SEMANTIC SIMILARITY: a Domain Analysis

Similaridade Semântica: uma Análise de Domínio

Rita Costa (1), Thiago Bragato (2), Renato Fileto (3)

(1) Universidade Federal de Santa Catarina, Brasil, rita.alamino@posgrad.ufsc.br

(2) Universidade Federal do Rio Grande do Sul, Brasil, bragato.barros@ufrgs.br

(3) Universidade Federal de Santa Catarina, Brasil, r.fileto@ufsc.br

Abstract

In the rapidly evolving field of Natural Language Processing (NLP), understanding the domain of semantic similarity is of paramount importance for both academic and industrial applications. This article presents a comprehensive domain analysis of semantic similarity, integrating a multidisciplinary approach that encompasses key concepts, interrelations among these facets, stakeholders, information practices, and existing classification systems. We elucidate the core ideas, such as lexical and syntactic similarity, embeddings, and various similarity metrics, and demonstrate their interrelatedness. The paper also identifies and characterizes the diverse array of stakeholders involved in this domain, from academic researchers and tech leads to policymakers and open-source communities. Furthermore, we explore how information is disseminated and used within this domain, including an examination of research publication trends and industry reports. Lastly, the article assesses existing classification systems and ontologies that structure the knowledge in this field. Our findings serve as a foundational framework for future research, development, and ethical considerations in the semantic similarity domain. This in-depth analysis aspires to guide both newcomers and seasoned experts through the intricate landscape of semantic similarity, thereby contributing to the field's holistic advancement.

Keywords: Domain Analysis; Semantic Similarity; Natural Language Processing; Knowledge Organization.

Resumo

No campo em rápida evolução do Processamento de Linguagem Natural (PLN), entender o domínio da similaridade semântica é de extrema importância tanto para aplicações acadêmicas quanto industriais. Este artigo apresenta uma análise abrangente do domínio da similaridade semântica, integrando uma abordagem multidisciplinar que abrange conceitos-chave, inter-relações entre essas facetas, partes interessadas, práticas de informação e sistemas de classificação existentes. Elucidamos as ideias centrais, como similaridade léxica e sintática, *embeddings* e várias métricas de similaridade, e demonstramos como elas estão inter-relacionadas. O artigo também identifica e caracteriza a diversa gama de partes interessadas envolvidas neste domínio, desde pesquisadores acadêmicos e líderes técnicos até formuladores de políticas e comunidades de código aberto. Além disso, exploramos como a informação é disseminada e usada dentro deste domínio, incluindo um exame das tendências de publicação de pesquisas e relatórios industriais. Por fim, o artigo avalia os sistemas de classificação e ontologias existentes que estruturam o conhecimento neste campo. Nossas descobertas visam servir como uma estrutura fundamental para futuras pesquisas, desenvolvimentos e considerações éticas no domínio da similaridade semântica. Esta análise profunda aspira orientar tanto recém-chegados quanto especialistas experientes pelo intrincado panorama da similaridade semântica, contribuindo assim para o avanço holístico do campo.

Palavras-chave: Análise de Domínio; Similaridade Semântica; Processamento de Linguagem Natural; Organização do Conhecimento.

1 Introduction

Semantic similarity, a subfield within Natural Language Processing (NLP) and computational linguistics, has garnered substantial attention over the past few decades. It operates at the intersection of linguistics and computer science to quantitatively assess the degree of relatedness between linguistic units, ranging from words to sentences and even entire documents (Teller, 2020). The significance of semantic similarity spans across various domains, including information retrieval, machine translation, text summarization, and recommender systems, among others.

1.1 Knowledge Organization and Domain Analysis

The multidisciplinary field of Knowledge Organization provides a powerful lens through which to approach the study of semantic similarity. At its core, Knowledge Organization is concerned with the representation, description, and organization of concepts, subjects, and documents across domains (Hjørland 2008). Within this realm, domain analysis emerges as a vital methodological approach, scrutinizing how knowledge is structured and organized intellectually within academic disciplines and research areas.

