LUMEN: Leveraging Large Language Models for Dynamic Ontologies in Wind Energy Domain Analysis

Andrea Lops[1], Serena Sassi[2]

[1]Politecnico di Bari, Dipartimento di Ingegneria Elettrica e dell'Informazione, Bari, 70125, Italy

[2]Università degli Studi di Bari Aldo Moro, Dipartimento di Ricerca e Innovazione Umanistica, Bari, 70121, Italy

Abstract

This study introduces LUMEN (Latent Understanding through Modeling Embeddings in Natural language), a structured approach leveraging Large Language Models (LLMs) and semantic embeddings to construct a hierarchical thesaurus for the wind energy domain. Using a domain-specific corpus generated via Sketch Engine, LUMEN applies a three-step methodology: corpus creation, semantic label identification through LLM-based analysis, and hierarchical term classification via embedding-based semantic similarity calculations. We outline our machine learning configuration, including the embedding techniques and similarity metrics employed. Results indicate that LUMEN can capture nuanced subdomains and semantic interrelations within wind energy, despite occasional misclassification issues. Future research directions include systematic benchmarking against established ontology-building tools and multilingual adaptation to broaden applicability.

Keywords

Terminology, Natural Language Processing, Large Language Model, Domain Analysis, Domain Tree, Wind Power Research

1. Introduction

The advancement of Natural Language Processing (NLP) techniques, particularly through the integration of Large Language Models (LLMs) and deep learning models, has opened new avenues for exploring and structuring knowledge domains. These technologies have demonstrated remarkable potential in understanding complex linguistic patterns, contextual relationships, and domain-specific terminologies. In this study, we leverage these advancements to construct a domain tree for wind energy, derived from a specialized corpus created using the Sketch Engine software. Our approach, named LUMEN (Latent Understanding through Modeling Embeddings in Natural language), employs LLMs and contextual embeddings to uncover latent subdomains and their interrelations, offering a comprehensive representation of the wind energy domain.

The use of LLMs allows a semantic analysis of the corpus [1], enabling the identification of key labels that define the domain's hierarchical structure. These labels are organized in a structured JSON format, providing a blueprint for subsequent classification tasks. Deep learning techniques, particularly those involving embeddings, are then utilized to compute semantic similarity between labels and corpus terms. This process facilitates the automatic distribution of terms in the domain tree, ensuring an accurate and scalable classification across multiple levels of granularity.

By applying these techniques, the latent structure of the wind energy domain is brought out, delineating distinct subdomains and highlighting their interconnections. This approach not only facilitates the construction of data-driven ontologies but also uncovers unexpected relationships within the field, enriching the understanding of its multidisciplinary nature. This study also aims to demonstrate the potential of these technologies, offering a scalable and adaptable framework for domain analysis while advancing ontological studies and fostering interdisciplinary collaboration in complex fields.

4th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2025, June 19-20, 2025, Thessaloniki, Greece.

andrea.lops@poliba.it (A. Lops); serena.sassi@uniba.it (S. Sassi)
https://sisinflab.poliba.it/people/andrea-lops (A. Lops); https://www.linkedin.com/in/serena-sassi (S. Sassi)
0009-0006-7461-5368 (A. Lops); 0009-0008-4894-0022 (S. Sassi)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

2. Genesis of the Issue

This research emerged from a series of theoretical and technical challenges encountered during some investigations on the terminological field of wind energy. The initial obstacle stemmed from the need to comprehend the status of the "domain" notion, a central concept to our study. The difficulty lies not only in defining the boundaries of the wind energy domain but also in understanding how to position this notion within a broader context characterized by rapidly advancing technologies [2, 3, 4]. The wind energy sector, like all renewable energy sources, is inherently dynamic and offers vast potential for technological and conceptual advancements. For these reasons, its dynamic nature poses significant challenges in defining clear boundaries. In particular, its fluidity and interdisciplinary nature further complicate this task, as the domain encompasses a broad range of disciplines, practices, and discourses, all contributing to a complex and ever-evolving terminological field.

2.1. Rethinking the Boundaries of Environmental Domains

The concept of "domain" occupies a central role across various academic fields, from linguistics and philosophy to sociology and information sciences [5]. Within the field of terminology, it serves as a foundational principle used to organize and classify scientific and technical terms, thereby structuring knowledge coherently and systematically. However, this traditional conception of the "domain", grounded in rigid classifications and boundaries, warrants reconsideration in light of contemporary shifts in both knowledge production and linguistic practices. Although such classifications may have once sufficed in contexts characterized by less terminological proliferation, the increasing complexity of science and technology, alongside the hybridization of specialized and general discourse, has rendered this notion increasingly problematic and open to debate.

