Towards the automated population of Thesauri using BERT: a use case on the Cybersecurity domain

Elena Cardillo[1], Alessio Portaro[1], Maria Taverniti[1], Claudia Lanza[2] and Raffaele Guarasci[3]

Abstract The present work delves into innovative methodologies leveraging the widely used BERT model to enhance the population and enrichment of domain-oriented controlled vocabularies as Thesauri. Starting from BERT's embeddings, we extracted information from a sample corpus of Cybersecurity related documents and presented a novel Natural Language Processing-inspired pipeline that combines Neural language models, knowledge graph extraction, and natural language inference for identifying implicit relations (adaptable to thesaural relationships) and domain concepts to populate a domain thesaurus. Preliminary results are promising, showing the effectiveness of using the proposed methodology, and thus the applicability of LLMs, BERT in particular, to enrich specialized controlled vocabularies with new knowledge.

Key words: Thesauri, Domain-specific language modeling, Semantic analysis, Knowledge Extraction, LLMs

---

[1] Institute of Informatics and Telematics, National Research Council, Italy, e-mail: `elena.cardillo@iit.cnr.it`, e-mail: `maria.taverniti@cnr.it`.
[2] University of Calabria (UNICAL), Italy, e-mail: `claudia.lanza@unical.it`.
[3] Institute for High Performance Computing and Networking, National Research Council of Italy (CNR), Italy, e-mail: `raffaele.guarasci@icar.cnr.it`.
Although the authors have cooperated in the research work and in writing the paper, Elena Cardillo contributed to the Conceptualization of the paper, and an overall revision of the paper; Alessio Portaro formalized the experimental pipeline and carried out the experiments described in section 4; Maria Taverniti contributed to conceptualization and revision; Claudia Lanza specifically dealt with Section 3.1.; Raffaele Guarasci contributed to the Conceptualization of the paper, and the Methodology.

1 Introduction

The field of Natural Language Processing (NLP) has experienced remarkable growth, largely driven by the emergence of Large Language Models (LLMs). The advent of Transformer-based models, starting with the release of models like BERT [5], has marked a new era in NLP approaches, consistently elevating the standards for achievable results across various tasks such as text classification [31, 8, 1], sentiment analysis [17], and anaphora and coreference resolution [29, 7, 22, 10]. The profound potential of LLMs to encode linguistic knowledge has become a focal point of scholarly interest [14]. However, within this landscape, a growing discrepancy between high-resource and low-resource languages has become evident. To address this imbalance, the exploration of transferring linguistic knowledge from one language to another has become a prominent topic in the NLP field [12, 11].

This work starts from the need to propose innovative methods that have been shown to improve performance in various areas of the NLP field, to support the creation and integration of controlled vocabularies such as Thesauri, by identifying specific (non-explicit) relations and domain concepts relevant for indexing. In particular, we propose a NLP-inspired pipeline based on the wide-popular LLM BERT to extract the knowledge graph from a specialized domain thesaurus and then apply a natural language inference (NLI) method to extract potential relationships. The resource chosen for this purpose is the cybersecurity thesaurus published on the Italian Cybersecurity Observatory (OCS) web platform [1], containing more than 200 Italian/English indexing terms.

The paper is organized as follows: Section 2 offers an overview of recent related works. Following that, Section 3 describes in details the materials and the research methodology, encompassing information on the structural probe, the chosen LLM, and the dataset employed. Section 4 outlines the experimental evaluation and subsequently presents and discusses some preliminary results. Lastly, Section 5 provides some conclusions and alludes to potential future developments.

