A Systematic Review of Automatic Term Extraction: What Happened in 2022?

Giorgio Maria Di Nunzio, Stefano Marchesin, Gianmaria Silvello
Department of Information Engineering
University of Padova, Italy
[giorgiomaria.dinunzio,stefano.marchesin,gianmaria.silvello]@unipd.it

Abstract

Automatic Term Extraction (ATE) systems have been studied for many decades as, among other things, one of the most important tools for tasks such as information retrieval, sentiment analysis, named entity recognition, and others. The interest in this topic has even increased in the recent years given the support and improvement of the new neural approaches. In this paper, we present a follow-up on the discussions about the pipeline that allows extracting key terms from medical reports, presented at MDTT 2022, and analyze the very last papers about ATE in a systematic review fashion. We analyzed the journal and conference papers published in 2022 (and partially in 2023) about ATE and cluster them into subtopics according to the focus of the papers for a better presentation.

Abstract

Les systèmes d'extraction automatique de termes (ATE) ont été étudiés pendant de nombreuses décennies comme, entre autres, l'un des outils les plus importants pour des tâches telles que la recherche d'informations, l'analyse des sentiments, la reconnaissance d'entités nommées, etc. L'intérêt pour ce sujet a même augmenté ces dernières années compte tenu du soutien et de l'amélioration des nouvelles approches neuronales. Dans cet article, nous présentons un suivi des discussions sur le pipeline qui permet d'extraire des termes clés des rapports médicaux, présentés à MDTT 2022, et analysons les tout derniers articles sur l'ATE de manière systématique. Nous avons analysé les articles de revues et de conférences publiés en 2022 (et partiellement en 2023) sur l'ATE et les avons regroupés en sous-thèmes en fonction de l'objet des articles pour une meilleure présentation.

preprint

1. Introduction

Computational Terminology (CT) is a multidisciplinary research area where computer scientists, information science specialists, linguists and, of course, terminologists design and develop automatic approaches applied to specialized texts (Bourigault et al., 2001). From an historical point of view, CT can be placed at the beginning of the 1990s with the first international conferences dedicated to this area. Given the computational character that strongly delineates this topic, it is no coincidence that these first initiatives were engineering and artificial intelligence conferences, such as the *Terminology and Knowledge Engineering* (TKE)[1], and the *Conférence Internationale Terminologie et Intelligence Artificielle* (TIA)[2].

The first volume entirely dedicated to this subject was published in 2001 (Bourigault et al., 2001). This collection of handpicked articles offers diverse perspectives on CT from researchers in a variety of fields: from automatic text parsing to terminology storage and use, from linguists to applied linguistics specialists, from information retrieval to artificial intelligence. In that volume, Automatic Term Extraction (ATE) is deeply studied and evaluated as a support to information retrieval problems, or to problems of translation and alignment of

[1] https://dblp.org/db/conf/tke/index.html

[2] https://www.irit.fr/TIA09/commencer_ici.htm

multilingual terminological databases. The special issues dedicated to CT (Drouin et al., 2015, 2018), in the same way, collect a selection of articles that offer a panorama of approaches oriented towards the automatic extraction of terms in which we observe the increasingly important presence of hybrid methods using artificial neural networks and a representation of words based on the distributional hypothesis (Mikolov et al., 2013).

Actually, in the mid-90s (Kageura & Umino, 1996) present the first systematic review of approaches to ATE starting from an analysis of automatic term indexing methods from the 1950s (Luhn, 1957), going through one of the most important quantitative methods in the history of information retrieval, the specificity of a term (Sparck Jones, 1972). In addition, the review offers hints for some hypotheses for the definition of term and termhood (a sort of quantification of how much `a term is a term'): i) A frequently appearing (terminological) unit in a domain is likely to be a term from that domain; ii) A unit that appears only in one domain is likely to be a term from that domain; iii) A unit that appears relatively more frequently in a specific domain than in general is likely to be a term of that domain; iv) A unit whose occurrence is somehow affected by (a) domain(s) is likely to be a term.

