de indexação baseados em palavras-chave, cujo resultado nem sempre atende às expectativas e necessidades dos usuários. A função básica de um mecanismo de busca é permitir que o usuário recupere documentos da *web* com base nas consultas realizadas por eles. Por sua vez, a principal limitação dos mecanismos de busca tradicionais reside na incapacidade de compreender o significado das palavras-chave e das expressões usadas pelo usuário na pesquisa (Grover; Kochar, 2019).

A falta de mecanismos capazes de captar o significado do conteúdo das páginas da *web* criou uma forte demanda de serviços de busca semântica, que passou a ser vista como uma alternativa factível para um melhor tratamento dos problemas relacionados à manipulação de informação na Internet. Portanto, a

[1] Este capítulo corresponde a uma versão adaptada do Trabalho de Conclusão de Curso (TCC) , apresentado ao Curso de Bacharelado em Ciências da Computação do Instituto Federal do Ceará, *Campus* Tianguá, por Daniel Albuquerque Carvalho.

associação dos dados compartilhados na *web* com informação semântica estruturada tem se mostrado uma solução promissora para fornecer dados conceitualmente similares, o que pode ser potencializado por meio do uso de ontologias, os quais visam atribuir sentido e significado ao conteúdo dos dados, atuando como ferramenta de representação do conhecimento (Berners-Lee *et al.*, 2001; Chishman *et al.*, 2006). Um mecanismo de busca semântica eficaz tenta analisar a intenção do usuário de pesquisar conteúdo e o significado esperado de determinado conteúdo de pesquisa. Ao vincular os dados, isso ajudará os usuários a encontrar e usar as informações com mais facilidade.

Desta forma, este trabalho tem como objetivo apresentar um *framework* para a solução do problema de busca semântica, estruturada na ontologia linguística WordNet. Para validação, foi implementado um protótipo, que suporta a ligação semântica de conceitos de propósito geral, sem um domínio específico, passível de ser reusado em diferentes sistemas que se utilizem de busca semântica.

REFERENCIAL TEÓRICO

Web Semântica

Dados são publicados na *web* por diferentes pessoas e estão armazenados em diferentes repositórios espalhados pelo mundo. Para facilitar a construção de um banco global com estes dados, é preciso que se estabeleça uma representação e uma forma padrão de conexão (Laufer, 2015).

A maneira na qual a maioria dos documentos *web* estão organizados na Internet faz com que apenas seres humanos possam entender seu significado, impossibilitando às máquinas de acessá-los e compreendê-los (Berners-Lee *et al.*, 2001). Nesse sentido, surgiu a *Web* Semântica, utilizada para expressar informações de forma precisa e passível de interpretação por máquinas, permitindo assim que agentes de *software* possam processar, compartilhar, reusar, além de compreender o sentido dos dados (Isotani; Bittencourt, 2015).

Para implementar a *Web* Semântica, é necessário um modelo de dados que permita que as informações sejam distribuídas pela *web*. O *Resource Description Framework* (RDF) é uma linguagem utilizada para representar tais modelos, usando declarações expressas na forma de triplas - compostas por um sujeito, um predicado e um objeto - que representam uma conexão entre dois conceitos (Segaran *et al.*, 2009).

O RDF fornece uma maneira simples de representar dados distribuídos. No entanto, para utilizar essa representação, é preciso um mecanismo que acesse os dados. Neste contexto, é possível utilizar a linguagem de consulta SPARQL. Pode-se fazer uma analogia entre SPARQL e a linguagem SQL, de consulta a bancos de dados relacionais: a diferença é que SPARQL tem uma sintaxe adequada a consultas a dados representados como um conjunto de triplas RDF (Laufer, 2015).

Ontologias

Ontologias são métodos flexíveis para representar o conhecimento. Na Computação, passou a significar os tipos de conceitos que podem ser descritos em um sistema ou contexto. Uma ontologia fornece os meios para classificar conceitos do mundo real, dando nomes e rótulos e definindo o tipo de propriedades e relacionamentos que podem ser a eles atribuídos. Uma ontologia, portanto, fornece um vocabulário de termos para uso em um domínio específico (Passin, 2004).