A central tenet of domain analysis is the recognition that a single object or document can be classified through multiple valid perspectives, each shaped by the unique viewpoints, theories, and informational needs of different scholarly communities. This pluralistic understanding highlights that distinct groups interpret and organize knowledge in diverse ways, informed by their guiding paradigms and epistemological frameworks.

This domain-analytic lens becomes indispensable for semantic similarity, a subdomain within Natural Language Processing. To advance holistically in this intricate field, a comprehensive grasp of the fundamental concepts, their intricate interrelationships, the array of stakeholders involved, and the existing classification systems and information practices is paramount. It is through this multifaceted domain analysis that the rich conceptual, theoretical, and applied landscapes of semantic similarity can be elucidated and navigated effectively.

1.2 Significance of the Domain

Understanding semantic similarity is not merely an academic exercise but a critical component in the development of intelligent systems. Its algorithms power search engines that sort through vast oceans of data, recommendation systems that personalize user experiences, and machine translation services that bridge language barriers. Moreover, with the increasing integration of machine learning in decision-making processes, ethical considerations like fairness and interpretability are pushing the boundaries of the domain towards more socially responsible algorithms.

1.3 Research Objectives

Given the complexities and myriad applications of semantic similarity, it is imperative to conduct a domain analysis to elucidate its multifaceted landscape. This paper aims to:

- Identify Key Concepts and Terms: To establish a comprehensive understanding of the fundamental building blocks, such as lexical and syntactic similarity, vector spaces, ontologies, and various similarity measures.

- Examine Stakeholder Needs: To understand the various actors involved in this domain, such as academic researchers, industry practitioners, and policymakers, and their respective informational necessities.
- Analyze Information Practices: To scrutinize how information is created, disseminated, and utilized, focusing on academic publications, conferences, and real-world applications.
- Investigate Domain-Specific Challenges and Issues: To highlight unique problems like ambiguity resolution, scalability, and ethical concerns, that the field faces.

By addressing these objectives, the paper aims to provide a holistic understanding of the domain of semantic similarity, thereby laying the groundwork for future research and practical applications.

1.4 Structure

The ensuing sections will delve into each of these objectives in detail, presenting a domain analysis aimed at both novice researchers and seasoned practitioners. Section 2 will provide a comprehensive review of the existing literature on semantic similarity, identifying the major approaches, techniques, and applications. Section 3 will elucidate the fundamental concepts and terminology of the domain in depth, establishing a solid foundation for subsequent analysis.

Section 4 will examine the interrelations and hierarchies among the various concepts and aspects of the domain, offering a structured view of the field. In Section 5, the diverse stakeholder groups involved, including academic researchers, industry professionals, policymakers, and others, along with their respective information needs, will be identified and characterized. Section 6 will conduct a detailed examination of how information is created, disseminated, and utilized within the domain, focusing on academic publications, conferences, and real-world applications. Finally, Section 7 will summarize the key findings and insights gained through this comprehensive domain analysis, highlighting its significance and implications for future research and applications.

2 Literature Review

Semantic Text Similarity (SST) is the semantic proximity between two blocks of text (Chandrasekaran and Mago 2021). SST measures the degree of semantic affinity based on shared semantic properties. It measures similarity rather than making a classification, for example, a binary decision of similar or not similar. Such a measure is based on the meaning and context of the text rather than just considering the presence of similar words or characters.

SST is used in various PLN tasks and is particularly useful in plagiarism detection, information retrieval, text summarization, text classification, machine translation, and (Wang et al. 2020) question answering. Semantic similarity can be calculated using various alternative methods, including knowledge-based, corpus-based, deep neural network-based, and hybrid methods. These methods can exploit various techniques, such as semantic networks and machine learning algorithms, to capture and measure the semantic properties of texts.

To identify and analyze pertinent studies that deal with "Semantic Similarity," a literature review was carried out using digital libraries. The research questions guided the investigation in three main directions:

- How is semantic similarity defined in the context of discourse?
- What are the current techniques and methods for measuring semantic similarity of texts?
- What are the applications of semantic similarity in discourse analysis?