In this context, it is essential to critically assess the relevance of the "domain" as it is presently conceived, especially in fields like environmental studies, where technical terms frequently overcome specialized spheres and infiltrate general language and public discourse. As Delavigne emphasizes, the attempt to precisely delineate a specific domain often leads to significant challenges, particularly in addressing complex subjects such as environmental issues [6]. The environment, by its very nature, is a multidimensional field that involves continuous interaction among various disciplines, relying on a vast network of interconnections with technical, scientific, social, and cultural discourses [7, 8]. As Myerson and Rydin observe, "in academic terms, 'environment' belongs to every discipline and none" [9].

These interactions transcend the connections between various domains and levels of expertise. Even highly specialized discussions on this matter inevitably draw on knowledge from a broad spectrum of associated fields [10, 11, 12]. Delavigne highlights not only the vast and interconnected network of disciplines, practices, discourses, and techniques that constitute environmental studies but also the considerable disparities in qualitative equivalence across these components [13]. Indeed, no domain, subdomain, or sub-subdomain within the environmental field can be addressed without acknowledging its intrinsic diversity and interdisciplinary nature. As Sager contends, "In practice no individual or group of individuals possesses the whole structure of a community's knowledge; conventionally, we divide knowledge up into subject areas, or disciplines, which is equivalent to defining subspaces of the knowledge space" [14].

In a field where the concept of "domain" remains fundamentally ambiguous and open to various interpretations, several critical questions naturally arise. How can these expansive and complex bodies of knowledge be segmented in a methodologically sound and pertinent manner? What methodologies can be adopted to quantify and delineate these domains, which are often fluid and interrelated? How can one adequately represent the meaning and scope of these diffuse networks of knowledge, practices, and scientific communities that collectively contribute to intellectual production? How can domain trees be constructed in an era when the boundaries between various fields are increasingly "permeable," and their overlap is both evident and significant [2, 3]?

Given these theoretical and technical challenges, we decided to integrate terminology and NLP to try a different approach to terminological analysis that can address the complexities inherent in domains. By combining traditional terminological methods with NLP techniques, we sought to develop a more systematic and scalable method to categorize and analyze terminology. This hybrid approach would enable us to create terminological resources, facilitating a deeper understanding of the semantic relationships within, in this case, the wind energy domain. NLP techniques, by processing large corpora of text and identifying complex patterns in language, offered the potential for potentially efficient categorization of terms. This methodological shift allowed us not only to enhance the precision of our terminological cataloging but also to streamline the overall process, making it more adaptable to ongoing developments within the field.

2.2. Enhancing Terminology through Ontologies

Another key aspect of our study involves the relationship between terminology and ontologies. Ontologies, as formal representations of knowledge within a particular domain, have long been recognized for their ability to structure information in a way that reflects both semantic relationships and conceptual hierarchies [15, 16, 17]. In the context of domain-specific terminological analysis, ontologies play a critical role in organizing and classifying terms based on their interconnections and shared meanings. The systematic classification of terms within a particular field has long been essential for structuring knowledge in a coherent and accessible way [18]. An ontology provides a formalized structure that organizes terms into categories such as domains, subdomains, and concepts, highlighting their relationships and interdependencies [19, 20].

The intersection of NLP and ontological studies presents an exciting opportunity for advancing the systematic representation of knowledge, especially in complex and evolving domains. The integration of NLP techniques offers, in fact, a different approach to this challenge. By leveraging LLMs and embedding-based methods, we were able to uncover latent relationships between terms and construct a dynamic, data-driven representation of a domain adaptive to future developments. By combining terminological analysis with ontological principles, we created a structured framework to organize the wind energy domain. This framework includes domain labels, subdomain classifications, and semantic relationships, which are essential for the categorization of terms.