Para construir uma ontologia, é necessário definir formalmente quais propriedades estão associadas a quais classes. Nos modelos semânticos, usa-se o termo "classe" para descrever grupos de entidades. Além disso, os modelos semânticos são orientados a propriedades, ou seja, entidades semânticas são consideradas membros de uma classe por causa de suas propriedades (Segaran *et al.*, 2009).

Uma das principais funções de uma ontologia é definir um conjunto de classes que, juntas, cobrem um domínio de interesse. Todas devem ser construídas com vocabulários e regras de construção conhecidas, possibilitando o reuso e facilitando a construção de novas ontologias (Passin, 2004).

Busca Semântica

Com uma grande variedade de fontes, organizações e estilos de informação, a *web* tornou-se o maior banco de dados do mundo. Neste contexto, a busca é uma ferramenta que permite que organizações e indivíduos explorem enormes quantidades de informações.

A ideia inicial da *web* foi a de prover um meio de navegação entre documentos dispostos em uma estrutura de hipertexto. O conteúdo das páginas é visto pelas máquinas de uma forma apenas sintática. Sendo assim, para conseguir identificar de forma individual, legível por máquinas, cada um dos dados agrupados nas páginas *web*, é necessário que se incluam informações extras sobre esses dados, os quais serão posteriormente consumidas por máquinas. Os principais motores de

busca têm uma atuação limitada: existem diversos dados compartilhados, e são necessários algoritmos para extrair esses dados a partir de informações geradas por seres humanos (Laufer, 2015).

Nesse sentido, o processo de recuperação de informação parte da comparação de dois elementos linguísticos: a representação dos documentos e a representação da expressão de busca. As ontologias inserem-se no processo de recuperação de informação visando prover um maior nível semântico de tais representações (Ferneda; Dias, 2017). Ontologias também podem ser usadas para melhorar a precisão das pesquisas na *web*, uma vez que o motor de busca pode procurar apenas as páginas que se referem a um conceito preciso, em vez de todas aquelas que usam palavras-chave ambíguas (Berners-Lee *et al.*, 2001).

A recuperação de informação baseada em ontologia já é um campo de pesquisa consolidado na Ciência da Computação e existe uma diversidade de trabalhos que abordam propostas para a utilização de ontologias no processo de recuperação de informação (Ferneda; Dias, 2017).

WordNet

No âmbito do Processamento de Linguagem Natural, um dos tipos de ontologias utilizadas são as ontologias linguísticas. A WordNet é uma ontologia linguística para o sistema léxico inglês, sendo um recurso léxico amplamente utilizado no processamento de linguagem natural e na recuperação de informações. Mais recentemente, também foi adotado na comunidade de pesquisa da *Web* Semântica (Van Assem *et al.*, 2006).

A WordNet pode ser entendida como um tesauro ou dicionário de sinônimos. No entanto, interliga não apenas formas de palavras, mas os sentidos específicos de palavras e, além disso, rotula as relações semânticas entre as mesmas. Na WordNet, conjuntos de sinônimos (ou *synsets*) são considerados como unidades básicas de organização. Outros conceitos básicos são *WordSense* e *Word*. *Word* são as unidades lexicais básicas, enquanto um *WordSense* é um sentido específico no qual uma palavra é usada. Por exemplo, *computer* como *computing machine* ou *computer* como *calculator* (Wu; Yuan, 2019).

A WordNet é dividida em classes (*synsets*) que agrupam os sentidos de palavras (*WordSense*) com um significado de sinônimos, e esses a seu correspondente (*Word*). O conceito central da WordNet é o *synset,* o qual agrupa palavras com seus sinônimos, como *car*, *auto*, *automobile*, *machine*, *motorcar*. Cada *WordSense* tem exatamente uma palavra que o representa lexicalmente e uma palavra pode estar relacionada a um ou mais *WordSenses* (Van Assem *et al.*, 2006).

Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é um ramo que processa e analisa dados da linguagem humana, tratando essencialmente de três aspectos: (i) som: fonologia; (ii) estrutura: morfologia e sintaxe; (iii) significado: semântica e pragmática (Vasiliev, 2020). Enquanto a sintaxe corresponde ao estudo de como as palavras agrupam-se para formar estruturas no nível de sentenças, a semântica está relacionada ao significado de uma palavra e do conjunto resultante delas. Conforme Gonzalez e Lima (2003) afirmam, o processamento semântico é considerado um dos maiores desafios do PLN .