Given the increasing interest in this research area, in this paper, we continue the analysis that we started at the 1st Multilingual Digital Terminology Today conference (Di Nunzio, Henrot, et al., 2022) about the advances in ATE that, in our case, was dedicated to the medical domain (Di Nunzio, Marchesin, et al., 2022).

Our objective is to analyze the most recent literature on ATE by means of a systematic review of the papers published in 2022 (and some at the beginning of 2023). The final goal is to give an overview of the most promising directions of this area of research as a fundamental bridge among different fields as well as a valuable tool for the creation of multilingual terminological databases.

The paper is organized as follows: in Section 2, we present the methodology that we followed to select the papers to review; in Section 3, we discuss our analysis of the state-of-the-art by presenting the papers divided into categories that highlight the focus of each contribution. In Section 4, we give our final remarks.

2. Methodology

In order to perform the systematic review, we proceeded with the following methodology: we used the Google Scholar database to retrieve documents with the key-phrase "automatic term extraction" or "automated term extraction". Subsequently, we filtered the papers that have been published since 2022 (and including 2023). The result was a list of 176 candidate papers[3]. From this list, we manually selected only those papers that were published in international journals or in international conferences where a peer review of the paper was explicitly mentioned in the call for papers of the conference. We intentionally did not include in this survey workshop papers, preprints, or arXiv papers. Then, we removed from this filtered set of papers, those who mentioned ATE only as a secondary topic of the paper or in the related works (i.e., the key phrase "automatic term extraction" was present in the papers, but the focus of the paper itself was on ATE). After this last filtering step, we obtained a list of 24 papers. A last comment about recent works on ATE, at the time of the preparation of this manuscript, we found on arXiv an interesting survey on recent advances on ATE that we feel obliged to mention giving the synchronicity with this manuscript (Tran et al., 2023). Despite the focus of this survey is mostly on neural transformer-based models, we found it very interesting also from the point of view of the analysis of the datasets and metrics used for the evaluation of the performance in ATE. Finally, we decided to organize the presentation of the analysis of the content of the papers according to a qualitative clustering of the papers in the following main topics: ATE tools, decision-making, knowledge modeling, multilinguality, word embeddings, evaluation. Of course, some papers address more than one of these topics, but we decided to present each paper in only one of these categories to avoid repetitions.

3. Analysis

In this initial part of the analysis, we want to list the different definitions of ATE that we found in the papers to show the slightly different nuances of this field, according to the authors. For example (Nugumanova et al., 2022) focuses on the pipeline for the ATE: "Automatic term extraction, also

[3] The last search was done on January 31 2023.

pre-print

known as automatic term recognition, is a task aimed at detecting domain terms in a given corpus of documents. Traditionally, methods for solving this problem include three stages: 1) preprocessing and term candidates extracting, 2) term candidates scoring, and 3) term candidate ranking." On the other hand (Nomoto, 2022) underlines the (shared) difficulty of the notion of keywords addressed in different areas: "The notion of 'keyword' has long defied a precise definition. […] History witnessed the rise of two major schools of thought, one in terminology science (TS) and the other in information retrieval (IR). […] Terminologists are generally concerned with finding terms that are specific to a particular technical domain, useful to organize knowledge relating to that domain, while people in information retrieval are focused more on identifying terms (which they call indexing terms) capable of distinguishing among documents to improve document retrieval". The difficulty of the distinction of a term is also expressed by (Terryn et al., 2022): "[…] ATE is usually considered a semi-automatic process that requires human validation, since it is such a difficult task that cannot yet be perfectly automated. One of the main difficulties for ATE lies in the ambiguous distinction between terms and general language." A difficulty that is also related to the task itself: "terminology extraction is a complex and difficult task, and requires certain linguistic knowledge and a related field background" (Zhao et al., 2022). Finally, other authors highlight the opportunities that ATE gives to support other research activities: "By easing the time and effort needed to manually extract the terms, ATE is not only widely used for terminographical tasks but also contributes to several complex downstream tasks (e.g., machine translation, […])" (Tran, Martinc, Pelicon, et al., 2022), or "The results [of ATE] can either be used directly to facilitate term management for, e.g., terminologists and translators, or as a preprocessing step for other tasks within natural language processing (NLP) […]." (Terryn et al., 2022)

In the following sections, we present a summary of the main objective and findings of each paper clustered by subtopics.