By employing NLP techniques, it is also possible to uncover latent knowledge structures that would be difficult to identify through manual analysis alone. In fact, in our study, LUMEN demonstrated its ability to detect and categorize multiple domains beyond the strictly technical scope of wind energy. Despite the field's highly specialized nature, LUMEN identified many distinct macro-domains: the environmental domain (encompassing terminology related to flora, fauna, etc.), the health domain, the political and economic domain, the social and cultural domain, and the juridical domain, each one with their subdomains and key terms. The integration of terminology studies with NLP-based ontological frameworks offers significant potential for advancing domain analysis, facilitating deeper interdisciplinary collaboration, and a more comprehensive understanding of emerging fields. We believe that this methodology can also be extended to other complex domains, allowing researchers and practitioners to continuously refine and expand their understanding of rapidly changing fields.

3. Our Approach

This section outlines the methodology used for latent subdomain identification. Our approach, illustrated in Figure 1, consists of three main stages: corpus creation using Sketch Engine, domain label identification via LLM, and term classification based on semantic similarity.

These stages contribute to the creation of a hierarchical domain tree that facilitates a comprehensive understanding of the field.

3.1. Step 1: Creation of an English Corpus

The first step of our methodology involves the compilation of a domain-specific corpus [21, 22, 23], leveraging Sketch Engine's [24] advanced text retrieval features. To create this corpus, we employed

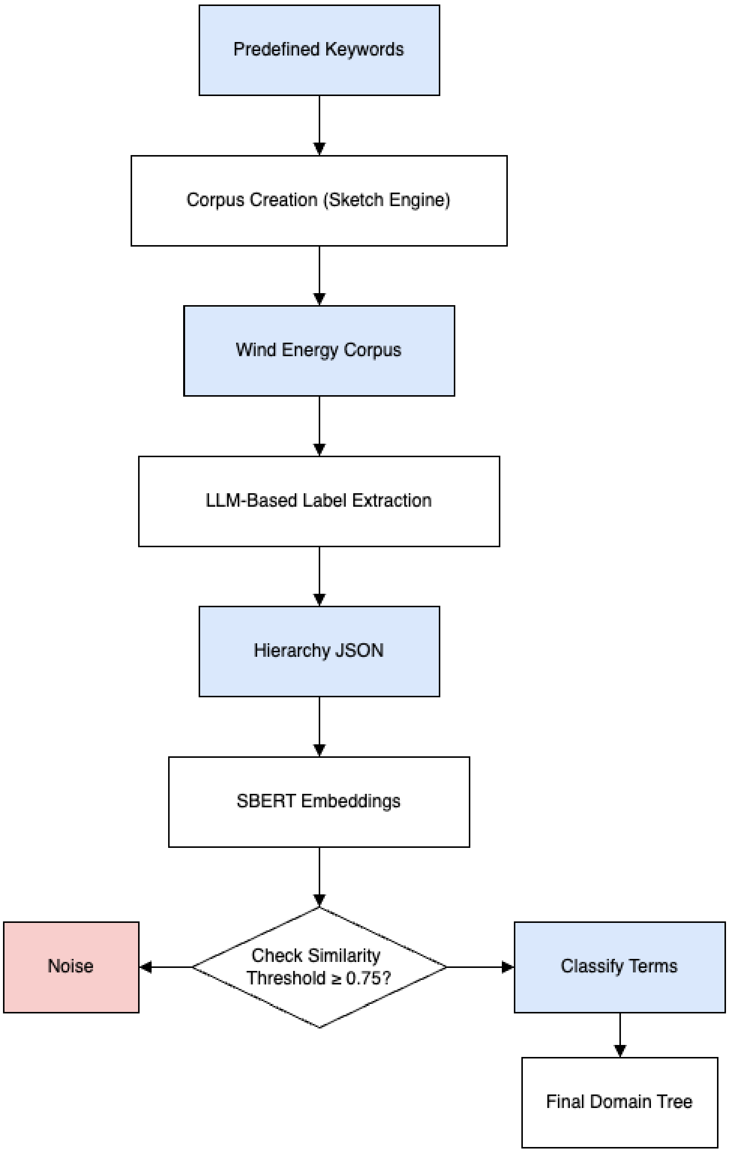
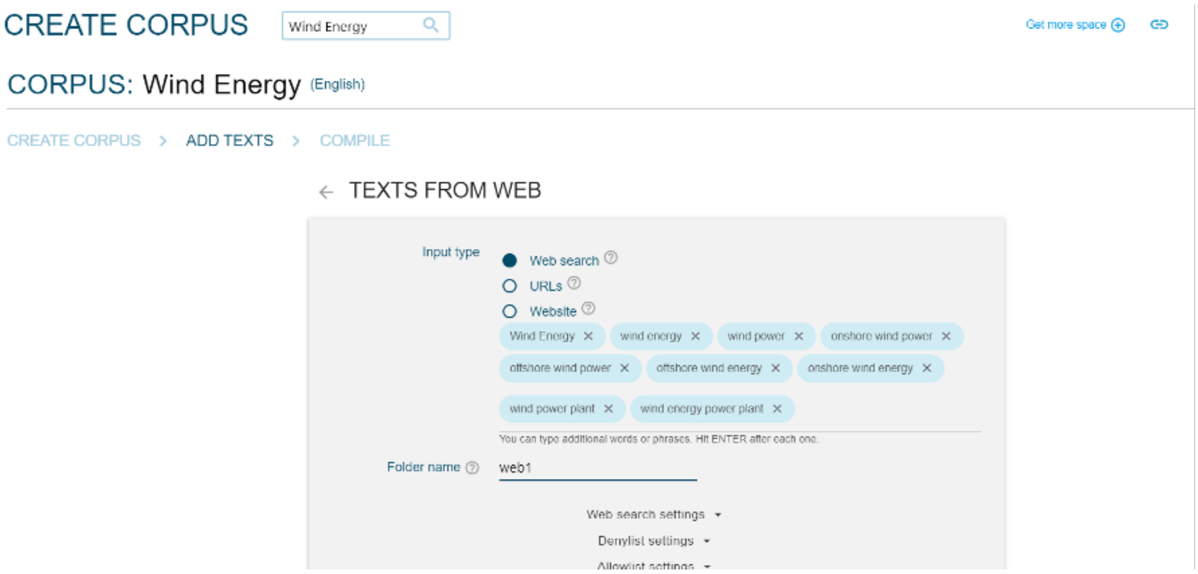