2 Related Works

Earning knowledge graphs has garnered significant attention recently, with advancements driven by LLMs. Early approaches to knowledge graph extraction primarily relied on rule-based systems and information retrieval techniques [21]. These methods often faced challenges in handling unstructured or ambiguous textual data and required extensive manual revision by experts. The release first of word embeddings [20] and then of large neural models of the language [5] shifted the paradigm by allowing unsupervised NLP techniques to be used in KG extraction as well. In particular, several studies have explored BERT's effectiveness in entity recognition

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

and relationship extraction. [34] and [36] show how fine-tuning BERT models on specific entity and relation extraction tasks significantly improve extraction accuracy. Moreover, BERT's ability to understand context and relationships is leveraged in classification tasks. Techniques such as attention mechanisms and multi-instance learning [35] showcase improvements in identifying and classifying complex relationships within textual data.

Concurrently, NLI has emerged as a complementary approach for knowledge graph construction. By leveraging BERT's contextual embeddings, NLI models aid in reasoning about entailment and contradiction, facilitating the extraction of implicit relationships within the text [3]. In recent years, studies have explored joint learning approaches that simultaneously address entity recognition and relation extraction tasks. As suggested by [18], BERT-based models exhibit the capacity to enhance both aspects concurrently, leading to more coherent and accurate knowledge graph construction.

3 Materials and Methods

This section describes in detail the resources and the neural model of the language (NLM) used in the experiment. In particular, the resources used are the following: *i)* the bilingual (Italian-English) cybersecurity thesaurus published on the mentioned Italian OCS website, namely the OCS Thesaurus, which is a controlled vocabulary offering a structured representation of the domain knowledge through semantic relationships between different indexed terms belonging to the field of study; *ii)* a corpus of cybersecurity regulations extracted from the OCS website using automated tools. Concerning the NLM method, we take into account the wide-popular BERT [5]. The choice was driven by the enormous popularity and versatility of the model, applied in recent years to a wide variety of tasks.

3.1 The OCS Thesaurus

Considering the specialized domain taken into account, that is Cybersecurity, characterized by a highly technical terminological distribution in the documentation [4], a preliminary task of this research activity was concentrated on the consultation of already existing resources aimed at formalizing this knowledge-domain. A first choice fell on the mentioned OCS bilingual Thesaurus realized as part of this project and built by taking into account an heterogeneous set of source texts from which a list of technical terms arranged through a semantic relationship structure were extracted (see [23], [26]). According to the ISO 25964:2013 standard (2013, p. 12) [25] 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 main purpose assigned to such a semantic resource is that of organizing the terminology characterizing a given sector-oriented domain of study in a way that supports the indexing operations and targeted knowledge discovery[2], as well as a terminological control over the information retrieval tasks (see [6]). Indeed, as Lykke (2001, 778)[19] argued, "the thesaurus is a tool that helps individual users to get an understanding of the collective knowledge domain". The terms reflecting the knowledge domain are semantically connected with each other following the rules provided by the ISO standards 25964-1:2011 [24] and 25964-2:2013 [25] according to which there are three main kinds of semantic relationships covered within thesauri:

1. Hierarchical relationship, marked with the tags Broader Term (BT) and Narrower Term (NT), points to the specificity connection between terms. This includes relationships of types "generic" (i.e., class-member relationship "IS-A"), "instance" and "partitive";
2. Equivalence relationship, marked with the tags Used (USE) and Used For (UF), manages the synonymy link between terms representing the same concept;
3. Associative relationship, marked with the tag Related Term (RT), denotes the coordination of terms belonging to the same category or to others.

These relationships are aimed at managing the conceptual framework of a specialized domain of study in order to guarantee a semantic tool able to normalize the information and avoid ambiguity in treating sector-oriented information [2]. Given these premises, constructing a semantic tool representing specialized fields of knowledge under the lens of a terminological network requires to rely on as much technical documentation as possible in order to detect the right amount of sector-oriented terms. The specialized corpus employed to populate the OCS Thesaurus and, above all, to shape its structure consists of 57 documents coming from authoritative legislative (e.g., national and international regulations, best practices) and popular sources (scientific journals on the subject), as well as glossaries on the domain (e.g., the NIST *Glossary of key information terms*[28] and ISO 27000:2016 Information technology — Security techniques — Information security management systems — Overview and vocabulary) [13] [15]. The number of terms in the OCS Thesaurus, extracted and selected alongside the supervision of domain experts and through the Term Frequency/Inverse Document Frequency (TF/IDF) statistical measure [33] is 238. To test the methodology proposed in this work we selected a sample of 5 documents in English, already part of the corpus used for the construction of the OCS Thesaurus, namely the legislative framework documentation on an European level and a specific set of relationships characterizing one Top Term (TT) of the resource, i.e., *Cybersecurity* consisting of 153 connecting terms. For instance, an example of a set of semantic relations taken into account is the following:

