Benchmarking Automatic Tools for Neologisms Extraction: Issues and Challenges

Giorgio Maria Di Nunzio[1]

[1]Department of Information Engineering, University of Padova, Via Gradenigo 6/b, 35131 Padova, Italy

Abstract

Human language is constantly evolving, driven by societal, technological, and cultural shifts, which lead to the creation of new terms and expressions. The rise of digital platforms, including social media and academic publications, has accelerated the introduction and spread of these neologisms. This paper explores current advancements and challenges in benchmarking automated and semi-automated tools for extracting neologisms. In particular, we will discuss challenges in dataset creation and evaluation procedures, such as defining neologisms, ensuring diverse text sources, managing annotation variability, and evaluating these tools.

Keywords

neologisms extraction, dataset creation, evaluation methodology

1. Introduction

Human language, by its nature, is always evolving and it generates newly coined terms or expressions that emerge in response to societal, technological, and cultural changes. The research into the detection and understanding of neologisms has increased exponentially in the last years due to the proliferation of digital communication platforms, including social media, academic publications, and technical documents. In fact, this panorama of available platforms has accelerated the introduction and dissemination of such novel linguistic elements which gives a unique opportunity for developing (semi-)automatic techniques for neologism extraction [1, 2].

Neologism extraction is a critical task for various fields, including computational linguistics, lexicography, and natural language processing (NLP). Traditional manual approaches to tracking linguistic evolution are labor-intensive and time-consuming, highlighting the need for automated or semi-automated methodologies. Identifying emerging terms can support the update of lexical resources, improve machine translation systems, and provide insights into societal trends[3, 4].

Fully automated approaches typically leverage large-scale corpora and machine learning techniques to identify candidate neologisms based on statistical analysis, linguistic patterns, or contextual novelty. These systems often incorporate dictionary comparisons, word frequency analysis, and morphological evaluation. Advances in deep learning and pretrained language models [5, 6, 7] have further enhanced the precision of such techniques by enabling context-aware evaluations of word novelty [3, 8, 9].

Semi-automated methods, on the other hand, combine computational efficiency with human expertise [1, 10]. These approaches may flag potential neologisms for manual validation, allowing domain experts to assess their linguistic legitimacy and relevance. By integrating human judgment, semi-automated systems to balance scalability and accuracy, making them particularly useful for specialized domains such as scientific literature or emerging technologies.

Multilinguality also plays a crucial role in automatic neologism extraction, as lexical innovation does not occur in isolation within a single language. Many neologisms emerge through cross-linguistic influence, such as borrowings from dominant languages or calques that adapt foreign terms into native structures. Moreover, different languages exhibit distinct morphological and syntactic processes for word formation, necessitating language-specific adaptation in extraction methodologies. A multilingual

1st International Workshop on Terminological Neologism Management (NeoTerm 2025), June 18, 2025, Thessaloniki, Greece.

giorgiomaria.dinunzio@unipd.it (G. M. Di Nunzio)

0000-0001-7116-9338 (G. M. Di Nunzio)

approach would enable comparative studies on neologism diffusion, tracking how new terms spread across linguistic and cultural boundaries. Developing resources that support multiple languages enhances the generalizability and applicability of neologism extraction tools, making them more robust for global linguistic research and practical applications in translation, lexicography, and information retrieval [3, 4, 8, 11, 10].

Despite recent progress, several challenges remain in the development of effective neologism extraction systems. These include distinguishing genuine neologisms from typographical errors, handling polysemy, and detecting subtle shifts in meaning for existing terms. Moreover, the rapid evolution of language in social media environments demands adaptive models capable of processing informal and creative language variations.

In this paper, we aim to provide a preliminary overview of the state-of-the-art techniques for benchmarking (semi-)automated tools for the extraction of neologisms, highlight their strengths and limitations, and suggest directions for future research. In particular, we will focus on the required specialized datasets that capture the dynamic nature of language in order to evaluate neologism extraction tools.

2. Current Datasets for Evaluating Neologism Extraction

When selecting a dataset for evaluation, it's crucial to consider the specific goals of your neologism extraction tool and choose resources that align with your target language and domain. The availability of well-structured datasets is essential for the evaluation and advancement of techniques designed to extract neologisms [12]. These datasets provide benchmarks for assessing the effectiveness of different methodologies and offer valuable insights into the linguistic characteristics of newly coined terms.

One notable resource is the Adjective-Noun Neologism Dataset [7], which accompanies research on identifying adjective-noun neologisms using pretrained language models. This dataset contains positive examples of adjective-noun neologisms alongside negative examples, making it suitable for supervised learning and evaluation tasks.

Another important dataset is the New York Times Word Innovation Types (NYTWIT), which includes over 2,500 novel English words published in the New York Times between November 2017 and March 2019[13]. The entries in this dataset are manually annotated according to different lexical innovation processes, such as derivation, blending, and compounding. This resource is valuable for tracking linguistic innovation in media discourse and evaluating automated extraction systems.

Additionally, the NEO-BENCH benchmark [14] provides a comprehensive evaluation framework for assessing how well NLP models handle neologisms across various language understanding tasks. The benchmark highlights the robustness and adaptability of systems when encountering unfamiliar lexical items.