Figure 1: Overview of LUMEN

Sketch Engine's "Find texts on the web" functionality, a tool designed to automate the collection of textual data from online sources based on a predefined set of keywords. This semi-automatic approach ensures a streamlined and scalable process for assembling a corpus that accurately reflects the terminological landscape of wind energy.

Figure 2 presents the set of keywords used for corpus generation. The selection of these keywords was guided by a principle of neutrality to minimize biases in data collection. Specifically, the keywords chosen—*"wind energy", "wind power", "onshore wind power", "onshore wind energy", "offshore wind power", "offshore wind energy", "wind power plant",* and *"wind energy power plant"*—were primarily two-word expressions. This decision was made to ensure broad coverage while avoiding over-reliance on highly specialized terms that could introduce distortions in the corpus composition. Moreover, the selection was informed by an effort to capture both general and specific terms relevant to the wind energy domain, ensuring a balance between comprehensiveness and precision.

As shown in Figure 3, following the keyword-based retrieval process, Sketch Engine compiled a corpus comprising 287 documents, exclusively in English. This corpus spans a total of 1,586,074 words, encompassing a diverse array of textual sources relevant to the wind energy sector, ensuring the capture of a large spectrum of discourse surrounding the field. Through Sketch Engine's automated



Wind Energy

CORPUS: Wind Energy (English)

Input type
● Web search
○ URLs
○ Website

Wind Energy ✕ wind energy ✕ wind power ✕ onshore wind power ✕ offshore wind power ✕ offshore wind energy ✕ onshore wind energy ✕ wind power plant ✕ wind energy power plant ✕

You can type additional words or phrases. Hit ENTER after each one.

Folder name: web1

Figure 2: Representing the keywords used to retrieve texts from the Web.

terminology extraction features, a subset of 100,000 terms was identified as domain-specific, serving as the foundational dataset for subsequent classification and hierarchical organization.

The creation of this corpus is a critical step in our methodology, as it provides a rich linguistic dataset from which domain labels can be inferred, and semantic relationships among terms can be analyzed. The extracted terms represent a comprehensive cross-section of the wind energy domain, encompassing technical concepts, industry-specific terminology, and related expressions that contribute to the identification of subdomains within the field.