Pode-se dizer que o PLN é um campo vasto uma vez que envolve diversas disciplinas do conhecimento. Um dos principais exemplos é o mecanismo de busca que utiliza, como parâmetros, palavras ou expressões e pode fornecer resultados ainda mais significativos se levarmos em consideração o significado do texto (e não apenas a sintaxe) em linguagem natural (Lane *et al.*, 2019).

Atualmente, existem algoritmos que analisam linguagens cuja semântica e as regras gramaticais são conhecidas, sendo possível construir aplicações que podem compreender as expressões da linguagem natural. Para conseguir fazer com que máquinas compreendam dados textuais, ao contrário dos humanos, as máquinas usam representações de palavras baseadas em vetores, o que permite fazer operações matemáticas em unidades de linguagem natural como palavras, frases e documentos (Vasiliev, 2020).

No entanto, antes de realizar a análise dos dados propriamente dita, é necessário realizar um pré-processamento, o qual prepara os dados de texto bruto para processamento posterior, já que frequentemente estes são incompletos, inconsistentes ou com ruídos (Kulkarni; Shivananda, 2021). Existem vários métodos e técnicas para pré-processar dados textuais antes da análise de dados. Entre as técnicas estão as seguintes:

- **Tokenização**: consiste em dividir o texto em segmentos significativos, chamados *tokens*. Esta é a primeira ação que aplicações de PLN normalmente executam em um texto: segmentá-lo em palavras, números ou sinais de pontuação (Vasiliev, 2020).
- **Lematização**: processo algorítmico que determina o lema, que é a forma básica de um *token* após a deflexão de uma palavra (Sumit, 2019). Por exemplo, o lema de *running* é *run*.
- **N-gramas**: coleção de N *tokens* de palavras de forma que estes sejam contíguos e ocorram em uma sequência (Sarkar, 2019). Uma palavra representa um único *token*, geralmente conhecido como unigrama ou 1-grama (Srinivasa-Desikan, 2018).

Sentence-BERT (SBERT)

O Sentence-BERT (SBERT) é uma modificação da rede pré-treinada BERT e faz uso de estruturas de rede siamesas e triplas para derivar *embeddings* (incorporações) de sentenças semanticamente semelhantes que podem ser comparadas pela similaridade de cosseno (Devlin *et al.*, 2018). Ou seja, consegue fazer vetorização de texto baseada em redes neurais, tornando os dados compreensíveis para as máquinas (Rothman, 2021).

Diversos modelos pré-treinados do SBERT são disponibilizados livremente, permitindo que sejam usados para novas tarefas que incluem comparação de semelhança semântica em larga escala, agrupamento e recuperação de informações por meio de pesquisa semântica, sem a necessidade de treinar uma rede neural (Reimers; Gurevych, 2019). O SBERT é utilizado para tarefas comuns de similaridade semântica. Assim, é capaz comparar a semelhança entre duas *embeddings* de sentenças, através dos modelos pré-treinados, retornando uma taxa de similaridade entre 0 e 1.

MATERIAIS E MÉTODOS

Este trabalho utiliza como método de pesquisa a *Design Science Research* (DSR), uma abordagem que tem duplo objetivo: primeiro, desenvolver um artefato para resolver um problema prático num contexto específico; e segundo, gerar novos conhecimentos técnicos e científicos. As principais atividades da DSR são: definir um problema, sugerir formas de contorná-lo, desenvolver uma solução e avaliá-la, concluindo assim a pesquisa (Pimentel *et al.*, 2020).

Segundo Vaishnavi e Kuechler (2007), os tipos de artefatos na DSR são: (i) Constructo: vocabulário conceitual de um domínio; (ii) Modelo: proposições que expressam relacionamentos entre os constructos; (iii) *Framework*: guia conceitual ou real que serve como suporte; (iv) Arquitetura: sistemas de estrutura de alto nível; (v) Princípio de projeto: princípios-chave e conceitos para guiar o projeto; (vi) Método: passos para executar tarefas – "como fazer"; (vii) Instanciação: implementações em ambientes que operacionalizam constructos, modelos, métodos e outros artefatos abstratos; e (viii) Teorias de projeto: conjunto prescritivo de instruções sobre como fazer algo para alcançar um determinado objetivo.