The literature review started with searches in digital libraries using the following keywords:

Text Similarity AND Semantic Similarity AND Semantic Textual Similarity

These searches allowed us to retrieve 21 articles relevant to the analysis of semantic similarity. The proposals related to semantic similarity, which we found in a systematic mapping of the literature, are based on structural analysis (Torkanfar and Azar 2020, Almuhaimeed et al. 2022), statistical (Mehndiratta and Asawa 2020), and ontological methods (Jha et al. 2022, Yang et al. 2021). However, as has been the case with many 1 tasks, there is a recent trend toward

transitioning to deep neural models (Joty et al. 2019, Lv et al. 2021, Cao et al. 2022, Wang and Zhang 2021, Wang et al. 2021, Malkiel et al. 2022, An et al. 2020, Chen et al. 2023, Peng et al. 2021, Sonawane and Kulkarni 2022, Xiao et al. 2022).

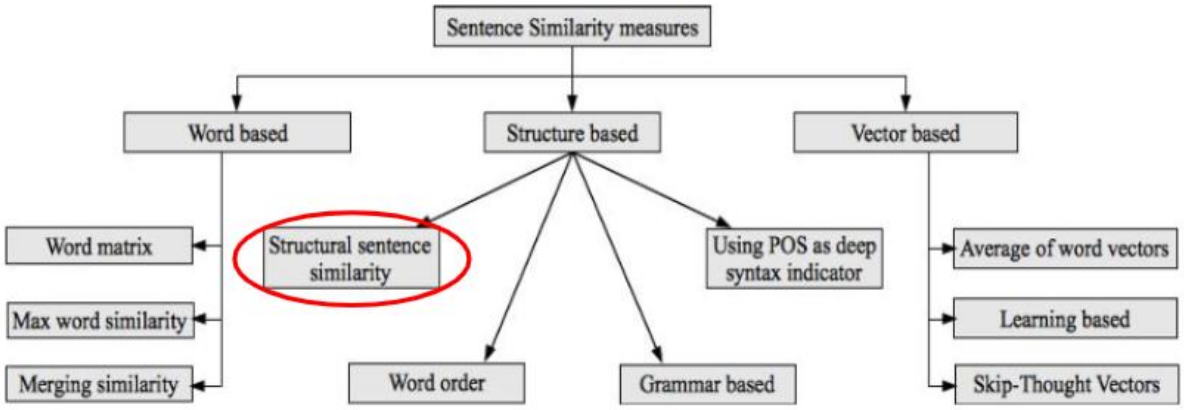
The solutions in the articles found are represented by various approaches to semantic similarity, each with distinctive features and methods. Categories include Ontological Approaches, Neural Network-Based Approaches, Graph-Based Approaches, Statistical Approaches, Ensemble Approaches, Hybrid Approaches, and Other Approaches.

- **Ontological Approaches:** These solutions use ontologies or knowledge graphs to represent concepts and their relationships.
- **Neural Network-Based Approaches:** These solutions use neural networks to learn representations of concepts and calculate similarity scores.
- **Graph-Based Approaches:** These solutions represent concepts as nodes in a graph and use graph algorithms to calculate similarity scores.
- **Statistical Approaches:** These solutions use statistical methods to calculate similarity scores based on features such as word co-occurrence.
- **Ensemble Approaches:** These solutions combine several methods to improve performance.
- **Hybrid Approaches:** These solutions combine several approaches. Example LDA - Latent Dirichlet Allocation (Blei 2003) and L-THM, a hybrid of LDA with TF-IDF - Term Frequency-Inverse Document Frequency (Wang 2019), which combines ontology-based approaches and neural networks.
- **Other Approaches:** These solutions need to fit more neatly into the above categories.

Figure 1 presents a taxonomy of approaches proposed in the literature to measure sentence similarity. They can be classified into three main classes, as illustrated in Figure 1: word-based, structure-based, and vector-based. Word-based similarity uses matrices, maximum values, or combinations of similarity between words to determine the similarity between sentences. Vector-based similarity depends on converting sentences into vectors (embeddings), which capture

semantic features, and then calculating the similarity between sentences based on these vectors. They are generated using machine learning techniques and eventually concatenating or averaging word vectors to create vectors of more significant portions of text. Lastly, the structure-based approach calculates the similarity based on the structure of (Farouk 2019) sentences.