3.2. Step 2: Domain Label Identification via LLMs

The second stage involves using a state-of-the-art LLM (specifically 'gpt-4o'[1]) to analyze the corpus and identify a hierarchical labeling structure. This step translates unstructured textual data into a structured hierarchical format.
The procedure includes:

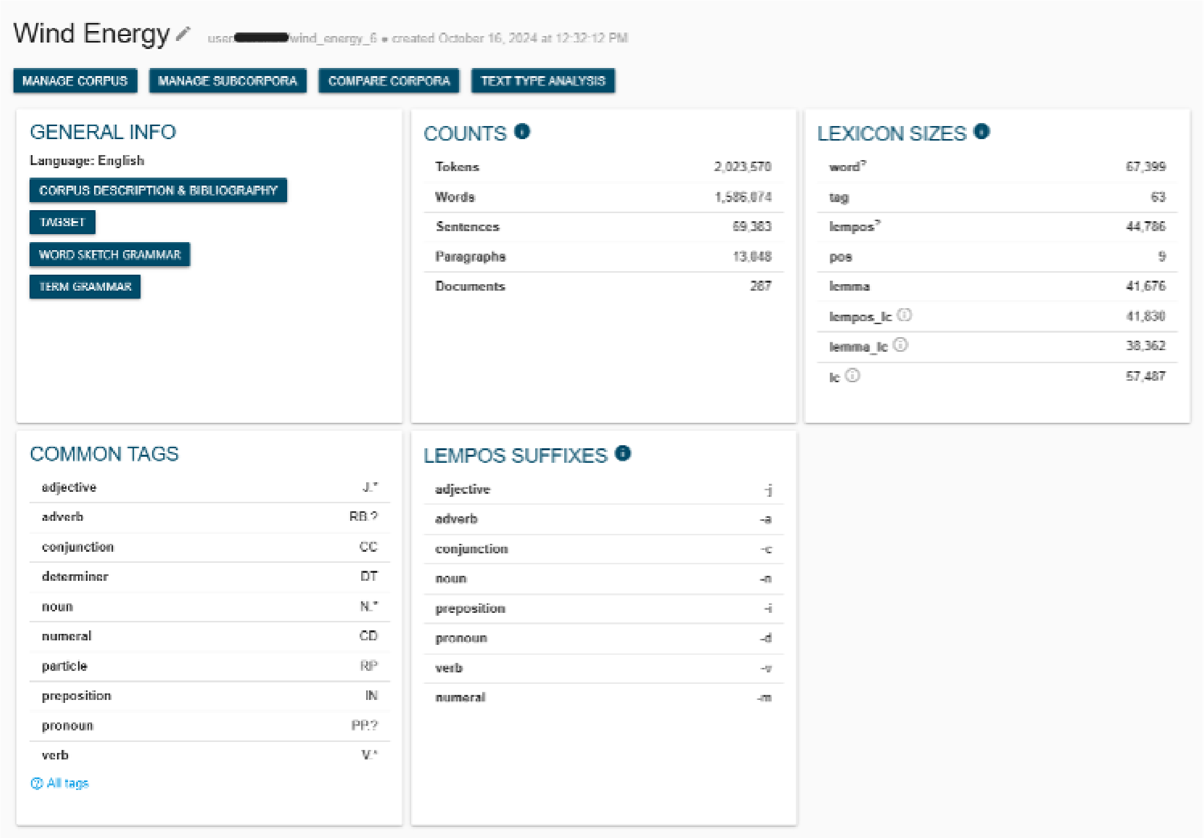
1. **Pre-processing and Prompt Engineering:** The text corpus is cleaned up and segmented into manageable documents, and prompt engineering techniques [25] are used to instruct the LLM to extract representative domain and subdomain labels taking into account the entire distribution of terms. Figure 4 shows the general structure of the prompt.
2. **Semantic Analysis and Label Extraction:** 'gpt-4o' generates potential labels by recognizing semantic clusters within the corpus, outputting structured labels hierarchically categorized into three levels (domain, subdomain, and sub-subdomain).
3. **Hierarchical JSON Generation:** The extracted labels are serialized into a hierarchical JSON format, constituting a structured thesaurus ready for semantic classification.

The hyperparameters used for 'gpt-4o' were: temperature set at 0 (favoring more deterministic outputs), top-p sampling at 0.95 (ensuring diversity while maintaining relevance), and a token limit of 128,000 to manage processing constraints.

