COMPARAÇÃO SEMÂNTICA: ANÁLISE DE *WORDNET* EM PORTUGUÊS PARA COMPOSIÇÃO DE UM BANCO DE DADOS RELACIONAL NA COMPARAÇÃO DE PRODUÇÕES TEXTUAIS

Rute Steffany da Silva[1], José Leonardo Oliveira Lima[1]
rute30.10@gmail.com, jjleo@ueg.br (https://orcid.org/0000-0001-8869-3056)

[1] Universidade Estadual de Goiás, Sistemas de Informação, Anápolis, Goiás

RESUMO: A Educação a Distância e a Educação Híbrida trazem alguns desafios para os professores na obtenção de *feedbacks* a respeito da aprendizagem dos alunos em relação aos conteúdos ministrados. Essa pesquisa tem por objetivo identificar a *Wordnet* em Português que melhor se adequa ao processo de comparação semântica – com foco no estudo da base de dados da *Wordnet,* o seu funcionamento e como essa pode ser modelada e disponibilizada em um banco de dados relacional – para ser utilizada em um futuro algoritmo para verificar o nível de absorção de conhecimento extraído dos textos de estudo pelo aluno. A pesquisa caracteriza-se como metodológica e descritiva tendo também caráter bibliográfico, com buscas em base de dados e artigos científicos para o levantamento metodológico do foco delimitado para o objetivo da pesquisa. O projeto ainda está em andamento sendo que, até o momento, já foram feitos levantamentos de referencial teórico e de metodologias importantes e também identificação de uma *Wordnet* candidata a ser utilizada.

Palavras-Chave: Comparação semântica, educação a distância, ensino híbrido, rede semântica e *Wordnet*.

SEMANTIC COMPARISON: WORDNET ANALYSIS IN PORTUGUESE FOR COMPOSITION OF A RELATIONAL DATABASE IN THE COMPARISON OF TEXTUAL PRODUCTIONS

ABSTRACT: Distance Education, and Hybrid Education bring some challenges for teachers in obtaining learning feedbacks about the contents taught. This research aims to identify the Wordnet in Portuguese that best suits the semantic comparison process – focusing the study of its database, its functioning, and how it can be modeled and aggregated in a relational database – to be used in a future algorithm to verify the level of absorption of knowledge extracted from the texts given for students studies. The research is characterized as methodological and descriptive and also has a bibliographic characteristic, with searches in scientific databases and articles for the methodological survey of the focus defined for the research objective. The project is still in progress and, up to now, surveys of theoretical references and important methodologies have been carried out and also the identification of a candidate Wordnet to be used.

Keywords: Semantic comparison, distance education, hybrid teaching, semantic network and Wordnet.

1. INTRODUÇÃO

A Educação a Distância (EaD) vem conquistando seu espaço no ensino superior por proporcionar uma estrutura de aprendizagem flexível e eficiente que se dá, segundo Moran (2002), em virtude do processo de ensino-aprendizagem ser mediados por tecnologias, tornando acessível para as pessoas que não podem estar em uma sala de aula presencial.

Por conta da grande demanda de ingressantes na modalidade EaD, os docentes vêm enfrentando dificuldades no processo de avaliação e consequentemente na obtenção do feedback de aprendizagem de cada aluno.

Além da modalidade EaD, tem-se também o ensino híbrido, no qual os alunos têm aula presenciais e não presenciais, mediadas pelo uso intensivo das tecnologias digitais.

A EaD, e também o ensino híbrido (a depender do grau de hibridismo e das peculiaridades de cada projeto de curso), pode usar das produções textuais em ambientes virtuais de aprendizagem para fazer a validação do aprendizado. Ressalta-se que os métodos de avaliação não são limitados somente às produções textuais, pois o docente tem autonomia para escolher a forma de avalição conforme projeto pedagógico do curso. Contudo, para o presente trabalho, foi priorizado o método de produção textual, com vistas a propor mecanismo para auxiliar o professor na obtenção do *feedback* sobre o aprendizado de cada aluno, conforme uma das necessidades de informação apontadas por especialistas em EaD que foi identificada na pesquisa de Lima (2016).