Figure 1: Approaches for measuring sentence similarity



Source: (Farouk 2019)

3 Key Concepts and Terms

The domain of semantic similarity plays a crucial role in various natural language processing tasks, including but not limited to text clustering, summarization, and machine translation. An understanding of key concepts and terms is fundamental for the effective analysis and implementation of semantic similarity methods. This chapter aims to elucidate these fundamental terms and concepts.

The concept of semantic similarity lies at the heart of many applications in text mining, natural language processing, and information retrieval. When dealing with semantic similarity specifically, one ventures into a highly interdisciplinary domain that borrows theories and methods from linguistics, computer science, and data science. A wide array of concepts and terminology forms the backbone of this area (Jurafsky and Martin, 2021).

Lexical Units: At the most fundamental level, lexical units serve as the basic entities for semantic comparison. These are the words or terms that carry meaning and are the building blocks of language and discourse.

Syntax Trees: Data structures that represent the grammatical relations within sentences. Although they are more closely related to syntactic rather than semantic similarity, understanding syntax trees can nonetheless aid in more complex semantic similarity models.

Embeddings: High-dimensional vector representations of words or sentences. These could be static, often referred to as Semantic Vector Spaces, or dynamic in the form of Contextual Embeddings. Embeddings capture the essence of a word's meaning within a mathematical construct.

Document Vectors: An aggregated form of word embeddings, document vectors aim to encapsulate the semantic essence of larger text units like paragraphs or full documents. These vectors provide a holistic approach to document-level semantic similarity.

Ontology: Structured framework that describes the relationships between different concepts within a particular domain. In semantic similarity, ontologies can be employed to understand and quantify semantic relations in a more structured manner.

Knowledge Graphs: A form of ontology, knowledge graphs are network-like structures that describe relationships between various entities and concepts. They often serve to augment semantic similarity assessments by providing a contextual backdrop against which semantic relations can be examined.

Cosine Similarity: Cosine similarity measures the cosine of the angle between two vectors in a multi-dimensional space. It is a frequently employed metric in document-level similarity assessments and even in comparing high-dimensional word vectors.

Jaccard Index: The Jaccard Index is a statistical measure used to evaluate the similarity and diversity between two sample sets. While it's generally used at the lexical or document level, it serves as a straightforward and effective measure of similarity.

Classification Algorithms: Various machine learning models, such as decision trees, Support Vector Machines (SVMs), or neural networks, are used for the purpose of categorizing levels of similarity. These models can be trained on a multitude of features, drawing from the aforementioned concepts.

Context: Finally, the notion of context refers to the surrounding textual environment in which a word or phrase appears. Context is critical for advanced semantic similarity models like BERT or GPT, as it captures the dynamic nature of language.

3.1 Semantic Relations

The domain of semantic similarity is inherently intertwined with the representation and modeling of semantic relations between linguistic units. Bräscher's (2014) seminal work on analyzing semantic relations in knowledge organization systems (KOS) provides a theoretical grounding that informs our understanding of this domain. Traditional KOS like thesauri and classifications have long captured hierarchical, equivalence, and associative relations between concepts and terms. As Bräscher elucidates, while valuable, these paradigmatic relations offer an incomplete view, lacking the richer contextualized syntagmatic relationships manifested in natural language.

This aligns with the evolutionary trajectory witnessed in semantic similarity research, which has seen a transition from simple lexical and ontological measures to embedding-based models that can capture contextualized semantic and syntactic associations between words, phrases, and sentences. Bräscher's call for an interdisciplinary synthesis from fields like linguistics and AI to theoretically fortify the modeling of semantic relations resonates with the multi-faceted landscape outlined in this domain analysis. Her emphasis on establishing robust theoretical frameworks for semantic relations can serve as a foundation for the holistic research and ethical development of semantic similarity technology envisioned herein.

Understanding the key concepts and terms in semantic similarity is paramount for any meaningful research or application in the field. This chapter has provided an overview of the most crucial terms, serving as a foundation for further study and implementation.

