Combining Knowledge graph and LLM to extract thesaural relationship and concepts on Cybersecurity

Elena Cardillo[1 [0000−0001−5003−205X]], Alessio Portaro[1 [0009−0003−2911−7472]], and Maria Taverniti[1 [0000−0001−7000−5817]]

Institute of Informatics and Telematics, National Research Council, Rende, Italy
elena.cardillo@iit.cnr.it, maria.taverniti@cnr.it
http://www.iit.cnr.it

Abstract. Interest in the use of knowledge graphs in the cybersecurity domain has been growing rapidly in recent years. However, due to the high specificity of the domain, some issues related to the robustness of these graphs are still open. This work analyses the results of a case study aimed at exploiting a BERT-based model - in particular word embeddings - combined to a Knowledge graph approach to enhance the population and enrichment of domain-oriented controlled vocabularies (i.e., Thesauri). Resources controlled and validated by domain experts, such as thesauri, are essential in high-risk domains like cybersecurity, where robustness and reliability are key factors. A Natural Language Processing inspired pipeline is presented, including knowledge graph extraction and inference to identify thesaural concepts and relationships. Although early findings suggest the model's potential to enhance controlled vocabularies with novel insights, the accuracy and quality of the extracted entities still underscore the need for an in-depth validation by a domain expert to select the candidates concepts and relationships.

Keywords: cybersecurity · knowledge graph · thesauri · BERT · relation extraction

1 Introduction

Since the advent of Tranformers-based Neural language models (NLMs) [45], Natural Language Processing (NLP) techniques have experienced unprecedented growth, ranging from "classic" linguistic tasks [7, 14, 16, 18, 35, 44] to more specific ones verticalized on restricted domains [15]. These models have recognized weakness in the enormous computational cost and the lack of control over the data needed for training. However, performances achieved are outstanding, especially in general-purpose approaches and wide-known tasks [47]. The situation is very different when moving to niche sub-domains and languages other than English [3]. In particular, in high-risk domains, such as cybersecurity, the presence of validated resources still remains a key element [4, 51].

Starting from these premises, this work has the aim to analyze the results of a case study developed in the domain of cybersecurity, where a hybrid approach combining elements from NLP / Large Language Models (LLMs) and knowledge graphs has been applied to enrich and update a domain-oriented Thesaurus [9]. A knowledge graph (KG) is typically defined as a semantic network comprising entities (i.e., concepts, objects, events, and so on), factual attributes of entities, and relationships between them, that are displayed as a graph and stored in a graph database. Unlike old semantic nets, these have an entity-centric view, focusing not only on the structured representation of the semantic knowledge of a specific domain, but also on how the entities are connected, and how they are interpreted and disambiguated. This helps in the verification of information consistency and correctness [1], even though they are less rigorous than ontologies. Alongside the advancements in NLP, KG-based approaches have been adapted to various sub-domains [52] yielding encouraging outcomes. Although, in recent years the few existing approaches adapted to cybersecurity, have shown mixed results [1,6,25,43]. Combining the linguistic abilities of NLMs with knowledge graphs can significantly improve tasks such as early warning predictions, decision-making, and health emergencies. In particular, the case study described in this paper aims to enhance the population and updating of a previously developed bilingual thesaurus on cybersecurity[1], by identifying semantic relationships and indexing terms relevant to the domain. This is driven by the necessity to adjust the Thesaurus' level of semantic representativeness with respect to the reference domain in order to guarantee appropriate terminology coverage, as the introduction of new guidelines, rules, and technical documents may also result in revision of the domain's terminology.
Although the thesaurus and the related corpus is available also in Italian, the described approach is focused on English for two reasons: i) because resources and models are much more narrowly available and tested in English [12]; ii) because the approach used to extract knowledge graphs from the legal corpus has been already tested for English [40]. Furthermore, while many studies have explored in recent years cross-lingual methods to compensate for the lack of resources [21] (including Italian [17, 19]), for specific domains we are still far from achieving outcomes equivalent to those of English in certain fields. [15, 19, 20, 31]. This is evident also in the pre-processing stages for languages other than English [34].
The paper is structured as follows: Section 2 provides an overview of recent related works. Next, Section 3 details the materials and research methodology, including information on the structural probe, the chosen LLM, and the dataset used. Section 4 presents the experimental evaluation along with a discussion of the preliminary results. Finally, Section 5 offers conclusions and suggests potential future developments.

[1] The Thesaurus ha been published in 2019 on the Italian Cybersecurity Observatory's website (OCS)

