Informology
Online ISSN: 2999 - 9790
Homepage: www.informology.org

Semantic Noise in Information Representation and Retrieval

Alireza Noruzi

Editor-in-Chief, Associate Professor, Informology Center, Marseille, France. E-mail: anoruzi@gmail.com

Article Info

Article type:
Editorial note

***Keywords*:**
semantic noise,
information retrieval,
indexing,
information representation,
natural language processing

ABSTRACT

Objective: In the field of Library and Information Science, the accurate representation and retrieval of information are of utmost importance. Information representation and indexing are critical processes that facilitate the efficient access and utilization of knowledge. However, these processes are not without challenges. One significant issue that arises is "semantic noise", a phenomenon that can distort the meaning of information and hinder effective communication between information retrieval (IR) systems and users. This study aims to explore the concept of semantic noise, its causes, and its implications for information representation and indexing.

Materials and Methods: The current study is primarily theoretical in nature, focusing on the conceptual exploration of semantic noise and its impact on information representation and retrieval. This study investigates the concept of semantic noise, its causes, and its implications for information representation and indexing in the field of library and information science.

Results: The results of the research highlight that semantic noise, caused by irrelevant, ambiguous, or conflicting elements in information representation and indexing, significantly disrupts the clarity and accuracy of information retrieval. Key causes include ambiguity in language and representation, varying contexts, inconsistent terminology, and cultural or linguistic barriers, which collectively introduce complexity and hinder effective communication between information retrieval systems and users. Semantic noise reduces retrieval accuracy, leads to inefficient query processing, and poses challenges for natural language processing (NLP) systems, often resulting in user frustration and diminished trust in information retrieval (IR) systems.

Conclusion: Semantic noise disrupts the clarity and accuracy of information representation and retrieval, leading to inefficiencies, misinterpretations, and user dissatisfaction. Addressing and mitigating semantic noise requires advanced techniques in natural language understanding, such as contextual analysis, semantic search, semantic modeling, and machine learning. These techniques ensure that information retrieval (IR) systems can effectively bridge the gap between user intent and stored data. These findings underscore the critical need for precision in language, standardized terminology, and context-aware approaches to minimize semantic noise and enhance the reliability of information representation and retrieval.

Cite this article: Noruzi, A. (2024). Semantic noise in information representation and indexing. *Informology*, *3*(2), 1-10.

Introduction

Information representation and *indexing* refers to the process of encoding and organizing data and documents in a structured format that can be easily understood and processed by both humans and machines (Fagbola, 2018). It involves selecting appropriate symbols, formats, structures, terms, descriptors, and subject headings to represent data in a meaningful way. The goal of information representation is to capture the essential characteristics and relationships of the data, allowing for efficient storage, retrieval, and management of information. Precisely, *subject indexing* involves assigning descriptive terms, keywords or subject headings to documents or information to facilitate efficient organization and retrieval. In subject indexing, a descriptor is a term used to represent or describe the content of an indexed document. It serves as a keyword or controlled vocabulary term assigned to a document to facilitate efficient and organized retrieval of information (Martins de Medeiros & Medeiros, 2020). Descriptors help in organizing and categorizing documents, making it easier for users to find relevant content within an information retrieval (IR) system.

Sometimes, *semantic noise* occurs in information representation and indexing. Semantic noise in information representation and indexing can occur when the assigned terms do not accurately represent the content or meaning of the documents or data. This can happen due to inconsistencies in terminology, ambiguity in the interpretation of terms, or subjective judgments made during the indexing process (de Jager, 2023). Semantic noise in information representation and subject indexing can lead to difficulties in retrieving relevant information or retrieving irrelevant information when searching for specific content. Efforts are made to minimize semantic noise in subject indexing by using standardized vocabularies, controlled vocabularies, thesauri, ontologies, authority files, or other techniques that ensure consistent and accurate representation of information. When semantic noise is present in information representation and indexing, it becomes challenging for users to retrieve relevant documents during information retrieval processes, as the index terms, descriptors, or subject headings may not adequately reflect the content of the documents. It can lead to reduced precision and recall in search results, making it more difficult for users to find the information they need. It is worth noting that precision measures the accuracy and relevance of the retrieved information, while recall measures the comprehensiveness or completeness of the query results that are successfully retrieved (López-Herrera et al., 2009; Owais et al., 2007).