3.1 ATE Tools

In this section, we review the works that deal with the design, implementation, and evaluation of tools for ATE, which is also the topic of the original paper presented at MDTT 2022 (Di Nunzio et al., 2022). ATE tools can be complex systems that allow users to perform a variety of operations for specific purposes related to the specific field. The case of the medical domain is tackled by (Marchesin et al., 2022) and (Thukral et al., 2023). In both cases, the main starting point is the electronic health record and the fact that essential information is contained in clinical narratives which are described in natural language. If clinical narratives were represented in a format understandable by AI applications, then a better diagnosis or decision could be obtained.

In this sense (Marchesin et al., 2022) propose a tool to overcome the limitation of the necessity to have annotated data by means of unsupervised NLP techniques to automatically extract critical information from pathology reports and use it for different digital pathology applications, such as automatic report annotation, pathological knowledge visualization. In this regard, they present the Semantic Knowledge Extractor Tool (SKET), an unsupervised hybrid knowledge extraction system that combines an expert system with pre-trained ML models to extract knowledge from pathology reports. (Thukral et al., 2023), on the other hand, try to translate clinical narratives effectively while retaining the medicinal vocabulary and semantics by means of a tool for Named Entity Recognition in conjunction with the validation of the medical expert.

Another example in a different domain is the one proposed by (Panoutsopoulos et al., 2022). The authors focus on the implementation of a custom Named Entity Recognition tool aiming to identify and extract agricultural terms from text in order to provide data-driven insights for this economic sector.

On the other hand, the contribution provided by (Martín-Chozas et al., 2022) is dedicated to the implementation of TermitUp, a tool that puts together pieces of language technology previously isolated, and improves them to build a pipeline that generates as output a multilingual terminology semantically enriched with data from the Linguistic Linked Open Data (LLOD)[4] - a movement about publishing data for linguistics and natural language processing - and published in open formats.

3.2 Decision Making

[4] https://linguistic-lod.org

In this section, we analyze the papers that tackle the issues related to decision-making systems and, in general, the problem of finding, or assess, the best alternative among different options. Three following papers are all collocated within the same domain: the maritime and naval domain. In this domain, the intelligent self-decision-making of naval operations is of great significance to the research of auxiliary decision-making for naval operations (Zhao et al., 2022; Andersen, 2022). The same importance is given to the recovery and enrichment of maritime heritage, such as documents and archives with drawings of ships and maps (Mouratidis et al., 2022). The fourth paper in this category concerns the identification of cybersecurity material and the consequent decisions to prevent digital threats (Prayogo et al., 2022).

In (Zhao et al., 2022), the authors present an approach for the extraction of domain terms in the operational planning field and the synonym extraction between terms. In the proposed method, the data to be processed comes from operational planning documents. Such documents (and domain) show sentences that have a rather different word segmentation that require a manual intervention to develop an effective and reusable set of terms.

(Andersen, 2022) discusses implementation aspects about the development of a terminology in the maritime domain as well as methodological questions like "How can we, with limited funding, to a maximal degree utilize existing language resources to develop a terminology at a relatively low cost?" In this respect, the author analyzes linguistic approaches that consider the fact that terms take certain syntactic forms and tend to follow certain morphosyntactic patterns.

Finally, in (Prayogo et al., 2022), the authors describe the process of building preventive measures against cybersecurity threats by discovering and understanding new vulnerabilities from cybersecurity-related material that are mainly communicated via textual channels online. The authors describe an architecture called "Attended over Distributed Specificity" that was introduced for ATE in cybersecurity.

3.3 Knowledge Modeling

In this section, we review two papers that deal with the automatic organization of knowledge in a specific domain and organize this knowledge into structures that can be used by humans or machines. In particular, the two works try to infer the semantic similarity among terms by means of functions, TextRank (Zhang et al., 2022) and ETBRrank (Wu et al., 2022), that extract meaningful terms and their relations within the domain.