4 Relationships and Hierarchies

Understanding the relationships and hierarchies between various concepts within the domain of semantic similarity is pivotal for effective research and application. This chapter elucidates the interrelations among key facets of semantic similarity, offering a structured lens through which the domain can be understood.

The domain of semantic similarity is multifaceted, with several key concepts and techniques playing unique yet interrelated roles. Understanding the hierarchy and relationships among these facets can provide insights into how different methods and concepts can be combined or adapted for more effective semantic analyses (Jurafsky and Martin, 2021). The chapter aims to dissect these interrelations, serving as a guide for future research and application.

4.1 Lexical Similarity

Syntactic Similarity - Often serves as the foundation for syntactic analyses, as it builds upon lexical units to represent structural aspects of language.

Similarity Measures - Simple metrics such as string matching and Levenshtein distance are commonly employed in lexical similarity tasks.

4.2 Syntactic Similarity

Lexical Similarity - Builds upon lexical units to form structures that represent sentence or document syntax.

Semantic Vector Spaces/Contextual Embeddings - May incorporate advanced embeddings to augment syntactic analyses, serving as features for more complex models.

4.3 Semantic Vector Spaces

Contextual Embeddings - Often act as a precursor to more complex contextual models which take into account the semantic coherence of a text.

Document-Level Similarity - Forms the basis for many document-level similarity measures and is often used alongside cosine similarity metrics.

Classification Methods - These vector spaces can be used as features in various machine-learning classification models.

4.4 Contextual Embeddings

Semantic Vector Spaces - This can be considered as a more advanced extension that incorporates contextual information.

Document-Level Similarity - Increasingly being used to derive features for document-level semantic analyses.

Classification Methods - Being integrated into neural classifiers to improve performance and precision.

4.5 Document-Level Similarity

Semantic Vector Spaces/Contextual Embeddings - Commonly employs these advanced embeddings for feature extraction.

Similarity Measures - Metrics like cosine similarity are often applied at this level to quantify similarity between documents.

4.6 Ontologies and Knowledge Graphs

Semantic Vector Spaces/Contextual Embeddings - This can enrich these embeddings by adding layers of conceptual meaning and relational data.

Document-Level Similarity - Helpful in identifying semantic fields or topics within documents, thereby enriching document-level analyses.

4.7 Similarity Measures

All Categories - Metrics like cosine similarity, Jaccard index, and Euclidean distance can be applied across various facets to provide a quantitative measure of similarity.

Classification Methods - Semantic Vector Spaces/Contextual Embeddings - These embeddings often act as feature vectors in machine learning classifiers or neural models.

Document-Level Similarity - Serve as the target variable in a classification problem, where the aim is to categorize documents based on their level of similarity.

5 Stakeholders

In the intricate landscape of semantic similarity, multiple stakeholders coalesce, each wielding distinct objectives and contributions. Academic researchers act as the bedrock of theoretical innovation, providing novel algorithms, evaluation methodologies, and high-quality, peer-reviewed publications. Their work often serves as a launching pad for industry practitioners and tech leads, who translate theory into practice. Academic researchers play a fundamental role in developing new techniques and approaches for calculating semantic similarity, as evidenced by recent works exploring models based on deep neural networks (Malkiel et al., 2022; Cao et al., 2022). The latter group is particularly concerned with the scalability, efficiency, and real-world applicability of semantic similarity models. Their work is not just a simple implementation but often involves customization and fine-tuning to meet specific business or operational requirements. Large technology companies, such as Google, Microsoft, and Amazon, are at the forefront of applying semantic similarity algorithms in commercial products, with a focus on scalability and efficiency.

Policymakers and ethicists introduce yet another layer of complexity. As algorithms become increasingly entwined in the fabric of society, these stakeholders examine the ethical ramifications, such as fairness, data privacy, and societal impact. With the increased use of AI in decision-making systems, regulatory authorities have been pushing for more transparency and accountability, requiring that semantic similarity models be interpretable and free of biases (Babaeianjelodar et al., 2021). They often act as a liaison between technology creators and the public, advocating for ethical practices and responsible AI deployment. Technology companies, motivated by commercial interests, seek technologies that are both innovative and market-ready. They are interested in aspects like ease of integration into existing systems, scalability, and the overall efficiency of the algorithmic solutions.