2 Related Work

The acquisition of knowledge graph has recently become a focal point, with advancements primarily driven by LLMs. Initially, knowledge graph extraction relied heavily on rule-based systems and information retrieval methods. However, these approaches often struggled with handling unstructured or ambiguous text data and necessitated significant manual intervention from experts. The introduction of word embeddings and subsequently large neural language models, such as BERT, revolutionized this field by enabling unsupervised NLP techniques for knowledge graph extraction. See [39] for a structured overview of the research landscape on KGs in NLP, including a useful taxonomy of the numerous NLP tasks used for knowledge acquisition and application. Several studies have investigated the effectiveness of fine-tuning BERT models for entity recognition and relationship extraction tasks. Notably, research by [50] demonstrated substantial improvements in extraction accuracy through this approach. Furthermore, BERT's contextual understanding and ability to discern relationships have been instrumental in various classification tasks. Techniques like attention mechanisms and multi-instance learning, as highlighted by [37], have demonstrated enhancements in identifying and categorizing intricate relationships within textual data. Concurrently, Natural Language Inference (NLI) has emerged as a complementary method for constructing knowledge graphs. By harnessing BERT's contextual embeddings, NLI models assist in deducing entailment and contradiction, thereby facilitating the extraction of implicit relationships embedded in the text. Recent studies have explored joint learning approaches that simultaneously tackle entity recognition, like in [49] and relation extraction tasks. According to [9, 38], BERT-based models exhibit the capability to enhance both aspects concurrently, resulting in more cohesive and precise knowledge graph construction. In recent research studies, hybrid NLP/NLM and knowledge graph approaches have been applied to the cybersecurity domain to extract entities and relations from textual data with the aim of contributing to the field of cyber threat intelligence (CTI). One example can be found in the study outlined by [2], in which they introduced theRoBERTa-BiGRU-CRF model for performing entity-relation extraction from a CTI corpus. This model was evaluated across various contexts pertaining to Advanced Persistent Threats (APT) and ransomware incidents, demonstrating encouraging outcomes.

3 Materials and Methods

This section describes in detail the resources and the NLM used in the study. Specifically, the resources employed include: *i)* The bilingual Thesaurus on cybersecurity (Italian-English) published on the aforementioned Italian OCS website[2], which offers a structured structured representation of the domain knowledge by establishing semantic relationships between terms extracted from technical documents and legislative acts related to cybersecurity, enabling users' comprehension

[2] `https://www.cybersecurityosservatorio.it/it/Services/thesaurus.jsp`

of specialized terminology in this field; *ii)* A sample corpus of cybersecurity regulations in English gathered from the OCS website utilizing automated tools. Regarding the NLM approach, the widely-acknowledged BERT model was chosen. This selection was influenced by the model's extensive popularity and flexibility, demonstrated through its application across a broad spectrum of tasks [11].

3.1 BERT Model

Currently, BERT [11] is a leading Transformer-based NLM in NLP, known for its efficiency and high performance. The foundational *BERT-base* version consists of 12 layers of decoder-only Transformers, each with 768 hidden dimensions and 12 attention heads, amounting to 110 million parameters, and supports input sequences up to 512 words. Based on the Transformer encoder architecture [46], BERT features a multi-layer bidirectional design, pre-trained on extensive unlabeled text using two main objectives: masked language modeling and next sentence prediction.

BERT's strength lies in generating robust context-dependent sentence representations, which can be adapted to various NLP tasks through fine-tuning. This process involves adjusting several hyperparameters, which significantly influence the outcomes. BERT's pre-training method focuses on masked language modeling, where random words in the training corpus are masked, enabling the model to learn from both directions of a sentence while predicting the masked words. BERT offers two pre-trained models based on either *cased* or *uncased* input vocabularies. Its bidirectional analysis provides substantial generative capacity in deep network layers, with outer layers tailored for task-specific fine-tuning, establishing BERT as a benchmark model in recent literature. In BERT, each input sequence begins with a special *[CLS]* token, which produces a vector of size H (hidden layer size) representing the entire input sequence. Each sentence within the input sequence must also end with a unique *[SEP]* token.

For an input sequence of words $t = (t_1, t_2, \ldots, t_m)$, BERT's output is $h = (h_0, h_1, h_2, \ldots, h_m)$, where $h_0 \in \mathbb{R}^H$ is the final hidden state of the *[CLS]* token, providing a pooled representation of the entire input sequence. The final hidden states of the remaining input tokens are denoted as $h_1, h_2, \ldots, h_m$.

When fine-tuning BERT for classifying input sequences into K distinct categories, the final hidden state h_0 is used to feed a classification layer, followed by a softmax operation to convert category scores into probabilities, as described by Sun et al. [42]:

$$P = softmax(CW^T) \tag{1}$$

where $W \in \mathbb{R}^{K \times H}$ is the parameter matrix of the classification layer.

3.2 The Thesaurus

As mentioned above, this research initially focused on an existing resource, developed in a previous project [29], i.e., the OCS bilingual Thesaurus, designed

to organize and formalize the knowledge domain of cybersecurity, characterized by highly technical terminology [27].

According to ISO 25964:2013 (2013, p. 12) [24], a thesaurus is a "controlled and structured vocabulary in which concepts are represented by terms organized so that relationships between concepts are made explicit and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms." The primary purpose of such a semantic resource is to organize the terminology of a specific domain to support indexing operations, targeted knowledge discovery [8], and terminological control in information retrieval tasks [13].

As Lykke (2001, 778) [33] stated, "the thesaurus is a tool that helps individual users understand the collective knowledge domain." The terms within the knowledge domain are semantically connected according to ISO standards 25964-1:2011 [23] and 25964-2:2013 [24], which define three main types of semantic relationships, summarized in table 1.