3.3. Step 3: Term Classification Using Semantic Similarity

The third step uses embedding-based semantic similarity for term classification in the hierarchical structure established previously. The workflow involves:

[1] https://openai.com/index/hello-gpt-4o/



Wind Energy user[redacted]wind_energy_6 • created October 16, 2024 at 12:32:12 PM

GENERAL INFO

Language: English

COUNTS

Tokens	2,023,570
Words	1,586,074
Sentences	69,383
Paragraphs	13,048
Documents	287

LEXICON SIZES

word?	67,399
tag	63
lempos?	44,786
pos	9
lemma	41,676
lempos_lc	41,830
lemma_lc	38,362
lc	57,487

COMMON TAGS

adjective	J.*
adverb	RB.?
conjunction	CC
determiner	DT
noun	N.*
numeral	CD
particle	RP
preposition	IN
pronoun	PP.?
verb	V.*

LEMPOS SUFFIXES

adjective	-j
adverb	-a
conjunction	-c
noun	-n
preposition	-i
pronoun	-d
verb	-v
numeral	-m

Figure 3: General data about our corpus.

1. **Embedding Generation:** Terms and hierarchical paths (concatenated labels, e.g., "Energy > Renewable Sources") are converted into vector embeddings using Sentence-BERT (SBERT) [26], chosen due to its effectiveness in capturing semantic nuances. Specifically, we employed the SBERT model *all-mpnet-base-v2*[2] with a vector dimension of 768, ensuring balanced performance between accuracy and computational efficiency.
2. **Similarity Calculation:** Cosine similarity [27] between the embeddings of each term and each hierarchical path is computed. This metric quantitatively measures the semantic closeness between terms and labels. We also explored alternative distance metrics (e.g., Euclidean distance) during an initial evaluation phase; however, cosine similarity yielded consistently higher alignment with expert judgments for this domain-specific setting.
3. **Threshold-based Classification:** Terms exceeding an experimentally determined similarity threshold (initially set at 0.75, optimized via grid search on a validation set) are automatically classified under corresponding labels. Terms below this threshold are provisionally categorized as "noise" and flagged for further human review. In addition to adjusting the threshold, we are developing a mechanism for incremental reclassification, whereby a term initially labeled as "noise" can be re-evaluated if the subsequent context or expert feedback indicates it is indeed relevant to a specific subdomain.

To ensure scientific rigor, a preliminary subset of classified terms was assessed following a terminological evaluation of the domain, providing initial feedback that enabled the estimation of precision and recall on a limited portion of the corpus (approximately 500 terms). In this context, *precision* refers to the proportion of terms labeled as relevant by the relevant system (i.e., the percentage of correct classifications), while *recall* measures the proportion of truly relevant terms that the system successfully

[2] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Prompt Structure

You are an experienced linguist with a deep understanding of natural language processing. You will be provided with a list of 100,000 lemmas. Your task is to return to me a JSON with all the labels of possible domains and sub-domains that you can unearth in this list. Be detailed, return me a JSON with as many domains and as many sub-domains as possible so as to include as many lemmas as possible. Take your time to answer don't be rushed, think it through. Always remember that I just want labels, don't return me lemmas. Follow this example:

```
{
"total_labels":[
    {
      "label": "label domain 1",
      "sub-labels":[
        "Label sub-domain 1":[ "Label sub-sub-domain 1", "Label sub-sub-domain 2", ...],
        "Label sub-domain 2": ["Label sub-sub-domain 1", "Label sub-sub-domain 2", ...],
        ...
      ]}
    },
    {
      "label": "Label domain 2",
      "sub-labes": {[
        "Label sub-domain 1": ["Label sub-sub-domain 1", "Label sub-sub-domain 2", ...],
        "Label sub-domain 2": ["Label sub-sub-domain 1", "Label sub-sub-domain 2", ...],
        ...
      ]}
    },
    {
      "label": "Label domain 3",
      "sub-labes": {[
        "Label sub-domain 1": ["Label sub-sub-domain 1", "Label sub-sub-domain 2", ...],
        "Label sub-domain 2": ["Label sub-sub-domain 1", "Label sub-sub-domain 2", ...],
        ...
      ]}
    },
    ...
    ]
}
```

Figure 4: Contracted structure of the prompt used for extracting representative domain labels.

identified (i.e., how many relevant items are captured in total). The results indicated a precision of about 0.82 and a recall of 0.78. Recognizing the preliminary nature of these findings, subsequent evaluations will expand the sample size, incorporate more diverse terms, and include comparative analysis with established ontology framework extraction as well as input from additional experts.