Semantic noise can occur due to factors like ambiguity, language barriers, ambiguous or misleading verbal symbols (e.g., thesaurus descriptors or subject headings), cultural differences, technical jargon, or errors in data transmission or processing (Kadir et al., 2018). This noise can lead to misunderstandings, misinterpretations, or incorrect conclusions when trying to extract

meaning from the represented information or when try to retrieve information. Effective information representation techniques aim to minimize semantic noise to ensure accurate and clear understanding of the intended message. For instance, if a descriptor is vague, open to interpretation, poorly chosen by indexers (and indexing systems), or lacks specificity and publication warrants, it can lead to misunderstanding and confusion by users.

Semantic noise can also arise due to factors such as: 1. ***Ambiguity***: It occurs when a descriptor or term has multiple meanings or interpretations. For example, the term "*Java*" can refer to the programming language or to the Indonesian island. This ambiguity can lead to confusion and inaccurate representation when indexing information and can be subject to misinterpretation when searching for specific information. 2. ***Synonymy***: It refers to different terms or descriptors that have similar meanings. For instance, "car" and "automobile" are synonyms. When multiple synonyms are used separately to index documents, it can result in redundancy and inefficiency in information retrieval (IR) systems (Nasir et al. (2019). 3. ***Polysemy***: It involves a descriptor or word having multiple related meanings. For example, the term "bank" can refer to a financial institution, the edge of a river, or a stack of money (Noruzi, 2006). 4. ***Granularity***: descriptors and subject headings can be too broad or too specific, causing problems in information representation and retrieval. If a descriptor or subject heading is overly general, such as "Science," it may encompass a wide range of topics and make it challenging to find specific information. Conversely, if a descriptor or subject heading is too detailed or specific, it may not accurately represent the content of a document and limit discoverability. If documents are not properly disambiguated or differentiated, the indexing process may introduce semantic noise and make it challenging to retrieve relevant information. To minimize semantic noise, it is important to choose descriptors, subject headings, and keywords carefully, ensuring they are precise, unambiguous, and relevant to the intended context. Providing additional context or examples can also help reduce confusion and ensure a shared understanding between the sender and recipient of the information.

As we navigate the vast landscape of digital information, acknowledging and addressing semantic noise is crucial for maintaining the integrity and effectiveness of information representation, especially in the artificial intelligence (AI) field and *generative information retrieval* (Gen-IR). By adopting standardized terminology, contextual clarifications, and advanced natural language processing (NLP) techniques, we can pave the way for clearer, more accurate information representation and retrieval, ultimately enhancing the quality of information and fostering a more robust information retrieval (IR) ecosystem.

This study aims to explore the concept of semantic noise, its implications, and strategies to mitigate its effects in information representation and subject indexing.

Materials and Methods

The current study is primarily theoretical in nature, focusing on the conceptual exploration of semantic noise and its impact on information representation and retrieval. This study investigates the concept of semantic noise, its causes, and its implications for information representation and indexing in the field of library and information science.

Results

Causes of Semantic Noise

Semantic noise occurs when irrelevant, ambiguous, or conflicting elements infiltrate the representation of information, obscuring the intended message. This noise can manifest in various forms, such as unclear language, inconsistent terminology, or the inclusion of irrelevant details. Ultimately, it introduces a layer of complexity that disrupts the seamless exchange of meaningful information. The causes of semantic noise vary depending on the nature of the information source, such as text or image data. These causes include, but are not limited to, the following:

1. **Ambiguity in Language and Representation**: Semantic noise often stems from the inherent ambiguity present in language. Words or phrases may have multiple meanings, leading to confusion and misinterpretation (Belfarhi, 2021). Resolving this issue requires precision in language and a standardized use of terminology. It also requires precision in both encoding (the delivery of the message by authors) and decoding (the representation of the message by indexers and indexing systems). In addition, semantic noise introduces ambiguity in how information is represented, indexed, categorized, or tagged, making it harder to organize and retrieve systematically. This can affect metadata, keywords, and other organizational structures that rely on clear and consistent meaning. Indexing is not an entirely objective process. In addition, human indexers inevitably interpret the content of a document differently, influenced by their own knowledge, biases, and perspectives. This inherent subjectivity can introduce inconsistencies and semantic noise into the indexing process. Moreover, language is dynamic, and the meanings of words evolve over time. Terms that were once widely understood may become obsolete or acquire new meanings, potentially creating semantic noise in older indexed materials.