In (Zhang et al., 2022), the authors discuss the problem of requirements analysis and, in particular, the terminology of requirements that helps the stakeholders share a common understanding of the key concepts within a specific domain. The authors use the smart home domain as a case study, and they construct an illustrative feature model to demonstrate that the terms extracted by the proposed adaptation of TermRank create an implicit hierarchy structure that can help the organization of the requirements analysis.

In (Wu et al., 2022), the authors propose an Automatic Biterm Extraction (ABE) – i.e., a word co-occurrence pattern – to discover emerging (and possibly unknown) topics in research papers. This is an interesting case study where the terminology itself is not consolidated and still in development, while the topics are either unnamed or named differently by several authors. Given two papers, the proposed approach - Emerging Topic BiTerm Rank - use paper titles to automatically backtrack the origin of the new topics from two co-occurring super topics.

3.4 Multilinguality

The questions related to multilinguality in terminology, which is also one of the main topics of the MDTT conference,[5] are tackled by four papers from different perspectives: scarce resource languages, accuracy and consistency, term "unithood" (Kageura & Umino, 1996) between languages.

In (Karaman et al., 2022), the authors study the problem of using the data of one language (English) to train an ATE model in a different language with limited linguistic resources (Turkish in this case). The results of a joint multilingual neural model trained on Turkish-English abstracts of theses, show a significant improvement in the multilingual ATE process.

(Jia et al., 2022) consider the perspective of domain-specific user-provided bilingual terminologies in the field of e-commerce. The authors propose a new task which is to discover bilingual terminologies

[5] http://mdtt2023.dei.unipd.it/en/

from comparable data in the e-commerce field. The task is to align a sentence in the source language with a sentence in the target language, extract the terms in the two languages, and link them.
In (Liwei, 2022), the author takes Chinese patent literature as the research object and proposes a method of extracting technical terms that combines grammatical rules and statistical methods. The challenge, in this context, is how to extend the ATE methods – usually studied in English or Romance languages – to the Chinese language. In particular, the author focuses on the difference of an English "word" used as the linguistic unit compared to the Chinese "character" to express a complete meaning.
(Barbero, 2022) presents a methodology that involves specialized corpora exploration in comparison to common language reference corpora. The authors base their work on the data collected from the compilation and treatment of a bilingual – European Portuguese/Italian – comparable corpus of specialized texts on Public Art. The main issues analyzed in the paper concern a better use of frequency analysis to improve the lists extracted from the specialized corpus (before submitting it to expert validation) and the evaluation of lexical/syntactic patterns to isolate specific semantic relations.

3.5 Word Embeddings

A word embedding is a learned representation, usually through neural networks, for text where words that have the same meaning have a similar representation. In this way, researchers can automatically discover and evaluate relationships among terms by measuring the "distance" between these representations. The use of word embeddings in terminology has been confronted in three papers.
(Vintar & Martinc, 2022) propose an interdisciplinary perspective with the aim of building a new multilingual and multimodal interactive knowledge base tailored to the needs of different types of users in the domain of karstology (a subfield of geomorphology). In particular, they use word embeddings to both identify words expressing a specific semantic relation and extract multiword units which contain the target relation.
In (Liu et al., 2022), the authors focus on the following problem: given two corpora, a domain corpus DC and general corpus, GC, the authors zot want to rank all terms, represented with word embeddings, in the shared vocabulary between DC and GC, such that the top of the ranking is enriched with domain terms when their meaning differs from common usage.
Finally (Li, 2022) focuses on the Internet of Things domain and the importance of the correct translation of terms for the exchange of scientific information. The author proposes an interactive approach where the expert analyze the results of a neural network model that produces a cluster of terms that are semantically related.