Table 1: Types of Semantic Relationships in Thesauri

Relationship Type	Description
Hierarchical	Marked with Broader Term (BT) and Narrower Term (NT) tags, these denote the specificity connection between terms, including "generic" (class-member "IS-A"), "instance", and "partitive" relationships. E.g.: Cybersecurity *NT* Cyber risk management
Equivalence	Marked with Used (USE) and Used For (UF) tags, these manage links between terms representing the same concept: in the first case the link is from a synonym or unauthorized / non-preferred term to a Preferred term (i.e., the term that has been selected to be included in the controlled vocabulary), and in the second case from a Preferred heading to a synonym / non-preferred term. E.g.: Cybersecurity *UF* Information Security
Associative	Marked with the Related Term (RT) tag, these indicate coordination of terms within the same category or across different categories. E.g.: Cybersecurity *RT* Privacy

These relationships play a crucial role in structuring the conceptual framework of the specialized domain. By explicitly defining the connections between terms (synonymy, hypernymy/hyponymy, meronymy/holonymy, etc.), in fact, the thesaurus acts as a semantic map, enabling the tool to:

- *Normalize information* - By identifying synonyms and preferred terms, the vocabulary ensures consistency in information representation, reducing redundancy and improving search accuracy.
- *Disambiguate sector-specific information* - By capturing hierarchical relationships (hypernymy/hyponymy) and part-whole relationships (meronymy/holonymy), the vocabulary clarifies the meaning of terms within the specific domain, preventing misinterpretations. [8].

After this short description of theoretical concepts related to Thesauri, it's important to give some details about the aforementioned OCS Thesaurus, selected for the presented use case. Extensive technical documentation was necessary to identify a comprehensive set of sector-specific terms to be included in the thesaurus. In particular, a specialized corpus of 57 heterogeneous documents

was used to populate it and shape its structure. More specifically, this corpus includes:

- Laws and regulations: especially national and European regulations proving the main references in the field of knowledge.
- Best practices: guidelines and recommendations for implementing effective cybersecurity measures.
- Technical Reports and papers: In-depth analyses and studies on specific aspects of cybersecurity.
- Standards and glossaries: unique definitions and taxonomies of technical terms related to cybersecurity, such as the NIST *Glossary of Key Information Security Terms* [41] and ISO 27000:2016 Information Technology — Security Techniques — Information Security Management Systems — Overview and Vocabulary [22, 28].

The OCS Thesaurus includes a total of 238 index terms, arranged through a semantic relationship structure [5], [36], selected after the application of Term Frequency/Inverse Document Frequency (TF/IDF) statistical measure [48] and a domain expert validation, which followed predefined guidelines set by international standards [26]. Figure 1 shows a branch of the Thesaurus which can be navigated on the OCS website, representing the semantic structure for the term VAPT, acronym of *Vulnerability Assessment and Penetration Testing* related to indexing term *Vulnerability*.

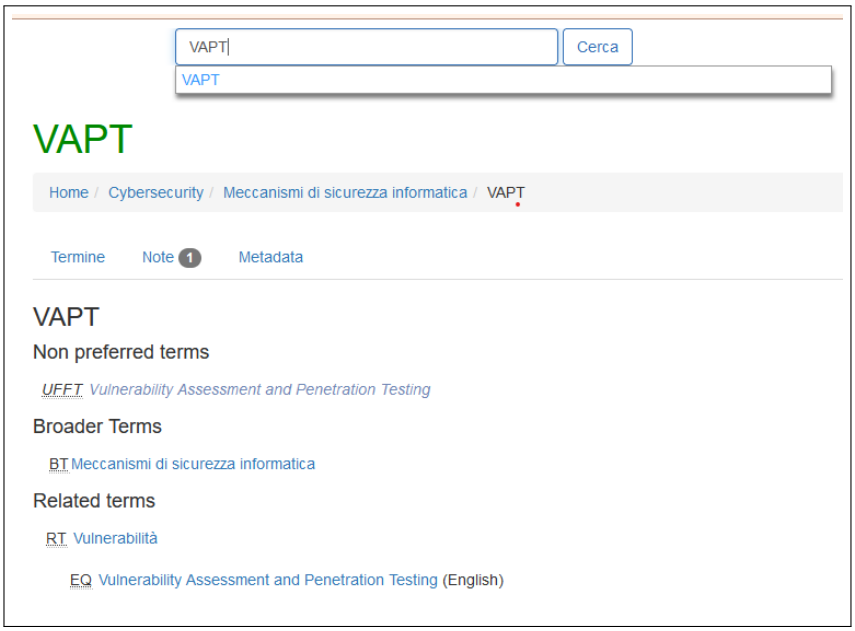

Fig. 1: OCS Thesaurus branch example

What can be observed here is that the acronym is used as the preferred term in the thesaurus and the long name as a synonym (equivalent relation UF), and that this concept is an hypernym (thus a more specific concept) of Information Security Monitoring Policy (in Italian "meccanismi di sicurezza informatica").

Moreover, as can be observed in the figure above, all the preferred terms in the thesaurus have the corresponding English term linked as an equivalence term (preceded by the notation "EQ") and extracted by the English sub-corpus. For the use case analyzed in the present work, a sample containing five documents in English has been extracted from The OCS Thesaurus corpus to evaluate the proposed methodology. Moreover, a specific set of relationships characterizing the Top Term (TT) "Cybersecurity", including 153 connecting terms has been used for testing.