Por outro lado, Peffers *et al*. (2007) dividem a atividade de avaliação de um artefato em duas atividades: demonstração e avaliação. A demonstração indica se o artefato funciona de maneira viável para "resolver uma ou mais instâncias do problema", ou seja, para atingir seu propósito em pelo menos um contexto. Já a avaliação propriamente dita é mais formal, verifica "quão bem o

artefato suporta uma solução para o problema" e inclui a coleta de medidas objetivas de desempenho.

Neste trabalho, é desenvolvido um artefato do tipo *framework* para a solução do problema de busca semântica. Para tanto, inicialmente, um estudo sobre os mecanismos de busca semântica foi realizado, com a finalidade de entender quais ferramentas seriam as mais utilizadas para consulta de ontologias, assim como no uso de PLN em buscas semânticas. Este *framework* considerou as seguintes tecnologias: processamento de linguagem natural, busca por expressões semanticamente equivalentes na ontologia linguística WordNet e cálculo de similaridade semântica entre as expressões utilizando-se o SBERT.

Para a demonstração e avaliação deste *framework* foi implementado um protótipo na forma de um sistema *web*, no qual o usuário fornece uma sentença em inglês e obtém, como resultado da busca, sentenças semanticamente semelhantes. Para tanto, foram utilizadas as seguintes ferramentas de desenvolvimento: (i) a linguagem de programação *Python*, em conjunto com o *framework* Django para criação de um sistema *web* (Santiago *et al.*, 2020); (ii) *spaCy*, uma biblioteca em Python para as técnicas de pré-processamento de PLN; (iii) *RDFLib*, uma biblioteca em Python para analisar, armazenar, consultar e serializar triplas RDF, através de consultas em SPARQL; (iv) a ontologia linguística *WordNet*; (v) *Intertools,* uma biblioteca em Python para utilização das funções de permutação e produto cartesiano; e, por fim, (vi) o modelo multi-qa-MiniLM-L6-cos-v1, da rede neural pré-treinada SBERT para verificação de frases semanticamente semelhantes.

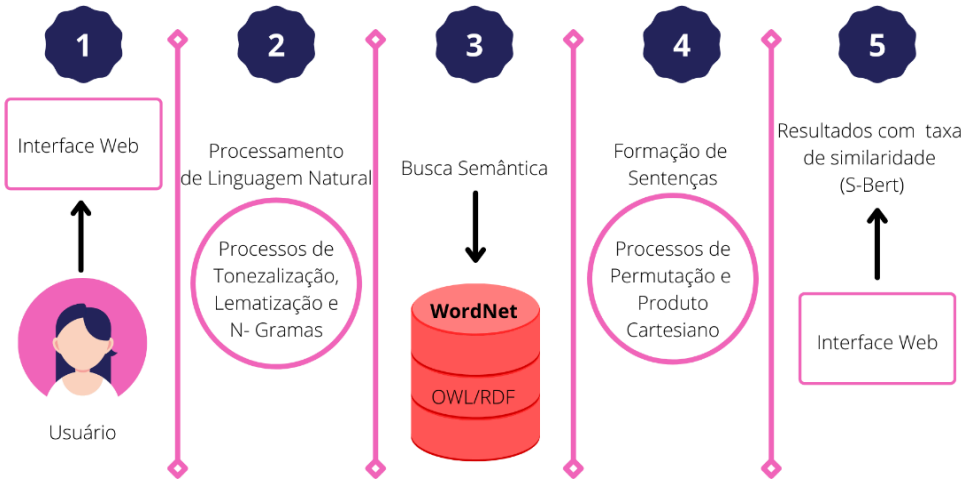
FRAMEWORK ON-SEMSE

O *framework* proposto como mecanismo de busca semântica para expressões em inglês, intitulado On-SemSe (*Ontology-based Semantic Search*), é estruturado na ontologia linguística WordNet e no uso de Processamento de Linguagem Natural.