2. **Varying Contexts**: Information representation may lose its intended meaning when viewed in different contexts and settings. What makes sense in one subject domain may become ambiguous or irrelevant when transferred to another, thus contributing to semantic noise (Liu et al., 2024). This highlights the importance of context in preserving the accuracy and relevance of information.

3. **Inconsistency in Terminology**: The divergent use of terminology within a dataset or across different sources introduces inconsistencies (Le & David Jeong, 2017). These discrepancies can lead to confusion, making it difficult to establish a cohesive understanding of the information being represented.

4. **Reduced Retrieval Accuracy**: In information retrieval systems, semantic noise can cause mismatches between user queries and the stored data, leading to irrelevant or incomplete results (Gulati & Garg, 2015). For example, synonyms, homonyms, or context-dependent terms may not align properly, reducing the system's ability to retrieve precise information. In other words, semantic noise can manifest in various forms, including synonyms, homonyms, polysemous words, misspellings, and contextually ambiguous terms. Such noise can lead to mismatches between user prompts/queries and indexed content, resulting in suboptimal retrieval outcomes.

5. **User Frustration and Reduced Trust**: When users encounter irrelevant or incomplete results due to semantic noise, their trust in the IR system may diminish. This can lead to frustration and a perception that the IR system is unreliable or ineffective.

6. **Cultural and Linguistic Barriers**: Semantic noise is often amplified in multilingual or multicultural contexts (such as Large Language Models (LLMs)), where differences in language use, idioms, or cultural references can further complicate information representation and retrieval (Shan et al., 2025). Information systems often serve diverse user populations with varying cultural and linguistic backgrounds. Consequently, terms and concepts that are clear in one culture may be misunderstood or misinterpreted in another, leading to semantic noise.

7. **Complexity of Subject Matter**: Some subjects are inherently complex and require specialized terminology. If human indexers (and indexing systems) and users do not share the same level of expertise, misunderstandings can arise, leading to semantic noise.

8. **Inefficient Prompt and Query Processing**: Queries may fail to retrieve the intended information if semantic noise causes a disconnect between the user's intent and the system's understanding of the query, even in *generative information retrieval* (Gen-IR). This can lead to inefficiencies in prompt and search processes and require additional effort to refine prompts or queries (Zhang et al., 2024).

9. **Impact of Semantic Noise on NLP**: Semantic noise also poses challenges for NLP systems, such as AI search engines or AI chatbots, which rely on understanding context and meaning. Ambiguities in language, cultural differences, '*culture influence*,' or jargon can hinder the Gen-IR system's ability to process and retrieve information accurately (Suadamara et al., 2010). Leveraging NLP techniques, such as part-of-speech tagging,

named entity recognition, and word sense disambiguation, can aid in disentangling semantic noise. By analyzing the context and semantics of words, NLP algorithms can enhance the accuracy of information representation and indexing.

Strategies to Address and Mitigate Semantic Noise

To mitigate semantic noise in information representation, various techniques are employed:

1. **Standardization of Terminology**: Using predefined and standardized vocabularies helps eliminate ambiguity and promotes consistency in descriptor usage. These vocabularies can include hierarchical structures and relationships that ensure accurate representation and indexing. Establishing and adhering to standardized terminology—such as by leveraging large language models (LLMs) and multilingual models—across datasets and communication channels can significantly reduce semantic noise, ensuring a consistent and accurate understanding of information. In addition, thesauri provide a controlled vocabulary that helps address synonymy. They offer alternative descriptors or terms that can be used for information representation and indexing, reducing redundancy and improving the precision of information retrieval. Thus, the semantic noise can be reduced through the use of controlled vocabularies, thesauri, contextual analysis, and computational methods. These approaches aim to minimize ambiguity, synonymy, and polysemy, thereby improving the efficiency and accuracy of information retrieval. It is worth noting that the "Scopus AI" now utilizes its internal thesaurus to verify the vocabulary of a user's query before generating an answer through its generative AI system.