4 Experiments

Starting from the methodology introduced in [10], experimental assessment is distributed in two phases, corresponding mainly to the two layers illustrated in Figure 2.

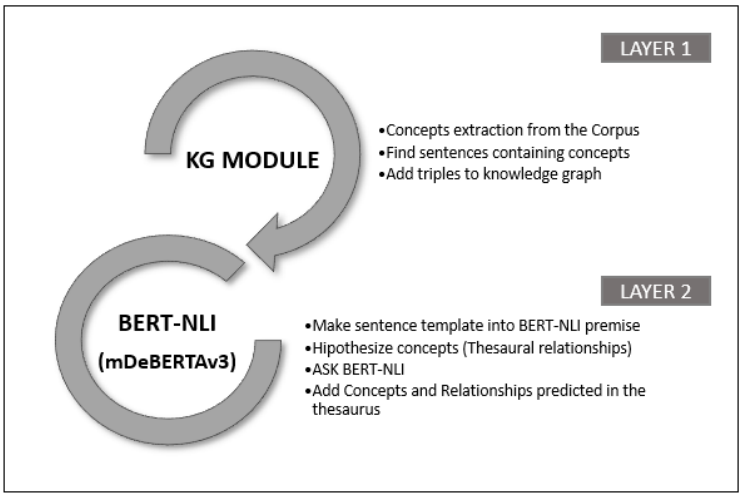

Fig. 2: Combined KG - BERT-NLI approach

4.1 Knowledge Graph Extraction

The initial phase of the experiment involves creating a Knowledge Graph (KG) by extracting concepts and their relationships from the sample of European legislative texts on cybersecurity mentioned in Section 3. To ensure semantically relevant KGs, the subset of documents derived from the OCS Corpus was selected by a team of legal and terminology experts. The KG extraction utilized

software from the EU Interlex project[3] [40], [30]. This tool extracs concepts from PDF texts using the **SpaCy**[4] library, which infers a PoS-tagged dependency tree (DPT). The Interlex module navigates the DPT to extract noun phrases as concepts. Once a set of concepts is established, the module identifies all those sentences connecting every possible pair of concepts. The sentences in their natural language form (from which we removed the particular instances of nouns playing the role of subject and object) are considered as the relations connecting the nouns of the KG. The retention of these relationships in their natural language form makes them suitable for use with LLMs. Considering the objective of our use case, this particular feature is the main reason why we have chosen to reuse the Interlex Portal library. The Interlex module, thus generates a set of triples (subject, predicate, object) that form the KG.

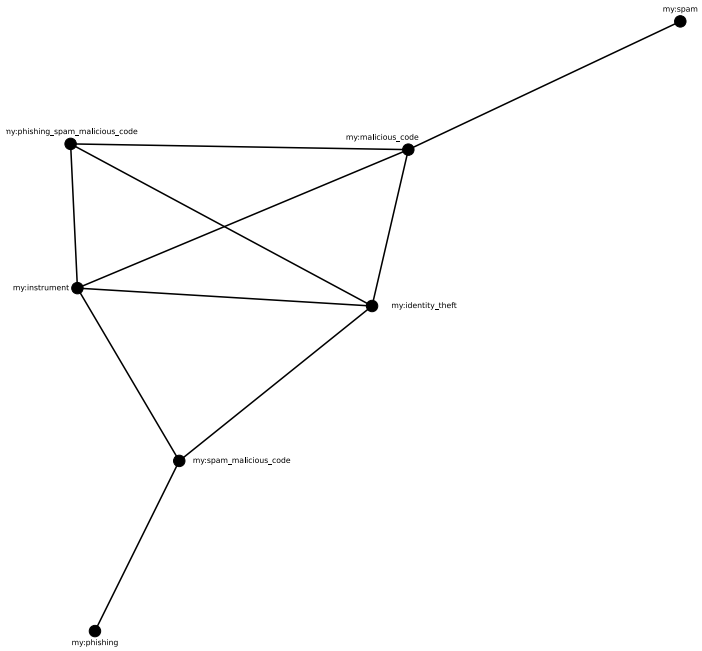

Fig. 3: KG nodes example

[3] InterLex is an EU funded project that aims at developing an online platform to provide information, decision support and training on private international law. It addresses the identification of the legal system having jurisdiction and of the national law to be applied to a specific case as well as the retrieval of relevant legal materials. Details are available at link:`https://interlex-portal.eu/en/index.html`

[4] `https://spacy.io`

Figure 3 above shows some example of triples represented in the extracted KG and gives an idea about the redundancy of semantic information, because many instances of the same relation can appear in the same graph.

4.2 NLI task

The second phase involves populating the OCS Thesaurus mentioned in Subsection 3.2, using LLMs fed with the extracted KG. As observed above, the KG model often contains redundant semantic information, with multiple instances of the same relation. Moreover, KG relations are typically subject-object relations, while thesaurus relations are more linguistic (i.e., hierarchical, equivalence, associative). To address redundancy, and to better generalize the information contained in the input text, linguistic relations need to be identified. Normally the identification of candidate concepts and thesaural relationships is a time consuming task, carried out by a team of linguistics and domain experts. Here we aimed at automating this task using modern LLMs like BERT. For this reason, a fine-tuned version of BERT for NLI tasks, **HuggingFace**'s mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 [30], was preferred.