Conforme representado na Figura 1, a solução desenvolvida neste trabalho é projetada com base em cinco diferentes etapas: (1) interação com interface; (2) processamento de linguagem natural; (3) busca semântica; (4) formação de sentenças; e (5) visualização de resultados obtidos ordenados por taxa de similaridade semântica.

As etapas (1) e (5) correspondem, respectivamente, à entrada de dados - através da informação de uma sentença na língua inglesa a ser pesquisada - e à saída de dados, que corresponde à listagem das sentenças semanticamente equivalentes à sentença original.

Figura 1 – Representação visual do *framework* On-SemSe



Fonte: Elaboração própria, 2024.

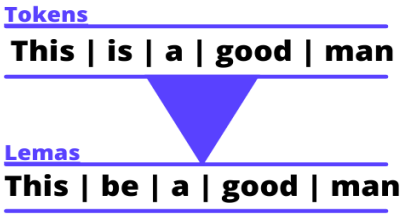
As demais etapas serão detalhadas nas seções a seguir. Para exemplificar a heurística utilizada em cada etapa, será utilizado um cenário ilustrativo com a sentença "*This is a good man*".

Etapa 2: Processamento de Linguagem Natural

Logo após a etapa de entrada de dados, na qual uma expressão na língua inglesa é informada, são utilizadas técnicas de PLN, como tokenização, lematização e formação de n-gramas. Estas técnicas são utilizadas com a finalidade de aumentar a relevância dos resultados da busca, gerando uma lista de palavras que será encaminhada para a etapa seguinte.

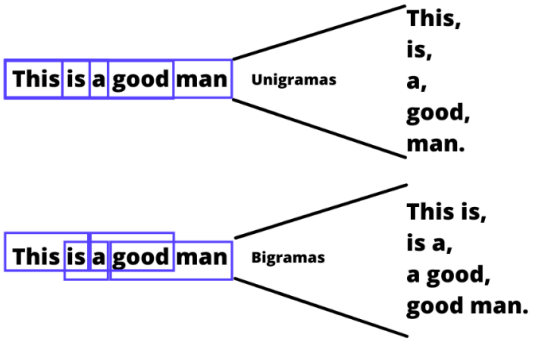
Nesta etapa, a sentença fornecida será pré-processada para separação em *tokens* e, posteriormente, cada *token* passará por um processo de lematização, como descrito na Figura 2. Por fim, na Figura 3, é mostrado o processo de segmentação em unigramas e bigramas, gerando listas de n-gramas. Na sequência, será realizada a busca semântica dos n-gramas na WordNet.

Figura 2 - Tokenização e Lematização



Fonte: Elaboração própria, 2024.

Figura 3 - Segmentação em unigramas e bigramas



Fonte: Elaboração própria, 2024.

Etapa 3: Busca Semântica na WordNet

Nesta etapa, recebe-se a lista de n-gramas para realização de consultas na WordNet. Após consulta à base WordNet, cada item dessa lista retorna uma lista de expressões semanticamente semelhantes. Assim, é formado um conjunto de *synsets* para os n-gramas obtidos na etapa anterior, conforme ilustra a Figura 4. Em seguida, ocorre a próxima etapa que contempla a formação de sentenças.

Figura 4 – Conjunto de *synsets*

```
"This" : ["this"]
"be" : ["be", "equal", "embody", "exist"]
"a" : ["a"]
"good" : ["good", "well", "fine"]
"man" : ["man"]
"This be" : ["this be"]
"be a" : ["be a"]
"a good" : ["a good"]
"good man" : ["good man"]
```

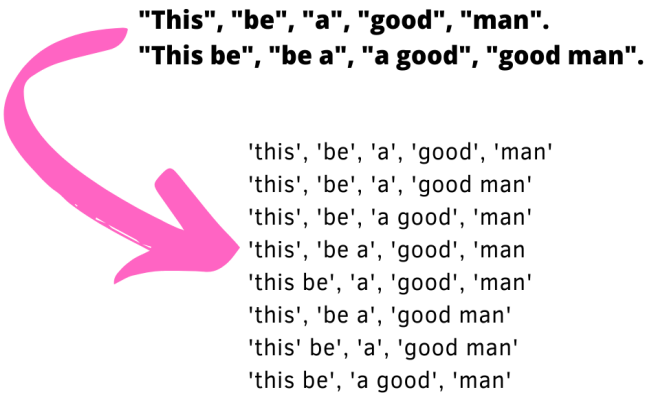
Fonte: Elaboração própria, 2024.