3.4. Output

The final output of the LUMEN methodology is a hierarchical JSON structure enriched with terms classified under each label node, offering a comprehensive representation of the wind energy domain and its subdomains. A selected extract from this structure is shown in Figure 5, this hierarchical representation allows the continued expansion of the *thesaurus* as new terms and concepts emerge within the domain.

In our experimentation, LUMEN successfully identified nine primary subdomains within wind energy:

Output

```
{
    "Environment" : [
        "Environmental Impact": [
            "On Fauna": [
                "farm development",
                "habitat loss",
                "cumulative impact",
                ...
            ],
            "On Flora": [
                "impact on the marine environment",
                "perspective on marine environmental impacts",
                "public land",
                ...
            ],
            "Visual": [...],
            ...
        ],
        ...
    ],
    "Energy": [...],
    "Geography" : [...],
    "Economy" : [...],
    "Technology" : [...],
    "Society" : [...],
    "Safety" : [...],
    "Fauna" : [...],
    "Research" : [...],
}
```

Figure 5: Excerpt from the hierarchical JSON output generated by LUMEN. In this example, environment-related labels, sub-labels, and terms are shown.

Environment, *Energy*, *Geography*, *Economy*, *Technology*, *Society*, *Safety*, *Fauna*, and *Research*. Each subdomain contains one or more sub-labels, ensuring that users can navigate from broader concepts (e.g., “Environmental Impact”) to increasingly granular topics (e.g., “On Fauna” or “On Flora”). By employing the semantic similarity approach described in Section 3.3, the final JSON structure not only organizes these terms efficiently but also highlights hidden thematic relationships and interdisciplinary links inherent in the field.

4. Conclusions and further research

This study illustrates the implementation of advanced NLP techniques and the construction of a domain tree for wind energy, based on a specialized corpus created with Sketch Engine. Through the application of the LUMEN methodology, latent subdomains within the wind energy sector have been identified. The findings highlight the potential of NLP methodologies to navigate the complexities of multidisciplinary domains. By offering a tool accessible to researchers, practitioners, and non-specialists, this study also aims to provide a more dynamic understanding of the renewable energy sector through a different methodology.

Nevertheless, a new challenge emerged during the testing phase: a substantial portion of terms was not accurately recognized by the software and was subsequently categorized as “noise.” While some of these terms were indeed unrelated to the domain (e.g., typographical errors, corpus artifacts), others were relevant but failed to meet our initial similarity threshold. Future iterations will incorporate a more adaptive threshold, guided by expert feedback and additional contextual cues extracted from the corpus.

To address concerns about the validation of our classification accuracy, we plan to establish a formal evaluation pipeline, including precision and recall metrics. Through platforms such as Google Forms or specialized annotation tools, experts and stakeholders will assess classification outcomes on both large-scale samples and targeted subdomains. The resulting quantitative evaluations will not only provide transparency and reliability but also guide incremental refinements of our similarity metrics and the threshold-based classification step.

Finally, while this case study has primarily focused on English-language data, the methodology is designed for scalability to multiple languages. Initial multilingual experiments will involve parallel corpora in languages such as French, Spanish, or Italian, enabling semi-automated extraction of corresponding term hierarchies. This extension is of particular interest for different stakeholders, where multilingual resources are essential for policy-making, industry collaborations, and cross-border academic research. By enhancing our approach to classification accuracy and expanding our linguistic scope, LUMEN aims to serve as an adaptive platform for terminological analysis and ontology construction in rapidly evolving technical domains.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 for grammar and spelling checks. The authors have subsequently reviewed and edited the content and take full responsibility for the publication's final version.

References

[1] D. Yu, L. Li, H. Su, M. Fuoli, Using LLM-assisted Annotation for Corpus Linguistics: A Case Study of Local Grammar Analysis, CoRR abs/2305.08339 (2023). URL: https://doi.org/10.48550/arXiv.2305.08339. doi:10.48550/ARXIV.2305.08339. arXiv:2305.08339.

[2] G. Bordet, Brouillage des frontières, rencontres des domaines: quelles conséquences pour l'enseignement de la terminologie et de la traduction spécialisée, ASp. la revue du GERAS (2013) 95–115.

[3] B. De Bessé, Le domaine, Le sens en terminologie (2000) 182–197.

[4] N. Lechevrel, L'écolinguistique: une discipline émergente, RELQ/QSJL 3 (2008) 16–38.