NLI fine-tuned LLMs are trained to predict the probability of a sentence called *thesis*(conclusion) to be a logical entailment of another text called *hypothesis* (context/premises). A hypothesis, as shown in table 2 is based on a premise and it is classified by a NLI model into three categories: entailment, contradiction, or neutral. For this task, the hypothesis consisted of questions about relations between concept pairs, and the premise included context from the KG. Relations were extracted as templates with placeholders for subjects and objects, replaced by corresponding concepts from the KG. The Interlex library ensured that resulting sentences were coherent in natural language, fitting the expected input for a transformer-based model like BERT. When the NLI model predicted "entailment", a relation in the domain thesaurus was established between the two concepts, corresponding to the hypothesized relation type.

Table 2: NLI Model Classifications

Class	**Description**
Entailment	The hypothesis can be inferred from the premises.
Contradiction	The hypothesis contradicts the premises.
Neutral	The hypothesis neither follows nor contradicts the premises.

For example, if a KG triple had "security" as the subject and "privacy" as the object, the type of thesaural relationship (e.g., RT for related term) was identified and verified for inclusion in the Thesaurus.

Some statistics regarding the main outputs are shown below. In particular in Table3 we can see the high number of domain related concepts extracted as subject in the KG (i.e., 530 entities).

Table 3: KG results

Type	Data
Text Sources	5
Sentences	1951
Triples	1123
Entities	530

On the other hand, Table4 highlights the high number of Hypotheses related to thesaural relationships verified (a total of 841,110 among six premises).

Table 4: BERT-NLI results

Type of Thesaural relation	Hypotheses verified
Synonyms	21368
Hierarchical relationships BT	31400
Hierarchical relationships NT	17016
Related Terms RT	38558

Actually, a more accurate evaluation is being performed by a domain expert to compare extracted concepts and relationships with the OCS Thesaurus and to select candidate new terms. This process is important to distinguish, for example in the case of equivalence terms, the preferred term and synonyms. It is a crucial task to create other connection branches for the thesaurus, as we know that only the preferred terms are used in the indexing process. Some preliminary results of this process led us to provide candidate terms / concepts to be integrated in the cybersecurity Thesaurus, as in the example showed below:

Table 5: Example of Thesaurus enrichment

Concept 1	Relationship type	Concept 2
Crime	BT	Fraud
Crime	BT	Forgery
Crime	BT	Identity theft
Fraud	RT	Forgery
Identity theft	UF	Identity fraud

These examples show how positive can be the outcomes at a conceptual level, facilitating the identification of numerous relevant concepts (both already existing entries of the Thesaurus and new candidates terms for its enrichment) related to the cybersecurity domain. For instance, the hierarchical relationship BT (Broader Term) between the concepts “Crime” and “Fraud” and between “Forgery” and “Identity theft” clarifies the hierarchical structure of concepts, essential for organizing information logically and accessibly. However, the analysis has highlighted also some limitations, particularly concerning the precision of the thesaural relationships extraction. In fact, relationships like RT (Related Term) and equivalence (USE/UF), as observed respectively between “Fraud” and “Forgery”, and between “Identity theft” and “Identity fraud”, have shown a de-

crease in precision. This suggests the need for more nuanced input from domain experts and terminologists to refine and clarify these links. Additionally, automatic extraction encountered challenges with compound nouns containing more than two terms, underscoring the need for further advancements in extraction technology to manage this complexity. These points reflect the ongoing challenge of balancing automation with accuracy in OCS Thesaurus results, emphasizing the importance of continuous monitoring and improvement in extraction and enrichment processes. In order to overcome some of the aforementioned limitations, a further experiment is underway aimed at testing the application of the Interlex semantic layer (i.e., Taxonomy construction), which allows structuring the extracted KG as a light ontology, giving it the basic form of a taxonomy by means of Formal Concept Analysis (FCA) [40]. This brings to the identification of the types/classes of a concept thus exploiting its hypernyms (NT relationships in the thesaurus) in the KG.

5 Conclusion and Future Works

This paper showed a case study firstly introduced in [9] for exploiting the association of KG and BERT models to enrich a thesaurus in the cybersecurity domain. In particular, the LLM has been used to extract thesaural concepts and relationships by asking to NLI fine tuned version of the model if the assumed relational hypothesis actually entails from the premises represented by the sentences containing the target concepts.

By leveraging BERT embeddings, our approach effectively extracts information from a corpus of cybersecurity-related documents. The proposed NLP-inspired pipeline seamlessly integrates NLMs, knowledge graph extraction, and NLI to identify domain concepts and implicit relationships among them, thereby allowing to integrate with new information the domain-specific thesaurus.

On one hand, preliminary results demonstrate the robustness of this methodology, highlighting the applicability of state-of-the-art LLMs in augmenting specialized controlled vocabularies with new knowledge. These findings underscore the potential of integrating BERT-based techniques to enhance the semantic richness and utility of domain-oriented thesauri without relying on outdated lexicons, and without loosing the peculiarity of their relational structure. This research can provide a contemporary framework for knowledge extraction and relationship identification in specialized contexts, such as cybersecurity which is recognized to be highly technical, actually enclosing different sub-domains (e.g., security management, identity and access management, compliance, cryptography, software development security, etc.) so requiring a wide range of different technical document sources to be analysed and high-performance techniques to identify domain-related semantic entities. Even in this preliminary stage, the promising results pave the way for further exploration and application in the rapidly growing landscape of knowledge management.