Etapa 4: Formação de Sentenças

Com base nos n-gramas gerados pela etapa 2 e os resultados da busca semântica obtidos na etapa 3, são formadas diferentes sentenças com a utilização de duas funções matemáticas: (1) a função de permutação, que realiza todas as combinações possíveis entre os elementos das listas de n-gramas, e (2) a função de produto cartesiano, que efetua as combinações de um elemento com ele mesmo.

Na Figura 5 está descrito como ocorre a função de permutação, que realiza as combinações entre os n-gramas para formar as sentenças-base. Por fim, com as sentenças-base construídas, é feita a função de produto cartesiano, na qual é realizada a combinação entre os *synsets* obtidos. Dessa forma, é produzida uma lista de sentenças semanticamente equivalentes (Figura 6).

Figura 5 - Permutações de n-gramas para formação de sentenças



Fonte: Elaboração própria, 2024.

Figura 6 - Sentenças semanticamente equivalentes

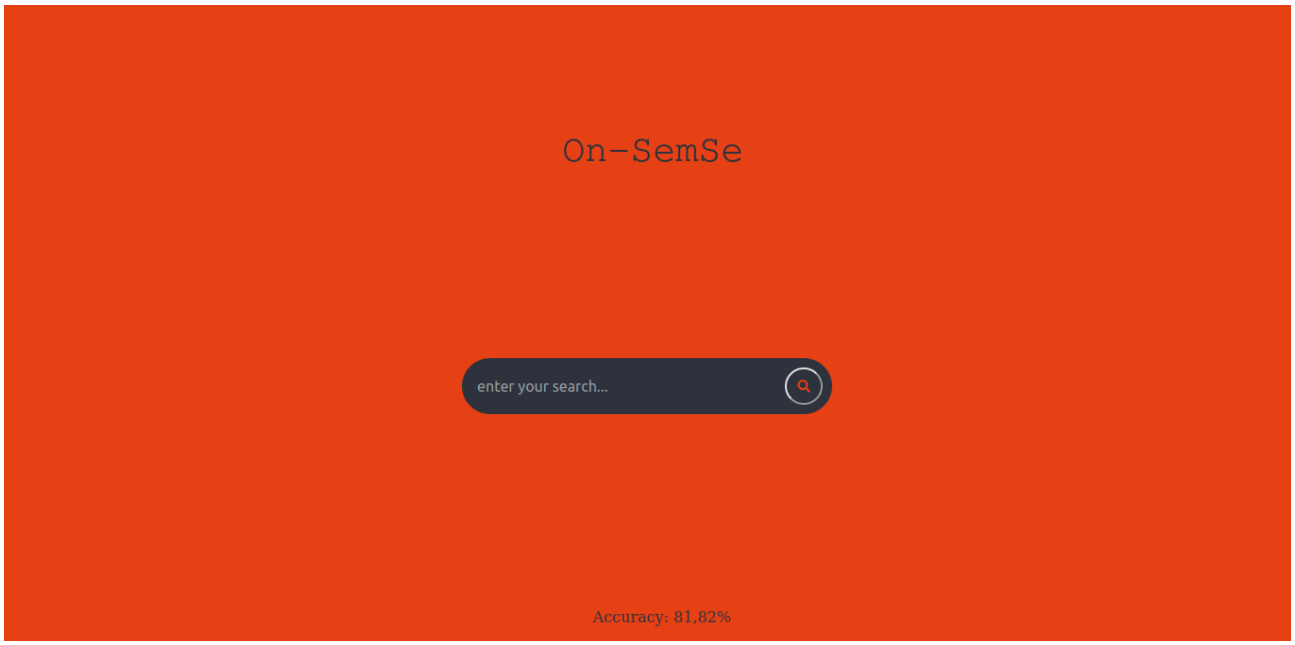
"This is a good man"
"This is a well man"
"This is a fine man"
"This be a good man"
"This embody a good man"
"This equal a good man"
.
.
.
"This exist a good man"

Fonte: Elaboração própria, 2024.