2. **Contextual Analysis and Clarifications**: Analyzing the context of a document, such as its surrounding text or other associated metadata, can help disambiguate descriptors and index terms. By understanding the document's content and intended meaning, better indexing decisions can be made to minimize semantic noise. Providing context-specific clarifications within information representations helps mitigate the impact of semantic noise, allowing users to interpret information within its intended context.

3. **Computational Methods and Advanced Data Processing Techniques**: Advanced algorithms and machine learning techniques can be applied to automatically extract or propose descriptors and index terms for information representation and indexing. These methods can help identify relationships, patterns, and disambiguate descriptors and index terms based on large datasets and context. Leveraging advanced data processing techniques—such as NLP/LLMs and machine learning algorithms and models—can help identify and filter out semantic noise in large datasets.

4. **Increased Complexity in Information Systems**: To mitigate semantic noise, IR systems may require more sophisticated algorithms, such as semantic search or machine learning models, to better understand context and meaning.
5. **User Feedback and Relevance Feedback Mechanisms**: Taking into account user feedback and conducting user studies can help identify and rectify semantic noise issues. Engaging users in providing feedback on search results can help in refining information representation and indexing. '*Relevance feedback'* (Nasir et al., 2019) mechanisms enable the system to adapt and learn from user interactions, consequently reducing the impact of semantic noise, especially in AI search engines.
6. **Ontologies and Knowledge Graphs**: Utilizing ontologies and knowledge graphs can provide structured representations of concepts and their relationships, mitigating semantic ambiguity. By incorporating semantic relationships and hierarchical classifications, ontologies offer a powerful means to combat noise in information representation and indexing.

Conclusion

Semantic noise in information representation and indexing refers to any interference or distortion in the meaning of information that occurs during the process of information representation and indexing (e.g., the irrelevant or misleading author-supplied keywords, terms, subject headings, descriptors, queries, and prompts), which can significantly impact information representation and retrieval and also affects the meaning or interpretation of the represented information. It occurs when there is a mismatch or inconsistency between the intended meaning of the information (descriptors or subject headings assigned to documents by human indexers or machines) and how it is actually conveyed or understood. It generally occurs when the assigned terms or descriptors fail to accurately represent the content or meaning of a document. Semantic noise can occur due to various factors such as ambiguity in language, subjective interpretations of indexers, improper selection of index terms, or mismatch between the vocabulary used by the indexer (and indexing systems) and the users searching for information.

The results of the study highlight that semantic noise, caused by irrelevant, ambiguous, or conflicting elements in information representation, significantly disrupts the clarity and accuracy of information retrieval. Key causes include ambiguity in language and representation, varying contexts, inconsistent terminology, and cultural or linguistic barriers, which collectively introduce complexity and hinder effective communication between users and information

systems. Semantic noise reduces retrieval accuracy, leads to inefficient query/prompt processing, and poses challenges for NLP/AI systems, often resulting in user frustration and diminished trust in IR systems. Additionally, mitigating semantic noise requires advanced techniques, such as semantic search and machine learning models, which increase the complexity and cost of developing and maintaining IR systems. These findings underscore the critical need for precision in language, standardized terminology, and context-aware approaches to minimize semantic noise and enhance the reliability of information representation and retrieval.

Despite advancements in tackling semantic noise, challenges persist in achieving robust and efficient information representation and indexing. Multilingual contexts, evolving language usage, and domain-specific terminologies present ongoing challenges in mitigating semantic noise. Future research endeavors may focus on developing adaptive, context-aware indexing systems and refining NLP/LLMs techniques to address these challenges.

Data Availability Statement

Not applicable.

Ethical considerations

The author avoided from data fabrication and falsification.

Funding

Not applicable.