End-users and consumers, often an overlooked but critical group, directly interact with applications driven by semantic similarity algorithms. Their feedback can offer invaluable insights into the usability and effectiveness of these systems. User experience studies have shown that recommendation systems based on semantic similarity can significantly improve consumer satisfaction and conversion rates (Konstan et al., 2012). Educational institutions rely on advancements from both the academic and industrial sectors to shape curricula, which ensures that the next wave of professionals is well-equipped to tackle emerging challenges in the field.

Data scientists and analysts operating in a space that straddles both academic research and practical application utilize semantic similarity for a range of data-intensive tasks, from feature engineering to complex analytics. Investors and funding bodies gauge the commercial potential and technological advantages, directing financial resources to areas of perceived future growth or societal impact. Open-source communities and legal professionals also contribute to shaping the field, the former by fostering collaboration and transparency and the latter by tackling issues related to intellectual property, potential misuse, and ethical considerations.

The symbiosis among these stakeholders is intricate and pivotal for the comprehensive development and dissemination of semantic similarity technologies. As the field undergoes rapid evolution, maintaining an open dialogue among these various entities will be crucial. It is this multi-stakeholder collaboration that will ultimately guide the ethical, academic, and practical trajectories of semantic similarity, addressing collective challenges and aligning disparate objectives for the greater good of the domain.

6 Information in the System

Within the domain of semantic similarity, the information ecosystem is both diverse and dynamic, governed by the practices and activities that facilitate the creation, dissemination, and utilization of knowledge. One can discern a multitude of avenues through which this occurs, each serving specific purposes and audiences.

Academic research, often the fountainhead of new theories and methodologies, propagates chiefly through scholarly journals, preprint archives, and conferences. These venues serve as

sanctuaries for rigorous peer review, fostering a culture of academic rigor and intellectual exchange. Conferences, in particular, serve as nexuses for interdisciplinary collaboration, offering platforms for researchers to interact and share emerging trends and technologies. The publication patterns often reveal cycles of interest in particular sub-domains, shaping the research agenda for years to come.

Parallel to the academic sphere, industry reports, and white papers provide actionable insights into the state-of-the-art applications of semantic similarity. These publications, generally less technical but more oriented toward business use cases, inform tech leads and industry practitioners about the scalable, efficient solutions that can be deployed in real-world settings (Chandrasekaran and Mago 2021; Jurafsky and Martin, 2021). Such reports are invaluable for deciphering market trends and identifying gaps where academic research can be translated into industrial solutions.

Government initiatives can also play a vital role, particularly through the lens of public policy and funding. It can accelerate technological innovation and responsible deployment by prioritizing semantic similarity in national research agendas and through strategic collaborations. The U.S. Defense Advanced Research Projects Agency (DARPA) has funded several programs that aim to improve information retrieval and question-answering capabilities. These programs likely involve research on semantic similarity as a key technique for achieving these goals. This often takes the form of grants, public-private partnerships, and sponsored research programs, creating a conducive environment for groundbreaking work in the field.

Information practices also extend to educational programs and curricula, which are designed to cultivate the next generation of professionals and researchers in the domain of semantic similarity. Top universities globally, including MIT, Stanford, and Carnegie Mellon, now offer specialized courses and degree tracks in natural language processing, equipping students with the latest semantic similarity techniques (Jurafsky and Martin, 2021) These curricula are regularly updated to incorporate recent advancements, ensuring a conversant workforce in both theory and practice.

Social media platforms, academic blogs, and webinars have also emerged as influential tools for the dissemination of knowledge. They democratize access to information and serve as repositories of informal yet highly impactful channels of communication. They are especially effective for continuous learning and for keeping the global community engaged with rapid advancements in the field.

In summary, the information practices and activities in the semantic similarity domain are not confined to traditional scholarly communication but are a rich tapestry of interconnected avenues that collectively contribute to the creation and propagation of knowledge. Understanding these practices is critical for anyone — be it an academic, a tech lead, or a policy maker—looking to navigate and contribute to this rapidly evolving field.