AVALIAÇÃO DO FRAMEWORK ON-SEMSE

Para a demonstração e avaliação do *framework* proposto, foi implementado um protótipo de tipo prova de conceito, na forma de um sistema *web,* com uma interface interativa, na qual o usuário fornece uma sentença em inglês e obtém, como resultado da busca, sentenças semanticamente semelhantes, conforme as Figuras 7 e 8.

Figura 7 – Interface de busca



Fonte: Elaboração própria, 2024.

Como resultado final da pesquisa, é retornada uma lista de sentenças em ordem decrescente de taxa de similaridade semântica, na qual a taxa de similaridade de cada item é calculada de acordo com a similaridade de cosseno entre os *embeddings* das sentenças, gerados utilizando o SBERT. A taxa de similaridade retorna valores entre 0 e 1, conforme mostra a Figura 8. Nessa lista, os itens que mais se assemelham à sentença inicial - com taxa de similaridade semântica próxima a 1 - são retornados primeiro, seguidos pelos itens menos semelhantes.

Este protótipo foi submetido a uma avaliação preliminar de tipo analítica, que busca avaliar o artefato e sua arquitetura, bem como sua maneira de interagir com o ambiente externo. O objetivo principal desta avaliação é verificar o desempenho do artefato (Dresch *et al*., 2015).

Para viabilizar a etapa de avaliação, ao lado de cada item foi incluída uma opção de voto na qual o usuário pode marcar se, em sua opinião, o item é similar ou não à sentença original (Figura 8). Os dados desta votação permitem calcular a acurácia da busca semântica neste protótipo, ou seja, a proximidade do resultado obtido com o resultado esperado. Neste caso, o cálculo do percentual de acurácia é obtido pela razão entre a (quantidade de respostas positivas / quantidade total de respostas) x 100.

Figura 8 – Resultados em ordem decrescente de taxa de similaridade

Fonte: Elaboração própria, 2024.

No período de outubro/2021 a fevereiro/2022, estudantes voluntários livremente informavam sentenças em inglês e obtinham os resultados de sua busca, avaliando os resultados retornados. Nesta ocasião, o protótipo chegou a 81,82% de acurácia, indicando que a maioria das sentenças foi avaliada pelos estudantes como semanticamente equivalente à sentença original, validando esta prova de conceito.

CONSIDERAÇÕES FINAIS

Apresentamos neste capítulo, como principal contribuição, o *framework* On-SemSe que visa propor uma solução para o problema de buscas semânticas. Este *framework* está estruturado na ontologia linguística WordNet e utiliza-se de técnicas de PLN. A finalidade deste artefato, desenvolvido com o uso do método DSR, é a de melhorar as formas atuais de busca, que tradicionalmente utilizam-se de palavras-chaves para a pesquisa ou realizam buscas semânticas apenas em um domínio específico.

Para avaliação da viabilidade técnica deste artefato, foi construído um protótipo, uma prova de conceito, que se constituiu de um sistema *web* no qual o usuário realiza buscas, obtém as sentenças semanticamente equivalentes e analisa os resultados de sua busca.

Como trabalhos futuros, está prevista a inclusão de ontologias linguísticas em português, a aplicação de técnicas mais especializadas de PLN e a utilização dos votos dos usuários na geração

de novas sentenças. Espera-se assim aprimorar o *framework* para melhorar a acurácia dos resultados da busca.

REFERÊNCIAS

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, New York, p. 29-37, 2001.

CHISHMAN, R.; ALVES, I. M. R.; BERTOLDI, A. **O Conhecimento Semântico Representado em Ontologias Aplicadas à Busca e Extração de Informações na Web**. 2006. Disponível em: http://www.filologia.org.br/ileel/artigos/artigo_285.pdf. Acesso em: 09 jun. 2024.

DEVLIN, J.; CHANG, M.; LEE, K.; TOUTANOVA, K. BERT: pre-training of deep bidirectional transformers for language understanding. **ArXiv**, 2018. http://dx.doi.org/10.48550/ARXIV.1810.04805.

DRESCH, A.; ANTUNES JUNIOR, J. A. V. **Design Science Research**: Método de Pesquisa para Avanço da Ciência e Tecnologia. São Paulo: Bookman, 2015. 177 p.