These datasets collectively address different aspects of neologism extraction, from structural and morphological innovation to semantic novelty and media-driven trends. They offer diverse testing environments for fully automated and semi-automated approaches, fostering the development of more accurate and context-aware systems.

3. Challenges for Datasets for Neologism Extraction

The creation of a dataset for the automatic extraction of neologisms presents multiple challenges related to the dynamic nature of linguistic innovation, the variability of textual sources, and the complexity of evaluation. One of the primary difficulties lies in defining what constitutes a neologism. Given that new words emerge and evolve over time, establishing temporal boundaries is essential but remains problematic, as some words gain acceptance while others disappear. Additionally, neologisms are highly domain-dependent, with technical fields generating specialized vocabulary that may not be perceived as new outside their respective disciplines. Their formation mechanisms, including affixation, blending, borrowing, and semantic shifts, add further complexity to their identification.

The choice of data sources significantly impacts the quality of a neologism dataset. While informal digital texts such as social media and blogs are rich sources of emerging words, they are also noisy, featuring spelling errors and non-standard language. On the other hand, curated sources like news articles or academic papers may offer greater linguistic stability but risk omitting the more ephemeral or subcultural neologisms. Ensuring a balanced representation across multiple text types is necessary but difficult to achieve. Ethical concerns also arise, particularly when mining from online communities where privacy regulations must be respected.

Annotation represents another major challenge. Human annotators must determine whether a term is genuinely new, rare, or simply a re-emergence of an older word. This process requires external validation, such as dictionary cross-referencing or frequency-based corpus comparison. Disagreements among annotators introduce variability in the dataset, reducing its reliability. Moreover, given that neologisms evolve, a static dataset may fail to capture their long-term usage trends. Therefore, an effective resource should support longitudinal tracking - repeated observation of the same variables -, enabling researchers to study word stabilization, meaning shifts, and eventual obsolescence.

4. Challenges for the Evaluation of Neologism Extraction Tools

Beyond data collection and annotation, evaluation poses additional obstacles. Unlike traditional NLP tasks, neologism extraction lacks standardized benchmarks, making performance assessment difficult.

Evaluating automatic tools for neologism extraction presents several challenges that must be addressed to improve their effectiveness and adaptability. One of the primary difficulties lies in defining the criteria for what constitutes a neologism across different domains. As new terms may emerge from slang, technical jargon, or creative word formations, establishing a universal benchmark for evaluation remains elusive.

Another significant challenge is the dynamic nature of language evolution, particularly on social media platforms. Rapid linguistic shifts, cultural memes, and ephemeral terms complicate the process of maintaining up-to-date evaluation datasets. Tools designed for neologism detection must therefore be adaptable and capable of processing large volumes of informal text while distinguishing between fleeting trends and enduring linguistic innovations.

Handling multilingual data adds an additional layer of complexity. Many neologisms emerge in one language and later diffuse into others, often undergoing transformations in spelling, morphology, or meaning. Evaluation frameworks must account for these cross-linguistic influences to assess the robustness of extraction tools in diverse linguistic environments.

Furthermore, distinguishing genuine neologisms from typographical errors, spelling variations, and non-standard word forms remains an issue. Automated systems require sophisticated mechanisms for contextual analysis to accurately filter out noise and identify meaningful linguistic innovations.

Semantic evaluation poses another challenge. Some neologisms involve new meanings for existing words rather than entirely novel forms. Automatic tools must therefore go beyond surface-level text analysis and incorporate semantic modeling techniques to capture these subtler shifts in usage.

Lastly, the human-in-the-loop approach remains critical for the evaluation process. While fully automated systems are efficient, expert validation is often necessary to ensure the linguistic validity and relevance of detected neologisms. Developing user-friendly interfaces and hybrid evaluation models that seamlessly integrate human expertise with machine efficiency is essential.

5. Conclusions

The continued creation and curation of high-quality datasets remain crucial for advancing research in the area of neologisms extraction. Future datasets should aim to capture neologisms from emerging domains, including social media and scientific literature, while incorporating multilingual perspectives to better understand global linguistic trends.

While there are limited datasets specifically dedicated to neologism extraction, several related resources can be utilized for this purpose. For example, tools like the NeoCrawler have been developed for semi-automatic neologism identification [1]. While not a dataset per se, it represents a methodological approach to neologism extraction. Sketch Engine also offers a feature called Trends, which is a diachronic analysis tool designed to study changes in word usage over time.[1] The NOW corpus (News on the Web) is another resource which was created from web-based newspapers and magazines from 2010 to the present time.[2] Google Trends[3] is also an alternative way of looking at how users search on the web rather than studying the content of the pages. In all these cases, a methodology for the evaluation of neologisms extraction is still to be studied.

Addressing these challenges will open the possibility for more accurate, scalable, and context-aware systems for neologism extraction. Future research should prioritize adaptive evaluation methodologies, cross-linguistic analysis, and enhanced semantic modeling to advance the state of the art in this domain. In particular, the representation of neologisms in Linked Open Data (LOD) is crucial for ensuring their integration into digital knowledge systems, enhancing both interoperability and accessibility across languages and domains [15].