[5] A. Rey, La terminologie: noms et notions, Que sais-je: Le point des connaissances actuelles, PUF, 1979. URL: https://books.google.it/books?id=7jJKAAAAYAAJ.

[6] V. Delavigne, La notion de domaine en question-à propos de l'environnement, Neologica 2022 (2022) 27–59.

[7] Y. Hamon, P. Paissa (Eds.), Discours environnementaux, volume 20 of *Lingue d'Europa e del Mediterraneo*, 2023. URL: https://www.aracneeditrice.eu/anteprime/9791221807769.pdf.

[8] C. Grimaldi, Le sfide linguistiche del cambiamento climatico, AIDAinformazioni (2021) 213–216.

[9] G. Myerson, Y. Rydin, The language of environment: A new rhetoric, Routledge, 2014.

[10] D. Pascaline, Étude en corpus de l'implantation de quelques emprunts à l'anglais et de leurs concurrents officiels, dans le domaine de l'environnement, Entre discours, langues et cultures: regards croisés sur le climat, l'environnement, l'énergie et l'écologie (2017) 61.

[11] J. Altmanova, E. Cartier, J. Luzzi, S. N. Pinto, S. Piscopo, et al., Innovations lexicales dans le domaine de l'environnement et de la biodiversité. le cas de bio en français et en italien, Neologica 16 (2022) 85–110.

[12] D. Candel, La présentation par domaines des emplois scientifiques et techniques dans quelques dictionnaires de langue, Langue française (1979) 100–115.

[13] V. Delavigne, Le domaine aujourd'hui. Une notion à repenser, 2002.

[14] J. C. Sager, A Practical Course in Terminology Processing, John Benjamins, Amsterdam/Philadelphia, 1990.

[15] C. Roche, Terminologie et ontologie, Langages (2005) 48–62.

[16] C. Roche, Ontological definition, in: Handbook of terminology, John Benjamins Publishing Company, 2015, pp. 128–152.

[17] N. Guarino, D. Oberle, S. Staab, What is an ontology?, in: S. Staab, R. Studer (Eds.), Handbook on Ontologies, International Handbooks on Information Systems, Springer, 2009, pp. 1–17. URL: https://doi.org/10.1007/978-3-540-92673-3_0. doi:10.1007/978-3-540-92673-3_0.

[18] I. Durán-Muñoz, M. R. Bautista-Zambrana, Applying ontologies to terminology: Advantages and disadvantages, Hermes-Journal of Language and Communication in Business (2013) 65–77.

[19] C. Roche, M. Calberg-Challot, L. Damas, P. Rouard, Ontoterminology - A new paradigm for terminology, in: J. L. G. Dietz (Ed.), KEOD 2009 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Funchal - Madeira, Portugal, October 6-8, 2009, INSTICC Press, 2009, pp. 321–326.

[20] R. Temmerman, K. Kerremans, Termontography: Ontology building and the sociocognitive approach to terminology description, Proceedings of CIL17 7 (2003).

[21] ISO 1087-1, Terminology work-Vocabulary-Part 1: Theory and application, Standard, International Organization for Standardization, Geneva, CH, 2000.

[22] ISO 704, Terminology work—Principles and methods, Standard, International Organization for Standardization, Geneva, CH, 2009.

[23] ISO 1087, Terminology work and terminology science–Vocabulary, Standard, International Organization for Standardization, Geneva, CH, 2019.

[24] A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, V. Suchomel, The Sketch Engine: ten years on, Lexicography 1 (2014) 7–36. URL: https://doi.org/10.1007/s40607-014-0009-9. doi:10.1007/s40607-014-0009-9.

[25] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, A. Chadha, A systematic survey of prompt engineering in large language models: Techniques and applications, CoRR abs/2402.07927 (2024). URL: https://doi.org/10.48550/arXiv.2402.07927. doi:10.48550/ARXIV.2402.07927. arXiv:2402.07927.

[26] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990. URL: https://doi.org/10.18653/v1/D19-1410. doi:10.18653/V1/D19-1410.

[27] A. Singhal, Modern information retrieval: A brief overview, IEEE Data Eng. Bull. 24 (2001) 35–43. URL: http://sites.computer.org/debull/A01DEC-CD.pdf.

A. Online Resources

The sources of LUMEN are available via https://anonymous.4open.science/r/LUMEN,