References

Belfarhi, K. (2021). Rethinking language ambiguity beyond the semantico-pragmatic interface. *Folia Linguistica et Litteraria*, (34), 211-233. https://www.researchgate.net/publication/351384801

de Jager, S. (2023). Semantic noise and conceptual stagnation in natural language processing. *Angelaki: Journal of the Theoretical Humanities*, 28(3), 111-132. https://doi.org/10.1080/0969725X.2023.2216555

Fagbola, O. O. (2018). Indexing and abstracting as tools for information retrieval in digital libraries: A review of literature. In Information Resources Management Association (Ed.), *Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications* (pp. 905-927). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-5225-5191-1.ch039

Gulati, N., & Garg, A. (2015, October). A proposed framework to optimize the query by filtering noise using Semantic information processing. In *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)* (pp. 1331-1338). IEEE. https://doi.org/10.1109/ICGCIoT.2015.7380673

Kadir, R. A., Yauri, R. A., & Azman, A. (2018). Semantic ambiguous query formulation using statistical Linguistics technique. Malaysian Journal of Computer Science, 48-56. https://ejournal.um.edu.my/index.php/MJCS/article/view/15487, https://doi.org/10.22452/mjcs.sp2018no1.4

Le, T., & David Jeong, H. (2017). NLP-based approach to semantic classification of heterogeneous transportation asset data terminology. *Journal of Computing in Civil Engineering*, *31*(6), 04017057. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000701

Liu, Y., Jiang, S., Zhang, Y., Cao, K., Zhou, L., Seet, B. C., Zhao, H., & Wei, J. (2024). Extended context-based semantic communication system for text transmission. *Digital Communications and Networks*, 10(3), 568-576. https://doi.org/10.1016/j.dcan.2022.09.023

López-Herrera, A.G. Herrera-Viedma, E., & Herrera, F. (2009). Applying multi-objective evolutionary algorithms to the automatic learning of extended Boolean queries in fuzzy ordinal linguistic information retrieval systems. *Fuzzy Sets and Systems*, 160(15), 2192-22051. https://doi.org/10.1016/j.fss.2009.02.013

Martins de Medeiros, G., & Medeiros, M.B.B. (2020). Subject indexing in arhival documents: Analysis of international definitions base on literature systematic review [A indexação de assunto em documentos arquivísticos: Análise das definições internacionais com base na revisão sistemática da literatura]. *Revista Digital de Biblioteconomia e Ciencia da Informacao*, 185(1), Article Number e020006.

Nasir, J. A., Varlamis, I., & Ishfaq, S. (2019). A knowledge-based semantic framework for query expansion. Information Processing and Management, 56(5), 1605-1617. https://doi.org/10.1016/j.ipm.2019.04.007

Noruzi, A. (2006). Folksonomies: (un) controlled vocabulary?. *Knowledge Organization*, 33(4), 199-203. https://www.nomos-elibrary.de/10.5771/0943-7444-2006-4-199.pdf

Owais, S.S.J., Snášel, V., & Krömer, P. (2007). Grow up precision recall relationship curve in IR system using GP and fuzzy optimization in optimizing the user query. *Neural Network World*, 17(4), 295-309. https://dspace.vsb.cz/handle/10084/63505

Shan, X., Xu, Y., Wang, Y., Lin, Y.S., & Bao, Y. (2025). Cross-cultural implications of large language models: An extended comparative analysis. In: Coman, A., Vasilache, S., Fui-Hoon Nah, F., Siau, K.L., Wei, J., Margetis, G. (eds.), *HCI International 2024 – Late Breaking Papers. HCII* 2024. *Lecture Notes in Computer Science*, Vol. *15375*, pp. 106-118. Springer, Cham. https://doi.org/10.1007/978-3-031-76806-4_8

Suadamara, R., Werner, S., & Hunger, A. (2010, June). Culture influence on human computer interaction-cultural factors toward user's preference on groupware application design. In *International Conference on Enterprise Information Systems* (Vol. 2, pp. 186-191). ScitePress. https://www.scitepress.org/PublishedPapers/2010/29727/

Zhang, W., Liu, Z., Wang, K., & Lian, S. (2024). Query expansion and verification with large language model for information retrieval. In: Huang, D.S., Si, Z., & Zhang, C. (eds.) *Advanced Intelligent Computing Technology and Applications. ICIC 2024. Lecture Notes in Computer Science*, *14878*, pp. 341-351, Springer, Singapore. https://doi.org/10.1007/978-981-97-5672-8_29