7 Conclusion

In conclusion, this domain analysis has offered a multi-faceted exploration of the field of semantic similarity within Natural Language Processing through the guiding lens of Knowledge Organization (KO) and domain analysis. By leveraging the core principles and methodological approaches of KO, we were able to conduct a comprehensive examination of key concepts, interrelations, stakeholders, and knowledge structures underpinning the semantic similarity domain.

Our analysis underscores that semantic similarity is not an isolated topic but an intricate field interwoven with various other domains, methods, and technologies. The KO perspective enabled us to untangle the complex relationships among lexical units, embeddings, similarity measures, and classification methods, unveiling an evolutionary trajectory from simplistic lexical matching to advanced machine learning techniques. Unraveling these technical nuances is crucial for researchers, developers, and practitioners engaged with semantic similarity and natural language processing technologies.

Moreover, a core strength of this KO-driven domain analysis lies in its emphasis on the human elements often overlooked in technological domains. By examining the diverse array of stakeholders, from academics and industry professionals to policymakers and end-users, we

highlighted the importance of recognizing their unique needs and constraints. This stakeholder-centric approach is vital for the inclusive and ethical development of semantic similarity technologies that can positively impact society.

The exploration of information practices and existing classification systems within the semantic similarity domain further exemplified the value of KO principles. Understanding how knowledge is organized, disseminated, and consumed can significantly influence the direction of future research and real-world implementation of these technologies across various sectors, from healthcare and education to finance and law.

Ultimately, this study has demonstrated that KO, with its domain-analytic methodology, provides a robust framework for comprehensively investigating complex, multidisciplinary fields like semantic similarity. By integrating technical concepts with human factors and knowledge structures, this domain analysis has laid a solid foundation for future research to build upon. As semantic similarity technologies continue to shape our society, maintaining a holistic, KO-driven perspective will be crucial for ensuring their ethical integrity and societal advancement.

Referências

An, Hongda, et al. "Hybrid Self-Interactive Attentive Siamese Network for Medical Textual Semantic Similarity." *Proceedings of the 2020 4th International Conference on Management Engineering, Software Engineering and Service Sciences*, p. 52-56, 2020 DOI: https://doi.org/10.1145/3380625.3380647.

Almuhaimeed, Abdullah, et al. "A modern semantic similarity method using multiple resources for enhancing influenza detection." *Expert Systems with Applications,* v. 193, p. 116466, 2022.

Babaeianjelodar, Marzieh. *Towards Fair and Transparent Decision Making and Machine Learning Systems.* Diss. Clarkson University, 2021.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." *Journal of machine Learning Research*, v. 3, p. 993-1022, 2003.

Bräscher, Marisa. "Semantic Relations in Knowledge Organization Systems." *Knowledge Organization*, v. 41, n. 2, p. 175–80, 2014.

Cao, Son, et al. "Hybrid Approach for Text Similarity Detection in Vietnamese Based on Sentence-BERT and WordNet. "*ITCC '22: Proceedings of the 4th International Conference on Information Technology and Computer Communications*, pp. 59-63, 2022. DOI: https://doi.org/10.1145/3548636.3548645,

Chandrasekaran, Dhivya, and Vijay Mago. "Evolution of semantic similarity—a survey." *ACM Computing Surveys (CSUR)*. v. 54, n. 2, p.1-37, 2021.

Chen, Qiang, et al. "Fine-grained semantic textual similarity measurement via a feature separation network." Applied Intelligence, v. 53, p. 18205-18218, 2023. DOI: https://doi.org/10.1007/s10489-022-04448-6.

Farouk, Mamdouh. "Measuring sentences similarity: a survey." *Indian Journal of Science and Technology*, v. 12, n. 25, 2019. DOI: https://doi.org/10.17485/ijst/2019/v12i25/143977.

Hjørland, Birger. "What is Knowledge Organization (KO)?" *Knowledge Organization*, v. 35, n. 2, p. 86-101, 2008. DOI: https://doi.org/10.5771/0943-7444-2008-2-3-86.