On the other hand, the analysis of the results showed some limitations of the model. In fact, even if good candidate cybersecurity-related concepts have been

identified in the KG for enriching the thesaurus, the precision decreased when entailing the pre-defined thesaural relationships between the extracted concepts, above all regarding the identification of hierarchical (type NT) and the equivalence relationships. One reason for this outcome can be due to the specificity of the sample corpus used for testing, which includes only legal documents (i.e., European regulations). Indeed, these types of documents present a fixed textual structure (e.g., title, preamble, general provisions, transitional provisions, etc.) which has proven to be unsuitable for such linguistic models or at least resulted to be less feasible when extracting predicates for the KG. One way to address this issue could be a more accurate customization of the Interlex's KG module parameters to improve the triples generation. Alternatively, considering the results of other studies found in the literature, the use of other KG modules and pre-trained BERT models for the domain knowledge could be tested (e.g., the Knowledge graph on Legal Data (SEC) proof of concepts[5]; the SecBert[6], the SecRoBERTa[7] or the CyBERT modules, trained on cybersecurity texts). Moreover, even reducing the human task in the preliminary stages of the approach, with respect to classical methods used for Thesaurus building, a final validation from domain experts is required. As mentioned in Section 4, we are trying to improve the results, and in particular this aspect, by testing the integration of more accurate semantic layers in the first phase of the approach, so to create the KG. This may allow us to understand whether by having a more expressive output in terms of semantics, the NLI model could be more performant in the verification of correct hypotheses, and thus in the identification of thesaural relationships useful to enhance the Thesaurus.

A planned future work will be the testing of the presented approach using Italian terms in the cybersecurity thesaurus, taking advantages of BERT's multilingual capabilities. This could open numerous research possibilities for parallel analyses on enriching translations within the resource.

Additionally, new methods that promise to surpass the current performance of NLMs will be explored, such as identifying and extracting specific types of hierarchical relationships (e.g., partitive relationship and the instance relationship, this one used to name particular instances of a class of things) to correctly connect concepts and properly update the thesaurus structure. Finally, the emergence of approaches beyond NLMs, such as Quantum NLP [17, 32], a sub-branch of NLP using quantum theory methods to boost performance, will be investigated to improve the accuracy and volume of new knowledge identified for populating the thesaurus.

Acknowledgments. This study was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

[5] https://github.com/AnjaneyaTripathi/knowledge_graph?tab=readme-ov-file
[6] https://huggingface.co/jackaduma/SecBERT
[7] https://huggingface.co/jackaduma/SecRoBERTa

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agrawal, G., Deng, Y., Park, J., Liu, H., Chen, Y.C.: Building knowledge graphs from unstructured texts: Applications and impact analyses in cybersecurity education. Information **13**(526) (2022). https://doi.org/https://doi.org/10.3390/info13110526
2. Ahmed, K., Khurshid, S.K., Hina, S.: Cyberentrel: Joint extraction of cyber entities and relations using deep learning. Computers Security **136**, 103579 (2024). https://doi.org/https://doi.org/10.1016/j.cose.2023.103579, https://www.sciencedirect.com/science/article/pii/S0167404823004893
3. Alam, F., Chowdhury, S.A., Boughorbel, S., Hasanain, M.: Llms for low resource languages in multilingual, multimodal and dialectal settings. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts. pp. 27–33 (2024)
4. Alawida, M., Mejri, S., Mehmood, A., Chikhaoui, B., Isaac Abiodun, O.: A comprehensive study of chatgpt: advancements, limitations, and ethical considerations in natural language processing and cybersecurity. Information **14**(8), 462 (2023)
5. Behrang, Q.Z., Handschuh, S.: The ACL RD-TEC: A dataset for benchmarking terminology extraction and classification in computational linguistics". In: Proceedings of the 4th International Workshop on Computational Terminology (Computerm). pp. 52–63. Association for Computational Linguistics and Dublin City University, Dublin, Ireland (aug 2014). https://doi.org/10.3115/v1/W14-4807, https://www.aclweb.org/anthology/W14-4807
6. Bolton, J., Elluri, L., Joshi, K.P.: An overview of cybersecurity knowledge graphs mapped to the mitre attck framework domains. In: 2023 IEEE International Conference on Intelligence and Security Informatics (ISI). pp. 01–06. IEEE (October 2023). https://doi.org/10.1109/ISI58743.2023.10297134
7. Bonetti, F., Leonardelli, E., Trotta, D., Guarasci, R., Tonelli, S.: Work hard, play hard: Collecting acceptability annotations through a 3d game. p. 1740 – 1750 (2022)
8. Broughton, V.: Essential Thesaurus Construction. Facet (2006). https://doi.org/10.29085/9781856049849
9. Cardillo, E., Portaro, A., Taverniti, M., Lanza, C., Guarasci, R.: Towards the automated population of thesauri using BERT: A use case on the cybersecurity domain. In: Barolli, L. (ed.) Advances in Internet, Data & Web Technologies - The 12th International Conference on Emerging Internet, Data & Web Technologies, EIDWT 2024, Naples, Italy, 21-23 February 2024. Lecture Notes on Data Engineering and Communications Technologies, vol. 193, pp. 100–109. Springer (2024). https://doi.org/10.1007/978-3-031-53555-0_10, https://doi.org/10.1007/978-3-031-53555-0_10
10. Chen, W., Ji, H.: Infer: Capturing implicit entity relations for knowledge graph completion using contextualized language models. arXiv preprint arXiv:2006.05295 (2020)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Pa-