- **Hierarchy**: Cybersecurity *NT* Cyber risk management;
- **Equivalence**: Cybersecurity *UF* Information Security;
- **Association**: Cybersecurity *RT* Privacy.

3.2 Neural Language Model

Presently, among Transformer-based Neural Language Models (NLMs) used in NLP, BERT [5] holds a prominent position, with its efficiency and high performance well-established in the literature. The *BERT-base* version of this deep neural network architecture comprises 12 layers of decoder-only Transformers, each with 768-hidden dimensional states and 12 attention heads, totaling 110 million parameters. Conceptually rooted in the Transformer encoder architecture [32], BERT employs a multi-layer bidirectional design, pre-trained on extensive unlabeled text through two primary objectives: masked language modeling and next sentence prediction.

The versatility of BERT lies in its ability to provide robust context-dependent sentence representations, subsequently adaptable to diverse NLP tasks through a fine-tuning process tailored to specific requirements. This fine-tuning necessitates the adjustment of various hyperparameters, directly impacting the achievable outcomes. BERT's pre-training methodology centers on masked language modeling, where a portion of words in the training corpus is randomly masked. This allows the model to learn from both sentence directions while predicting the masked words. The choice between *cased* and *uncased* input vocabularies leads to two distinct pre-trained models. BERT's bidirectional analysis maintains significant generative capacity in deep constituent network layers, with outer layers adapted for task-specific fine-tuning. This dual-layered approach has established BERT as a benchmark model in recent literature. In BERT, each input word sequence begins with a special *[CLS]* token, generating a vector of size H (hidden layer size) as the output, representing the entire input sequence. Additionally, a unique *[SEP]* token must conclude each sentence within the input sequence.

Given an input sequence of words $t = (t_1, t_2, \dots, t_m)$, BERT's output is denoted as $h = (h_0, h_1, h_2, \dots, h_m)$, where $h_0 \in R^H$ serves as the final hidden state of the *[CLS]* token, offering a pooled representation for the entire input sequence.

Subsequently, $h_1, h_2, \dots, h_m$ represent the final hidden states of the remaining input tokens. In the context of fine-tuning BERT for classifying input sequences into K distinct text categories, the utilization of the final hidden state h_0 facilitates feeding a classification layer, followed by a softmax operation to convert category scores into likelihoods, as expounded by Sun et al. [30]:

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

where $W \in R^{KxH}$ is the parameter matrix of the classification layer.

4 Experimental Assessment

The proposed methodology starts from the work proposed in [3]. It is divided in two macro-steps as shown in figure 1:

1. **Concepts Identification and Extraction**: The first layer of the application aims to extract concepts and their relationships in the form of a Knowledge Graph (KG). As mentioned in 3, the corpus of documents used to run the experiments comprises legislative texts containing cybersecurity regulations. In order to have more semantically relevant knowledge graphs, a smaller set of documents was later selected by legal experts of the team from the initial texts' corpus.
   For the extraction of the KG we used part of the software developed by [27] within the scope of the Interlex project [16] . This tool first extracts concepts from the text in PDF files, leveraging the capabilities of the **SpaCy** [2] library which can infer PoS-tagged dependency tree (DPT). Using the DPT, the Interlex module extracts relevant concepts navigating nodes tagged as noun phrases. Once the concepts' set is created, it can search the tree for those tokens connecting concepts. The latter is used to describe a relation between the two concepts. This relation is simply in the same form as the sentence from which it was extracted, with the only difference of having two blanks as placeholders for the subject and the object of the sentence. This method for relations extraction has the apparent advantage of preserving the natural language structure, that is very useful when the resulting data is fed into an LLM. The Interlex module was preferred among other alternatives for KG extraction since triples composed of the two concepts (subject and object) and the relation connecting the two (predicate) come together to form a KG.