FERNEDA, E.; DIAS, G. A. OntoSmart: um modelo de recuperação de informação baseado em ontologia. **Perspectivas em Ciência da Informação,** v. 22, n. 2, p.170-187, jun. 2017. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/1981-5344/2081.

GONZALEZ, M.; LIMA, V. L. S. Recuperação de Informação e Processamento da Linguagem Natural. **XXIII Congresso da Sociedade Brasileira de Computação**, v. 3, p. 347-395, 2003.

GROVER, D.; KOCHAR, B. Information Retrieval on Web: ontology based vs traditional search engines. **International Journal of Recent Technology and Engineering (IJRTE)**, v. 8, n. 3, p. 901-903, set. 2019. http://dx.doi.org/10.35940/ijrte.c4085.098319.

ISOTANI, S.; BITTENCOURT, I. I. **Dados Abertos Conectados:** Em Busca da Web do Conhecimento. Novatec, 2015.

KULKARNI, A.; SHIVANANDA, A. **Natural Language Processing Recipes**. 2. ed. Apress, 2021. 358 p.

LANE, H.; HOWARD, C.; HAPKE, H. **Natural Language Processing**: in action. New York: Manning Publications Co., 2019.

LAUFER, C. **Guia de Web Semântica**. 2015. Disponível em: https://ceweb.br/guias/web-semantica/. Acesso em: 09 jun. 2024.

PASSIN, T. B. **Explorer's Guide to the Semantic Web**. Manning Publications, 2004. 300 p.

PEFFERS, K.; TUUNANEN, T.; ROTHENBERGER, M. A.; CHATTERJEE, S. A Design Science Research Methodology for Information Systems Research. **Journal of Management Information Systems**, v. 24, n. 3, p. 45-77, 2007. Informa UK Limited. http://dx.doi.org/10.2753/mis0742-1222240302.

PIMENTEL, M.; FILIPPO, D.; SANTOS, T.. Design Science Research: pesquisa científica atrelada ao design de artefatos. **Re@D - Revista de Educação A Distância e Elearning**, v. 31, p. 37-61, 2020. http://dx.doi.org/10.34627/VOL3ISS1PP37-61.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: sentence embeddings using siamese bert-networks. **ArXiv**, 2019. ArXiv. http://dx.doi.org/10.48550/ARXIV.1908.10084.

ROTHMAN, D. **Transformers for Natural Language Processing**. Packt Publishing, 2021. 384 p.

SANTIAGO, C.; VERAS, N.; ARAGÃO, A.; CARVALHO, D.; AMARAL, L. Desenvolvimento de sistemas Web orientado a reuso com Python, Django e Bootstrap. **Minicursos da ERCEMAPI 2020**, p. 97-120, set. 2020. SBC. http://dx.doi.org/10.5753/sbc.11.5.5.

SARKAR, D. **Text Analytics with Python:** a Practical Real-World Approach to Gaining Actionable Insights from Your Data. Apress, 2016. 385 p.

SAYED, A.; MUQRISHI, A. A. IBRI-CASONTO: Ontology-based semantic search engine. **Egyptian Informatics Journal,** 2017. http://dx.doi.org/10.1016/j.eij.2017.01.001.

SEGARAN, T.; EVANS, C.; TAYLOR, J. **Programming the Semantic Web:** Build Flexible Applications with Graph Data. O'Reilly Media, 2009. 302 f.

SRINIVASA-DESIKAN, B. **Natural Language Processing and Computational Linguistics**: a practical guide to text analysis with python, Gensim, spaCy and Keras. Packt Publishing, 2018.

SUMIT, R. **Building Chatbots with Python**. Apress, 2019. 214 p.

VAISHNAVI, V. K.; KUECHLER, W. **Design Science Research Methods and Patterns Innovating Information and Communication Technology.** New York: Auerbach Publications, 2007.

VAN ASSEM, M. F. J.; GAMGEMI, A.; SCHREIBER, A. T. Conversion of WordNet to a standard RDF/OWL representation. **Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)**, p. 1-6, 2006.

VASILIEV, Y. **Natural language processing with Python and SpaCy**: a practical introduction. No Starch Press, 2020. 216 p.

WU, G.; YUAN, Y. **Lexical Ontological Semantics**. Routledge, 2019. 248p.