Jha, Akshita, et al. "Supervised Contrastive Learning for Interpretable Long-Form Document Matching." ACM Transactions on Knowledge Discovery from Data, v.17. n. 2, p. 27, 2023. DOI: https://doi.org/10.1145/3542822.

Joty, Shafiq, et al. "Discourse analysis and its applications." *I: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: tutorial abstracts."*, Florence, Italy, 2019, p. 12-17. DOI: https://doi.org/10.18653/v1/P19-4003.

Jurafsky, Dan, and James H. Martin. Speech and Language Processing (3rd ed. draft). 2021, https://web.stanford.edu/~jurafsky/slp3/. Accessed 17 June 2024.

Konstan, Joseph A., and John Riedl. "Recommender systems: from algorithms to user experience." *User modeling and user-adapted interaction*, v. 22, p. 101-123, 2012.

Lv, Chao, et al. "Siamese Multiplicative LSTM for Semantic Text Similarity." *In: ACAI'20: Proceedings of the 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence*, p. 28, 2021, DOI: https://doi.org/10.1145/3446132.3446160.

Malkiel, Itzik, et al. "Interpreting BERT-Based Text Similarity via Activation and Saliency Maps." *In: WWW'22: Proceedings of the ACM Web Conference 2022*, p. 3259-3268, 2022. DOI: https://doi.org/10.1145/3485447.3512045.

Mehndiratta, Akanksha, and Krishna Asawa. "Spectral Learning of Semantic Units in a Sentence Pair to Evaluate Semantic Textual Similarity." *In:* Bellatreche, L., Goyal, V., Fujita, H., Mondal, A., Reddy, P.K. (eds) *Big Data Analytics: 8th International Conference, BDA 2020*. Sonepat, India, 2020. DOI: https://doi.org/10.1007/978-3-030-66665-1_4.

Peng, Deguang, et al. "Learning Long-Text Semantic Similarity with Multi-Granularity Semantic Embedding Based on Knowledge Enhancement." In: *Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System*, 2021, p. 19-25, DOI: https://doi.org/10.1145/3437802.3437806.

Sonawane, Sheetal S., and Parag Kulkarni. "Concept based document similarity using graph model." *International Journal of Information Technology*, v. 14, n.1, p. 311-322, 2022. DOI: https://doi.org/10.1007/s41870-019-00314-w.

Teller, Virginia. "Book Reviews: Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.". *Computational Linguistics*, v, 26, n. 4, p 638–641, 2000. https://direct.mit.edu/coli/article/26/4/629/1682/On-Coreferring-Coreference-in-MUC-and-Related.

Torkanfar, Navid, and Ehsan Rezazadeh Azar. "Quantitative similarity assessment of construction projects using WBS-based metrics." *Advanced Engineering Informatics*, v. 46, p. 101179, 2020. DOI: https://doi.org/10.1016/j.aei.2020.101179

Yang, Jiaqi, et al. "Measuring the short text similarity based on semantic and syntactic information." *Future Generation Computer Systems*, v. 114, p. 169-180, 2021.

Wang, Jiangyao, et al. "Text similarity calculation method based on hybrid model of LDA and TF-IDF." *In*: *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence.* 2019. DOI: https://doi.org/10.1145/3374587.3374590.

Wang, Jing, et al. "Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: bibliometric study on PubMed." *Journal of Medical Internet Research*, v. 22, n. 1, p. e16816, 2020.

Wang, Keyang, et al. "Comparison between Calculation Methods for Semantic Text Similarity Based on Siamese Networks." *In: 4th International Conference on Data Science and Information Technology*, 2021, p. 389-395. DOI: https://doi.org/10.1145/3478905.3478981.

Wang, Zhongguo, and Bao Zhang. "Chinese Text Similarity Calculation Model Based on Multi-Attention Siamese Bi-LSTM." In: *Proceedings of the 4th International Conference on Computer Science and Software Engineering*, 2021, p. 93-98. DOI: https://doi.org/10.1145/3494885.3494902.

Xiao, Qi, et al. "An unsupervised semantic text similarity measurement model in resource-limited scenes." *Information Sciences*, v. 616, p. 444-460, 2022.

Received: 17/05/2024 Accepted: 02/08/2024