2. **Natural Language Inference using KG**: In the second layer, the extracted KG populates a Thesaurus, in this case the OCS Thesaurus mentioned in 3.1 using LLMs. The KG model is known to have much redundant semantic information because many instances of the same relation can appear in the same graph. Furthermore, the relations in a KG are typically subject-object relations, whereas the relations in a thesaurus are more linguistic correlations between concepts (as mentioned above, can be hierarchical, equivalence and associative relationships). To resolve this redundancy issue, we need to identify said linguistic relations. An expert would identify and classify the thesaural relationships using a classic approach to build a thesaurus. However, our scope was to automate this process, leveraging the enormous linguistic capabilities of modern LLMs like BERT. More specifically, we used a fine-tuned version of the BERT model further trained for NLI tasks. In particular, the model used is **HuggingFace**'s mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 [16]. The classic input for an NLI model is composed of a premise and a hypothesis, which the model can classify in three ways:

   - **Entailment**, meaning the hypothesis that can be inferred form the premises;
   - **Contradiction**, meaning the hypothesis that contradicts the premises;
   - **Neutral**, meaning the hypothesis that neither descends or contradicts the premises;

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

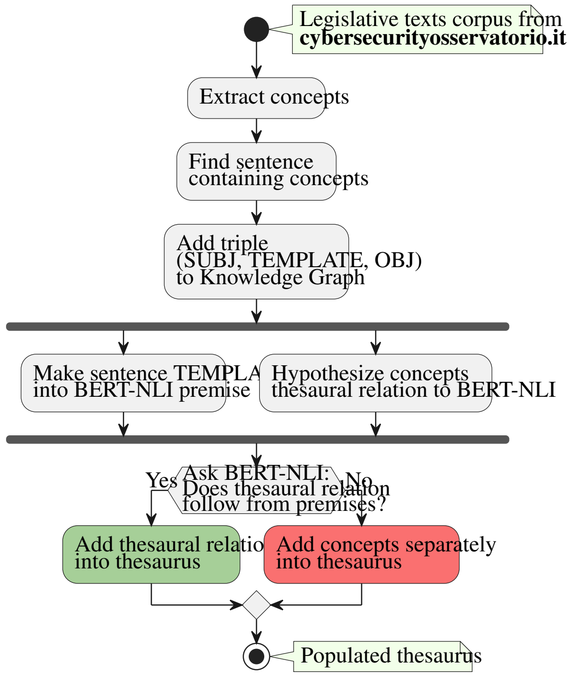

Fig. 1 Activity Diagram

Since the hypothesis we wanted to verify is a series of questions about the relations between each pair of concepts, the premise consisted in some context about the two concepts. For this task, we decided to include every relation in the KG connected to each pair concept as context. The relations were extracted as a template with some blanks for the subject and object of the sentence, so we replaced the blanks with the concepts that the said relation connects in the KG. Thanks to the Interlex library, once this operation is completed, the resulting sentence can be perfectly understandable in natural language, perfectly fitting the input expected by an NLP transformer-based model like BERT. Given a context and a hypothesis, the NLI model predicts a class. So in our task, when the class corresponded to **Entailment**, we created a relation in the resulting domain thesaurus between the two concepts. The type of relation matched the one we hypothesized as input for the NLI model. To give an example, from a triple of the KG extracted having as subject ”Security” and object ”Privacy”, we could identify the type of thesaural relationships among them (in this case RT) and verify its presence in the Thesaurus.