Sendo assim, a pesquisa baseia-se no seguinte problema de pesquisa: Qual *Wordnet* em Português pode ser integrada a um banco de dados para ser usado em um algoritmo de comparação semântica de textos, de forma a verificar se o discente, por meio de suas produções textuais, está construindo conhecimento a partir dos textos disponibilizados pelo docente?

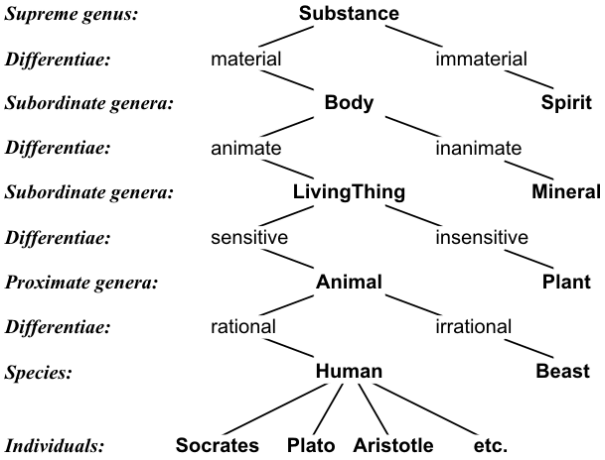
Com isso, tem-se como objetivo: Identificar a *Wordnet* em Português que melhor se adequa ao processo de comparação semântica – com foco no estudo da base de dados da Wordnet, o seu funcionamento e como essa pode ser modelada e disponibilizada em um banco de dados relacional – para ser utilizada em um futuro algoritmo para verificar para verificar o nível de absorção de conhecimento extraído dos textos de estudo pelo aluno. Ressalta-se que a verificação do possível algoritmo é delimitada aos textos produzidos pelos discentes a partir dos textos disponibilizados pelo professor para estudo de um determinado conteúdo, para se chegar ao nível de absorção de conhecimento pelo aluno, o que direciona a identificação da Wordnet em Português e sua análise.

2. REVISÃO BIBLIOGRÁFICA

Atualmente, existem diferentes estudos sobre comparação de textos, em áreas distintas, que buscam analisar a comunicação humana. Para a presente pesquisa, buscou-se o estudo da comparação semântica, que busca representar os textos por seus significados.

A representação do conhecimento pode ser feita de diversas formas, dentre as quais as redes semânticas que, segundo Rosa (2016) e Sowa (1992 apud ROSA, 2016), são padrões de representação do conhecimento e de processamento de linguagem natural baseados em grafos, onde os vértices representam as palavras e as arestas os relacionamento entre elas conforme exemplo na figura 1.

Figura 1 – Exemplo de grafo e aresta de rede semântica



Fonte: Sowa (1992).

Conforme as definições de Cunha (2013), a rede semântica tem como seu elemento principal a relação entre palavras, adequando-se a variedades de métodos computacionais, sendo assim considerado um dicionário legível para máquinas.

Seguindo esse conceito, as *Wordnets* podem ser consideradas um tipo de rede semântica, pois integram um grande conjunto de palavras, com suas relações semânticas.

As *Wordnet*s, segundo Miller (1995), são formadas por conjuntos de substantivos, adjetivos, verbos e advérbios, organizados em grupos de sinônimos denominados "*synsets*", sendo interligados por relações semânticas. Moraes (2008) define que cada *synset* reúne itens lexicais compartilhando um mesmo conceito. As definições das palavras da *Wordnet* são organizadas de forma conceitual (LEACOCK; CHODOROW, 1998 apud FELLBAUM, 1998), diferenciando-se da organização alfabética de um dicionário comum.