3.6 Evaluation

In this last section, we want to present those works the aim of which was to evaluate ATE approaches using standard datasets. In particular, we found six papers that used the same dataset, the Annotated Corpora for Term Extraction Research (ACTER)[6] dataset.
The authors of the papers (Tran, Martinc, Pelicon, et al., 2022) and (Tran, Martinc, Doucet, et al., 2022) compare the multilingual learning to the monolingual learning in the cross-domain sequence-labeling term extraction task (Gooding & Kochmar, 2019). They examine the cross-lingual effect of rich-resource training language over fewer resources one, such as Slovenian. The results demonstrate a promising impact of multilingual and cross-lingual cross-domain transfer learning.
In a similar fashion (Terryn et al., 2022) interpret ATE as a sequential labeling task, where each token in a text is classified as (part of) a term or not. The authors employ this strategy for ATE in a monolingual and multilingual setting and evaluate different models.
(Nugumanova et al., 2022) use the ACTER dataset to evaluate the performance of the term extraction approach that uses a different mathematical model. In particular, they use a non-negative matrix factorization approach to represent the documents in the collection. This approach does not require training data and is invariant both to the domain and to the language.
(Hazem et al., 2022) perform an extensive study of neural approach named Bidirectional Encoder Representations from Transformers BERT for ATE as a sequence labeling method. They study both the cross-domain and cross-lingual scenarios thanks to transfer learning. The results show that BERT can transfer learning across domains and languages,

[6] https://github.com/AylaRT/ACTER

even when there is a limited availability of annotations.

A different perspective is described by (Awwad et al., 2022). The main issue is how to build English Arabic scientific glossaries based on ATE. The authors contextualize this approach within the domain of technology and science – where new concepts and terminologies emerge very quickly – and, in particular, in terms of the urge for technological resources to increase the pace of the translation output at lower costs.

As a final remark, we want to highlight the fact that these state-of-the-art approaches, on this specific ACTER dataset, achieve precision and recall values that range between 35% to 70% at most, with a weighted average (F1 score) between the two that is very rarely greater than 60%. These results show that, at the present time, the ATE approaches are good but not excellent, and that a human-in-the-loop approach is necessary to recover terms that were discarded or remove unwanted terms.

4. Conclusions

In this systematic survey, we have given an overview of the most recent literature on Automatic Term Extraction, and we have tried to focus on the current challenges (and possibly future directions) of this very active research field. Neural models, in particular pre-trained models and transfer learning have been dominating the scene in the last months. These models are very promising in terms of the impact on scarce resource languages; at the same time, their performance – in terms of proportion of terms correctly recognized – shows that the intervention of the expert of the field is still the key point in producing a high-quality multilingual terminology. At the same time, multilinguality has been gaining a lot of attention, especially in terms of the re-use of ATE models that are trained in one language or specific domain in other languages and domains.

In the future, we believe that a major aspect would be that of building a collaborative task for the creation of annotated datasets for both the training and evaluation of models as well as a shared repository of terminological database.

References

Andersen, G. (2022). Utilising heterogeneous language resources for term extraction in maritime domains. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *28*(1), 1–36. https://doi.org/10.1075/term.20024.and

Awwad, H., Sawalha, M., Allawzi, A., & Yagi, S. (2022). Building translator-oriented English-Arabic physics glossary from domain corpus. *International Journal of Speech Technology*. https://doi.org/10.1007/s10772-022-10001-0

Barbero, C. (2022). CQL Grammars for Lexical and Semantic Information Extraction for Portuguese and Italian. In V. Pinheiro, P. Gamallo, R. Amaro, C. Scarton, F. Batista, D. Silva, C. Magro, & H. Pinto (Eds.), *Computational Processing of the Portuguese Language* (pp. 376–386). Springer International Publishing. https://doi.org/10.1007/978-3-030-98305-5_35

Bourigault, D., Jacquemin, C., & L'Homme, M.-C. (2001). *Recent Advances in Computational Terminology*. John Benjamins. https://www.jbe-platform.com/content/books/9789027298164

Costa, R. (2013). Terminology and Specialised Lexicography: Two complementary domains. *Lexicographica*, *29*(2013), 29–42. https://doi.org/10.1515/lexi-2013-0004

Di Nunzio, G. M., Henrot, G. M., Musacchio, M. T., & Vezzani, F. (Eds.). (2022). In *Proceedings of the 1st International Conference on Multilingual Digital Terminology Today* (Vol. 3161). CEUR. https://ceur-ws.org/Vol-3161/#preface

Di Nunzio, G. M., Marchesin, S., & Silvello, G. (2022). Terminology Extraction in Electronic Health Records. The ExaMode Project (poster). In G. M. D. Di Nunzio, G. M. Henrot, M. T. Musacchio, & F. Vezzani (Eds.), *Proceedings of the 1st International Conference on Multilingual Digital Terminology Today* (Vol. 3161). CEUR. https://ceur-ws.org/Vol-3161/#poster1