pers). pp. 4171–4186. ACL, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
12. Diandaru, R., Susanto, L., Tang, Z., Purwarianti, A., Wijaya, D.T.: Could we have had better multilingual llms if english was not the central language? In: Proceedings of the Second International Workshop Towards Digital Language Equality (TDLE): Focusing on Sustainability@ LREC-COLING 2024. pp. 43–52 (2024)
13. Gabler, S.: Thesauri – a toolbox for information retrieval. Bibliothek Forschung und Praxis **47**(2), 189–199 (2023). https://doi.org/doi:10.1515/bfp-2023-0003, https://doi.org/10.1515/bfp-2023-0003
14. Gargiulo, F., Minutolo, A., Guarasci, R., Damiano, E., De Pietro, G., Fujita, H., Esposito, M.: An electra-based model for neural coreference resolution. IEEE Access **10**, 75144–75157 (2022). https://doi.org/10.1109/ACCESS.2022.3189956
15. Guarasci, R., Catelli, R., Esposito, M.: Classifying deceptive reviews for the cultural heritage domain: A lexicon-based approach for the italian language. Expert Systems with Applications **252** (2024). https://doi.org/10.1016/j.eswa.2024.124131, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85192240278&doi=10.1016%2fj.eswa.2024.124131&partnerID=40&md5=12d7d5089bd1ce00331a65fc7ac332bd
16. Guarasci, R., Damiano, E., Minutolo, A., Esposito, M., De Pietro, G.: Lexicon-grammar based open information extraction from natural language sentences in italian. Expert Systems with Applications **143**, 112954 (2020). https://doi.org/https://doi.org/10.1016/j.eswa.2019.112954, https://www.sciencedirect.com/science/article/pii/S0957417419306724
17. Guarasci, R., De Pietro, G., Esposito, M.: Quantum natural language processing: Challenges and opportunities. Applied Sciences (Switzerland) **12**(11) (2022). https://doi.org/10.3390/app12115651
18. Guarasci, R., Minutolo, A., Damiano, E., De Pietro, G., Fujita, H., Esposito, M.: ELECTRA for neural coreference resolution in italian. IEEE Access **9**, 115643–115654 (2021). https://doi.org/10.1109/ACCESS.2021.3105278
19. Guarasci, R., Silvestri, S., De Pietro, G., Fujita, H., Esposito, M.: Bert syntactic transfer: A computational experiment on italian, french and english languages. Computer Speech & Language **71**, 101261 (2022)
20. Guarasci, R., Silvestri, S., De Pietro, G., Fujita, H., Esposito, M.: Assessing bert's ability to learn italian syntax: A study on null-subject and agreement phenomena. Journal of Ambient Intelligence and Humanized Computing **14**(1), 289–303 (2023)
21. Guarasci, R., Silvestri, S., Esposito, M.: Probing cross-lingual transfer of xlm multi-language model. Lecture Notes on Data Engineering and Communications Technologies **193**, 219 – 228 (2024). https://doi.org/10.1007/978-3-031-53555-0_21, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85186474840&doi=10.1007%2f978-3-031-53555-0_21&partnerID=40&md5=2dd368fbe6cc5fc20da4886433d53fc6
22. Hazem, A., Daille, B., Claudia, L.: Towards automatic thesaurus construction and enrichment. In: Daille, B., Kageura, K., Terryn, A.R. (eds.) Proceedings of the 6th International Workshop on Computational Terminology. pp. 62–71. European Language Resources Association, Marseille, France (may 2020), https://aclanthology.org/2020.computerm-1.9
23. ISO - International Organization for Standardization: ISO 25964-1:2011 Information and documentation — Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval (August 2011), ISO 25964-1:2011(en)