Quando se fala em *Wordnet*, é necessário fazer a distinção de qual se trata, pois ao decorrer dos anos, foram criadas várias *Wordnets* para diversas línguas, cada uma com suas peculiaridades. A primeira a ser desenvolvida foi a *Wordnet* de *Princeton*, que se encontra disponível desde a década de 90 (MILLER, 1995 apud FELLBAUM, 1998), a qual serve como norteadora e base para todas as *Wordnets* da atualidade.

Por conta da grande popularidade da *Wordnet* de *Princeton* como base de conhecimento, foi criada uma organização não comercial, a *Global WordNet Association* (GWA), uma plataforma que partilha, discute e faz ligação com as *Wordnet* de todo o globo (BOND; PAIK, 2012; OLIVEIRA et al. 2015). Com a criação da GWA o processo de classificação e criação das *Wordnets* se tornou mais fácil.

O processo de desenvolvimento de algumas *Wordnets,* foi feito de modo manual, ou seja, cada palavra era selecionada manualmente, conforme seu significado e semântica, com as outras comparando uma a uma. Essas palavras classificadas foram sendo armazenadas em base de dados lexicais.

Para Greghi (2002) as Bases de Dados Lexicais (BDL) são normalmente utilizadas como repositório central de informações lexicais de uma determinada língua. Tais informações armazenadas podem ser de âmbito sintático, semântico, morfológico, podendo também expressar relações dentre lexicais de mesma língua, ou entre línguas distintas, sendo que essas bases de dados lexicais facilitam a manutenção e manipulação das informações (GREGHI, 2002).

A comparação entre textos é feita pela comparação e verificação da similaridade de ambos, ou seja, se o assunto de ambos os textos é igual ou parecido, pode-se classificá-los como semelhantes. Silva (2008) e Costa (2017) esclarecem que o processo de identificação das similaridades semânticas é feito seguindo diversos tipos de abordagens. De acordo com Wang (2005 apud SILVA, 2008) e Petrakis et al. (2006 apud SILVA, 2008), as medidas de similaridade semântica podem ser divididas em quatro categorias principais, sendo:

- Abordagem baseada em ontologias: Categorizada como todas as abordagens que utilizam recursos e bases de conhecimento, como as ontologias e *Wordnets*, calculando o grau de similaridade;
- Abordagem baseada no índice de informações compartilhadas: Quanto mais um termo tende a ser similar a outro mais se assemelham, ou seja, compartilham o mesmo grau de informação;
- Abordagem baseada em características: Os termos só são semelhantes se compartilharem das mesmas características, exemplo: a palavra “estudante”, é semelhante a aluno, discente, acadêmico, entre outras, que remetem a uma pessoa que estuda;

- Abordagem híbrida: Junção de duas ou mais das abordagens anteriores, tornando o seu processo de similaridade mais seguro, pois faz a comparação dos termos de mais de uma forma possível.

A similaridade semântica tem uma ampla área de aplicação, dentre as quais pode-se destacar: recuperação de informação; integração de dados; integração de base de dados; detecção de ambiguidades e duplicação de informação (COSTA et al.,2014).

3. METODOLOGIA

Seguindo os critérios de classificação de pesquisa apresentados por Vergara (2013), a presente pesquisa é classificada: **quanto aos fins**, como uma pesquisa metodológica e descritiva; **quantos aos meios**, como uma pesquisa bibliográfica.

Para o levantamento bibliográfico, foram utilizadas as plataformas *online* de pesquisas, sendo elas: o *Google Academics,* o Portal de Periódicos da CAPES e a Biblioteca Digital de Tese e Dissertações (BDTD). Os dados levantados, até o presente momento, foram tratados de forma qualitativa, por meio das sínteses dos materiais pesquisados.

Busca-se fazer a identificação da *Wordnet* que melhor se adequa no processo de comparação semântica, fazendo a análise de sua base de dados e como ela pode ser conectada em um banco de dados para o processo de comparação semânticas das produções textuais.