5 Conclusion and Future Works

This paper showed the preliminary steps of an approach that exploits modern neural models of language to extract/add information to a controlled vocabulary. Specifically, BERT was used to validate a pipeline that extracts concepts from a thesaurus on cybersecurity to conduct an NLI task. Exploiting BERT embeddings, the approach is able to extract information from a corpus of Cybersecurity-related documents. The proposed NLP-inspired pipeline seamlessly integrates NLMs, knowledge graph extraction, and NLI to discern implicit relations and domain concepts, enriching a domain thesaurus.

Preliminary findings demonstrated the robustness of the proposed methodology, highlighting the applicability of state-of-the-art Large Language Models in augmenting specialized controlled vocabularies with new knowledge. The outcomes underscore the potential for integrating BERT-based techniques to enhance the semantic richness and utility of domain-oriented thesauri, without relying on outdated lexicons. This research contributes to the advancement of methodologies for constructing and enriching, thus updating controlled vocabularies, providing a contemporary framework for knowledge extraction and relationship identification in domain-specific contexts. Even in a preliminary stage, the promising results obtained open up for further exploration and application in the rapid growing landscape of domain-specific knowledge management. For future work, given the great versatility of BERT and its availability in multilingual versions, the approach here proposed will be tested taking into account Italian terms contained in the thesaurus. It could open up numerous research possibilities regarding parallel analyses on the enrichment of the translations provided in the resource.

Furthermore, new approaches will be tested that promise to exceed the current performance achievable with NLMs, in order to be able to identity and extract, for example, precise types of hierarchical relationships (i.e., partitive, instance) to correctly connect the concepts extracted and properly shape/update the thesaurus structure. Moreover, novel perspectives on analysis have opened up with the emergence of Quantum NLP, or that sub-branch of NLP that uses methods derived from quantum theory to increase performance [9]. This could be investigated to improve the results in terms of accuracy and quantity of new knowledge that can be identified to populate the thesaurus.

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

References

1. 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)

2. Broughton, V.: Essential Thesaurus Construction. Facet (2006). DOI 10.29085/9781856049849
3. Chen, W., Ji, H.: Infer: Capturing implicit entity relations for knowledge graph completion using contextualized language models. arXiv preprint arXiv:2006.05295 (2020)
4. Claudia, L., Elena, C., Maria, T., Roberto, G.: Terminology management in cybersecurity thought knowledge organization systems: an italian use case. International journal on advances in security (1-2), 17–27 (2020)
5. 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 Papers), pp. 4171–4186. ACL, Minneapolis, Minnesota (2019). DOI 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423
6. Gabler, S.: Thesauri – a toolbox for information retrieval. Bibliothek Forschung und Praxis **47**(2), 189–199 (2023). DOI doi:10.1515/bfp-2023-0003. URL https://doi.org/10.1515/bfp-2023-0003
7. 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**, 75,144–75,157 (2022). DOI 10.1109/ACCESS.2022.3189956
8. 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**, 112,954 (2020). DOI https://doi.org/10.1016/j.eswa.2019.112954. URL https://www.sciencedirect.com/science/article/pii/S0957417419306724
9. Guarasci, R., De Pietro, G., Esposito, M.: Quantum natural language processing: Challenges and opportunities. Applied Sciences (Switzerland) **12**(11) (2022). DOI 10.3390/app12115651
10. Guarasci, R., Minutolo, A., Damiano, E., De Pietro, G., Fujita, H., Esposito, M.: ELECTRA for neural coreference resolution in italian. IEEE Access **9**, 115,643–115,654 (2021). DOI 10.1109/ACCESS.2021.3105278
11. 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**, 101,261 (2022)
12. 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)
13. Hazem, A., Daille, B., Claudia, L.: Towards automatic thesaurus construction and enrichment. In: B. Daille, K. Kageura, A.R. Terryn (eds.) Proceedings of the 6th International Workshop on Computational Terminology, pp. 62–71. European Language Resources Association, Marseille, France (2020). URL https://aclanthology.org/2020.computerm-1.9
14. Jawahar, G., Sagot, B., Seddah, D.: What does BERT learn about the structure of language? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3651–3657. ACL, Florence, Italy (2019). DOI 10.18653/v1/P19-1356. URL https://www.aclweb.org/anthology/P19-1356
15. Lanza, C.: Semantic Control for the Cybersecurity Domain: Investigation on the Representativeness of a Domain-Specific Terminology Referring to Lexical Variation. CRC Press (2022). DOI https://doi.org/10.1201/9781003281450citation-key
16. 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 (2022). URL https://osf.io/74b8k. Publisher: Open Science Framework
17. Li, W., Zhu, L., Shi, Y., Guo, K., Cambria, E.: User reviews: Sentiment analysis using lexicon integrated two-channel cnn–lstm family models. Applied Soft Computing **94**, 106,435 (2020). DOI https://doi.org/10.1016/j.asoc.2020.106435. URL https://www.sciencedirect.com/science/article/pii/S1568494620303756

18. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D.: Joint entity recognition and relation extraction as a multi-head selection problem. arXiv preprint arXiv:2201.10208 (2022)
19. Lykke, M.: A framework for work task based thesaurus design. Journal of Documentation **57**, 774–797 (2001). DOI 10.1108/EUM0000000007100
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
21. Miller, G.A.: Introduction to wordnet: An on-line lexical database. International Journal of Lexicography **3**(4), 235–244 (1990)
22. 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), 22,493 – 22,518 (2022). DOI 10.1007/s00521-022-07641-3
23. Nielsen, M.L.: Thesaurus construction: Key issues and selected readings. Cataloging & Classification Quarterly **37**(3-4), 57–74 (2004). DOI 10.1300/J104v37n03_05. URL `https://doi.org/10.1300/J104v37n03_05`
24. Organization, I.S.: ISO 25964-1:2011 Information and documentation — Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval (2011)
25. Organization, I.S.: ISO 25964-2:2013 Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies (2013)
26. Q. Z., B., S., H.: 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 (2014). DOI 10.3115/v1/W14-4807. URL `https://www.aclweb.org/anthology/W14-4807`
27. 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)
28. of Standards, N.I., Technology: Glossary of key information security terms. Tech. rep., NIST Interagency or Internal Report (NISTIR) 7298 Rev. 2 (May 2013)
29. Sukthanker, R., Poria, S., Cambria, E., Thirunavukarasu, R.: Anaphora and coreference resolution: A review. Information Fusion **59**, 139–162 (2020). DOI https://doi.org/10.1016/j.inffus.2020.01.010. URL `https://www.sciencedirect.com/science/article/pii/S1566253519303677`
30. Sun C. Qiu X., X.Y., X., H.: How to fine-tune bert for text classification? In: China National Conference on Chinese Computational Linguistics, p. 194–206. Springer (2019)
31. Trotta, D., Guarasci, R., Leonardelli, E., Tonelli, S.: Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In: M.F. Moens, X. Huang, L. Specia, S.W.t. Yih (eds.) Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 2929–2940. Association for Computational Linguistics, Punta Cana, Dominican Republic (2021). DOI 10.18653/v1/2021.findings-emnlp.250. URL `https://aclanthology.org/2021.findings-emnlp.250`
32. 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). URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need`
33. 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). URL `https://api.semanticscholar.org/CorpusID:18303048`
34. Wu, S., Dredze, M.: Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 833–844. Association for Computational Linguistics, Hong Kong, China (2019). DOI 10.18653/v1/D19-1077. URL `https://www.aclweb.org/anthology/D19-1077`

35. Zhang, S., Wang, Z., Tang, J.: Bert for joint entity and relation extraction via context-aware coreference resolution. Information Processing & Management **58**(5), 102,356 (2021)
36. Zhang, Y., Zhang, Y., Ji, D.: Kg-bert: Bert for knowledge graph completion. arXiv preprint arXiv:2002.00388 (2020)