Drouin, P., Grabar, N., Hamon, T., & Kageura, K. (2015). Introduction to the Special Issue: Terminology across languages and domains: *Terminology*, *21*(2), 139–150. https://doi.org/10.1075/term.21.2.01dro

Drouin, P., Grabar, N., Hamon, T., Kageura, K., & Takeuchi, K. (2018). Computational terminology and filtering of terminological information: Introduction to the special issue: *Terminology*, *24*(1), 1–6. https://doi.org/10.1075/term.00010.dro

Gooding, S., & Kochmar, E. (2019). Complex Word Identification as a Sequence Labelling Task. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1148–1153. https://doi.org/10.18653/v1/P19-1109

Hazem, A., Bouhandi, M., Boudin, F., & Daille, B. (2022). Cross-lingual and Cross-domain Transfer Learning for Automatic Term Extraction from Low Resource Data. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 648–662. https://aclanthology.org/2022.lrec-1.68

Humbley, J. (1997). Is terminology specialized lexicography? The experience of French-speaking

countries. *HERMES - Journal of Language and Communication in Business*, *18*, Article 18. https://doi.org/10.7146/hjlcb.v10i18.25410

Jia, H., Gu, S., Zhang, Y., & Duan, X. (2022). Bilingual Terminology Extraction from Comparable E-Commerce Corpora. *2022 International Joint Conference on Neural Networks (IJCNN)*, 01–08. https://doi.org/10.1109/IJCNN55064.2022.9892544

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *3*(2), 259–289. https://doi.org/10.1075/term.3.2.03kag

Karaman, I. N., Cicekli, I., & Ercan, G. (2022). Automatic Term Extraction with Joint Multilingual Learning. *2022 7th International Conference on Computer Science and Engineering (UBMK)*, 159–164. https://doi.org/10.1109/UBMK55850.2022.9919455

Li, Y. (2022). Construction of Internet of Things English terms model and analysis of language features via deep learning. *The Journal of Supercomputing*, *78*(5), 6296–6317. https://doi.org/10.1007/s11227-021-04130-7

Liu, Y., Medlar, A., & Głowacka, D. (2022). Lexical ambiguity detection in professional discourse. *Information Processing & Management*, *59*(5), 103000. https://doi.org/10.1016/j.ipm.2022.103000

Liwei, Z. (2022). Chinese technical terminology extraction based on DC-value and information entropy. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-23209-6

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, *1*(4), 309–317. https://doi.org/10.1147/rd.14.0309

Manandise, E. (2022). Extracting Domain Terms from Data Model Elements. In P. Rosso, V. Basile, R. Martínez, E. Métais, & F. Meziane (Eds.), *Natural Language Processing and Information Systems* (pp. 267–278). Springer International Publishing. https://doi.org/10.1007/978-3-031-08473-7_24

Marchesin, S., Giachelle, F., Marini, N., Atzori, M., Boytcheva, S., Buttafuoco, G., Ciompi, F., Di Nunzio, G. M., Fraggetta, F., Irrera, O., Müller, H., Primov, T., Vatrano, S., & Silvello, G. (2022). Empowering digital pathology applications through explainable knowledge extraction tools. *Journal of Pathology Informatics*, *13*, 100139. https://doi.org/10.1016/j.jpi.2022.100139

Martín-Chozas, P., Vázquez-Flores, K., Calleja, P., Montiel-Ponsoda, E., & Rodríguez-Doncel, V. (2022). TermitUp: Generation and enrichment of linked terminologies. *Semantic Web*, *13*(6), 967–986. https://doi.org/10.3233/SW-222885

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In Y. Bengio & Y. LeCun (Eds.), *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. http://arxiv.org/abs/1301.3781

Mouratidis, D., Mathe, E., Voutos, Y., Stamou, K., Kermanidis, K. L., Mylonas, P., & Kanavos, A. (2022). Domain-Specific Term Extraction: A Case Study on Greek Maritime Legal Texts. *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*, 1–6. https://doi.org/10.1145/3549737.3549751