24. ISO - International Organization for Standardization: ISO 25964-2:2013 Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies (2013), Iso25964-2:2013
25. Jia, Y., Qi, Y., Shang, H., Jiang, R., Li, A.: A practical approach to constructing a knowledge graph for cybersecurity. Engineering **4**(1), 53–60 (2018). https://doi.org/https://doi.org/10.1016/j.eng.2018.01.004, https://www.sciencedirect.com/science/article/pii/S2095809918301097, cybersecurity
26. de Keyser, P.: Indexing from Thesauri to the Semantic Web. Chandos Publishing (2012). https://doi.org/10.1080/01639374.2013.841786
27. Lanza, C., Cardillo, E., Taverniti, M., Guarasci, R.: Terminology management in cybersecurity thought knowledge organization systems: an italian use case. International Journal on Advances in Security (1-2), 17–27 (2020)
28. Lanza, C.: Semantic Control for the Cybersecurity Domain: Investigation on the Representativeness of a Domain-Specific Terminology Referring to Lexical Variation. CRC Press (2022). https://doi.org/https://doi.org/10.1201/9781003281450citation-key
29. Lanza, C., Cardillo, E., Taverniti, M., Guarasci, R.: Knowledge representation frameworks for terminology management in cybersecurity: The ocs project use case (2019), https://api.semanticscholar.org/CorpusID:249558982
30. Laurer, M., Atteveldt, W.v., Casas, A.S., Welbers, K.: Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. Preprint (Jun 2022), https://osf.io/74b8k, publisher: Open Science Framework
31. Licari, D., Comandè, G.: Italian-legal-bert: A pre-trained transformer language model for italian law. EKAW (Companion) **3256** (2022)
32. Lorenz, R., Pearson, A., Meichanetzidis, K., Kartsaklis, D., Coecke, B.: Qnlp in practice: Running compositional models of meaning on a quantum computer. Journal of Artificial Intelligence Research **76**, 1305–1342 (2023)
33. Lykke, M.: A framework for work task based thesaurus design. Journal of Documentation **57**, 774–797 (12 2001). https://doi.org/10.1108/EUM0000000007100
34. Marulli, F., Pota, M., Esposito, M., Maisto, A., Guarasci, R.: Tuning syntaxnet for pos tagging italian sentences. Lecture Notes on Data Engineering and Communications Technologies **13**, 314 – 324 (2018). https://doi.org/10.1007/978-3-319-69835-9_30
35. Minutolo, A., Guarasci, R., Damiano, E., De Pietro, G., Fujita, H., Esposito, M.: A multi-level methodology for the automated translation of a coreference resolution dataset: an application to the italian language. Neural Computing and Applications **34**(24), 22493 – 22518 (2022). https://doi.org/10.1007/s00521-022-07641-3
36. Nielsen, M.L.: Thesaurus construction: Key issues and selected readings. Cataloging & Classification Quarterly **37**(3-4), 57–74 (2004). https://doi.org/10.1300/J104v37n03_05, https://doi.org/10.1300/J104v37n03_05
37. Qi, K., Du, J., Wan, H.: Learning from both structural and textual knowledge for inductive knowledge graph completion. Advances in Neural Information Processing Systems **36** (2024)
38. Ranade, P., Piplai, A., Joshi, A., Finin, T.: Cybert: Contextualized embeddings for the cybersecurity domain. In: 2021 IEEE International Conference on Big Data (Big Data). pp. 3334–3342. IEEE (2021)
39. Schneider, P., Schopf, T., Vladika, J., Galkin, M., Simperl, E., Matthes, F.: A decade of knowledge graphs in natural language processing: A survey (2022)

40. Sovrano, F., Palmirani, M., Vitali, F.: Legal knowledge extraction for knowledge graph based question-answering. In: Legal Knowledge and Information Systems, pp. 143–153. IOS Press (2020)
41. of Standards, N.I., Technology: Glossary of key information security terms. Tech. rep., NIST Interagency or Internal Report (NISTIR) 7298 Rev. 2 (May 2013)
42. Sun, C., Qiu, X., Xu, Y., Huang, Xuanjing", e.M., Huang, X., Ji, H., Liu, Z., Liu, Y.: How to fine-tune bert for text classification? In: China National Conference on Chinese Computational Linguistics. p. 194–206. Springer (2019)
43. Sun, N., Ding, M., Jiang, J., Xu, W., Mo, X., Tai, Y., Zhang, J.: Cyber threat intelligence mining for proactive cybersecurity defense: a survey and new perspectives. IEEE Communications Surveys & Tutorials (2023)
44. Trotta, D., Guarasci, R., Leonardelli, E., Tonelli, S.: Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 2929–2940. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.findings-emnlp.250, https://aclanthology.org/2021.findings-emnlp.250
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. pp. 5998–6008. Long Beach, CA, USA (2017), http://papers.nips.cc/paper/7181-attention-is-all-you-need
46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
47. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: Superglue: A stickier benchmark for general-purpose language understanding systems. Advances in neural information processing systems **32** (2019)
48. Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L.: Interpreting tf-idf term weights as making relevance decisions. ACM Trans. Inf. Syst. **26**, 13:1–13:37 (2008), https://api.semanticscholar.org/CorpusID:18303048
49. Würsch, M., Kucharavy, A., David, D.P., Mermoud, A.: Llms perform poorly at concept extraction in cyber-security research literature (2023)
50. Yang, S., Yoo, S., Jeong, O.: Denert-kg: Named entity and relation extraction model using dqn, knowledge graph, and bert. Applied Sciences **10**(18), 6429 (2020)
51. Yigit, Y., Buchanan, W.J., Tehrani, M.G., Maglaras, L.: Review of generative ai methods in cybersecurity. arXiv preprint arXiv:2403.08701 (2024)
52. Zhang, Y., Hu, B., Chen, Z., Guo, L., Liu, Z., Zhang, Z., Liang, L., Chen, H., Zhang, W.: Multi-domain knowledge graph collaborative pre-training and prompt tuning for diverse downstream tasks. arXiv preprint arXiv:2405.13085 (2024)