Essa identificação será feita por meio do levantamento das *Wordnets* desenvolvidas para o português, buscando, conforme critérios que estão em processo de definição, dentre eles o fato de ser livre e de código aberto, selecionar a que mais se adequa as necessidades de comparação textual foco da pesquisa.

4. CONSIDERAÇÕES FINAIS

Considerando as análises de trabalhos até o presente momento, pode-se expor que o processo de comparação semântica é uma área de estudo complexa, podendo ser identificada e estuda por diferentes áreas e métodos.

Para o processo de comparação baseada em *Wordnet* está sendo feito o estudo da base de dados de uma *Wordnet* candidata (dentre quatro *Wordnets* em Português encontradas), para verificar o potencial dela para que, ao final do estudo dessa base de dados, ela possa ser modelada e integrada em um banco de dados relacional, utilizando a *Standard Query Language* (SQL) como forma de busca. Contudo, ainda se faz necessário realizar análises mais detalhadas em algumas

metodologias especificas da área, para que o processo seja plenamente sustentado.

REFERÊNCIAS

BOND, Francis; PAIK, Kyonghee. **A survey of wordnets and their licenses**. 2012 p. 64-71. Disponível em: < https://www.researchgate.net/profile/Francis_Bond/publication/267427763_A_Survey_of_WordNets_and_their_Licenses/links/549cccaa0cf2fedbc30fe027 >. Acesso em: 30 maio 2020.

COSTA, Douglas de Jesus. **Análise de algoritmo de comparação semântica de textos:** Uma abordagem sob a ótica cognitivista de Piaget. 2017. 120 p. Monografia (Curso de Sistemas de Informação) - Universidade Estadual de Goiás, Anápolis, 2017.

CUNHA, Marcelo do Vale. **Redes semânticas baseadas em títulos de artigos científicos.** 2013. 127p. Dissertação de Mestrado (Mestrado em Modelagem Computacional e Tecnologia Industrial) – SENAI CIMATEC, Salvador, 2013.

FELLBAUM, Christiane; MILLER, George A. **WordNet:** an electronic lexical database language. 1. ed. MIT Press, 1998. p. 1-423. Disponível em: <https://books.google.com.br/books?id=Rehu8OOzMIMC&lpg=PA265&ots=Irp8HhWXc6&dq=Leacock%20e%20Chodorow%201998&lr&hl=pt-BR&pg=PR3#v=onepage&q&f=false>. Acesso em: 30 maio 2020.

GREGHI, Juliana Galvani et al. **Projeto e desenvolvimento de uma base de dados lexicais do português.** 2002. Tese de Doutorado. Dissertação de Mestrado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

LIMA, J. L. O. **Avaliação discente em cursos de graduação a distância mediados por ambientes virtuais de aprendizagem:** Necessidade de informações dos docentes na visão de especialistas europeus e brasileiros. 2016. 298f. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, Brasília, 2016.

MILLER, George A.; **Wordnet**: a lexical database for english. In: _____. Communications of. 1995. v. 38. p. 39-41. Disponível em: < https://dl.acm.org/doi/10.1145/219717.219748 >. Acesso em: 30 maio 2020.

ROSA, Marcos G.; **Modelo empírico para analisar a robustez de redes semânticas.** 2016. 134p. Tese (Doutorado Multi-institucional e Multidisciplinar em Difusão do Conhecimento) – Universidade Federal da Bahia. Faculdade de Educação, Salvador, 2016.

SILVA, Daniel Ferreira da. **Estudo de Funções de Similaridade Semântica de Termos Aplicadas a um Domínio.** 2008. p. 45. Monografia (Curso de Ciência da Computação) – Universidade Federal de Pernambuco, Pernambuco, 2008. Disponível em: < https://www.cin.ufpe.br/~tg/2007-2/dfs3.pdf>. Acesso em: 25 set 2020.

VERGARA, Sylvia Constant. **Projetos e relatórios de pesquisa em administração**. 14. ed. São Paulo: Atlas, 2013.