Nomoto, T. (2022). Keyword Extraction: A Modern Perspective. *SN Computer Science*, *4*(1), 92. https://doi.org/10.1007/s42979-022-01481-7

Nugumanova, A., Akhmed-Zaki, D., Mansurova, M., Baiburin, Y., & Maulit, A. (2022). NMF-based approach to automatic term extraction. *Expert Systems with Applications*, *199*, 117179. https://doi.org/10.1016/j.eswa.2022.117179

Panoutsopoulos, H., Brewster, C., & Espejo-Garcia, B. (2022). Developing a Model for the Automated Identification and Extraction of Agricultural Terms from Unstructured Text. *Chemistry Proceedings*, *10*(1), Article 1. https://doi.org/10.3390/IOCAG2022-12264

Prayogo, N., Amjadian, E., McDonnell, S., & Abid, M. R. (2022). Context-Aware Attended-over Distributed Specificity for Information Extraction in Cybersecurity. *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 0359–0367. https://doi.org/10.1109/IEMCON56893.2022.9946567

Sparck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Terryn, A. R., Hoste, V., & Lefever, E. (2022). Tagging terms in text: A supervised sequential labeling approach to automatic term extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *28*(1), 157–189. https://doi.org/10.1075/term.21010.rig

Thukral, A., Dhiman, S., Meher, R., & Bedi, P. (2023). Knowledge graph enrichment from clinical narratives using NLP, NER, and biomedical ontologies for healthcare applications. *International Journal of Information Technology*. https://doi.org/10.1007/s41870-022-01145-y

Tran, H. T. H., Martinc, M., Caporusso, J., Doucet, A., & Pollak, S. (2023). *The Recent Advances in Automatic Term Extraction: A survey* (arXiv:2301.06767). arXiv. https://doi.org/10.48550/arXiv.2301.06767

Tran, H. T. H., Martinc, M., Doucet, A., & Pollak, S. (2022). Can Cross-Domain Term Extraction Benefit from Cross-lingual Transfer? In P. Pascal & D. Ienco (Eds.), *Discovery Science* (pp. 363–378). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-18840-4_26

Tran, H. T. H., Martinc, M., Pelicon, A., Doucet, A., & Pollak, S. (2022). Ensembling Transformers

for Cross-domain Automatic Term Extraction. In Y.-H. Tseng, M. Katsurai, & H. N. Nguyen (Eds.), *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries* (pp. 90–100). Springer International Publishing. https://doi.org/10.1007/978-3-031-21756-2_7

Vezzani, F. (Ed.). (2022). *Terminologie numérique: Conception, représentation et gestion*. Peter Lang International Academic Publishers. https://doi.org/10.3726/b19407

Vintar, Š., & Martinc, M. (2022). Framing karstology: From definitions to knowledge structures and automatic frame population. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *28*(1), 129–156. https://doi.org/10.1075/term.21005.vin

Wu, J., Huang, G., & Zarei, R. (2022). ETBTRank: Ranking Biterms in Paper Titles for Emerging Topic Discovery. In G. Long, X. Yu, & S. Wang (Eds.), *AI 2021: Advances in Artificial Intelligence* (pp. 775–784). Springer International Publishing. https://doi.org/10.1007/978-3-030-97546-3_63

Wüster, E. (1968). *The Machine Tool: An Interlingual Dictionary of Basic Concepts ; Comprising and Alphabetical Dictionary and a Classified Vocabulary with Definitions and Illustrations ; Prepared Under the Auspices of the United Nations Economic Commission for Europe and Under the Direction of*. Technical Press.

Zhang, J., Chen, S., Hua, J., Niu, N., & Liu, C. (2022). Automatic Terminology Extraction and Ranking for Feature Modeling. *2022 IEEE 30th International Requirements Engineering Conference (RE)*, 51–63. https://doi.org/10.1109/RE54965.2022.00012

Zhao, X., Wang, C., Cui, P., & Sun, G. (2022). Operational Rule Extraction and Construction Based on Task Scenario Analysis. *Information*, *13*(3), Article 3. https://doi.org/10.3390/info13030144

preprint