Another possible line of research may involve interdisciplinary collaboration between digital humanities and the study of neologisms is pivotal in understanding and analyzing the evolution of language in the digital era. By integrating computational tools with linguistic research, scholars can effectively track, analyze, and interpret the emergence and usage of new words. For example, in [16] researchers employed computational methods to identify new elements of the hybrid language Surzhyk. In another work, [17], authors tackled the inaccuracies introduced by Optical Character Recognition (OCR) software when digitizing historical newspapers. This issue is crucial for accurately studying the evolution of language and the emergence of neologisms over time. The fusion of digital humanities and linguistic studies may offer a robust framework for exploring neologisms. Computational tools enable researchers to process vast textual datasets, identify new linguistic patterns, and understand the socio-cultural factors influencing language change. This interdisciplinary approach not only enriches our comprehension of language evolution but also enhances the methodologies employed in linguistic research.

Acknowledgments

This work is partially supported by the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074, and it is part of the initiatives of the Center for Studies in Computational Terminology (CENTRICO) of the University of Padua and in the research directions of the Italian Common Language Resources and Technology Infrastructure CLARIN-IT. This work is also partially supported by the "National Biodiversity Future Center - NBFC" project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of 16 December 2021, rectified by Decree n. 3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU. Project code CN_00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP F87G22000290001.

Declaration on Generative AI

During the preparation of this work, the author used Chat-GPT-4 in order to: Grammar and spelling check. After using these tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

[1] https://www.sketchengine.eu/english-trends-corpus/
[2] https://www.english-corpora.org/now/
[3] https://trends.google.com/home

References

[1] D. Kerremans, J. Prokić, Mining the Web for New Words: Semi-Automatic Neologism Identification with the NeoCrawler, Anglia 136 (2018) 239–268. URL: https://www.degruyter.com/document/doi/10.1515/ang-2018-0032/html?utm_source=chatgpt.com. doi:10.1515/ang-2018-0032, publisher: De Gruyter.

[2] J. Zhu, D. Jurgens, The structure of online social networks modulates the rate of lexical change, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2201–2218. URL: https://aclanthology.org/2021.naacl-main.178/. doi:10.18653/v1/2021.naacl-main.178.

[3] I. Falk, D. Bernhard, C. Gérard, From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers, in: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 4337–4344. URL: https://aclanthology.org/L14-1260/.

[4] B. Babych, Unsupervised Induction of Ukrainian Morphological Paradigms for the New Lexicon: Extending Coverage for Named Entities and Neologisms using Inflection Tables and Unannotated Corpora, in: T. Erjavec, M. Marcińczuk, P. Nakov, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger, R. Yangarber (Eds.), Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1–11. URL: https://aclanthology.org/W19-3701/. doi:10.18653/v1/W19-3701.

[5] A. Webson, Z. Chen, C. Eickhoff, E. Pavlick, Are "Undocumented Workers" the Same as "Illegal Aliens"? Disentangling Denotation and Connotation in Vector Spaces, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4090–4105. URL: https://aclanthology.org/2020.emnlp-main.335/. doi:10.18653/v1/2020.emnlp-main.335.

[6] M. Ryskina, E. Rabinovich, T. Berg-Kirkpatrick, D. Mortensen, Y. Tsvetkov, Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods, in: A. Ettinger, G. Jarosz, J. Pater (Eds.), Proceedings of the Society for Computation in Linguistics 2020, Association for Computational Linguistics, New York, New York, 2020, pp. 367–376. URL: https://aclanthology.org/2020.scil-1.43/.

[7] J. P. McCrae, Identification of Adjective-Noun Neologisms using Pretrained Language Models, in: A. Savary, C. P. Escartín, F. Bond, J. Mitrović, V. B. Mititelu (Eds.), Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), Association for Computational Linguistics, Florence, Italy, 2019, pp. 135–141. URL: https://aclanthology.org/W19-5116/. doi:10.18653/v1/W19-5116.

[8] M. Mizrahi, S. Yardeni Seelig, D. Shahaf, Coming to Terms: Automatic Formation of Neologisms in Hebrew, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4918–4929. URL: https://aclanthology.org/2020.findings-emnlp.442/. doi:10.18653/v1/2020.findings-emnlp.442.

[9] Y. Li, J. Cheng, C. Huang, Z. Chen, W. Niu, NEDetector: Automatically extracting cybersecurity neologisms from hacker forums, Journal of Information Security and Applications 58 (2021) 102784. URL: https://www.sciencedirect.com/science/article/pii/S2214212621000302. doi:10.1016/j.jisa.2021.102784.

[10] P. Lerner, F. Yvon, Towards the Machine Translation of Scientific Neologisms, Technical Report Rapport D2-3.1, ISIR, Université Pierre et Marie Curie UMR CNRS 7222, 2025. URL: https://hal.science/hal-04852293.

[11] L. Camacho, A primer on getting neologisms from foreign languages to under-resourced languages, 2023. URL: http://arxiv.org/abs/2304.10495. doi:10.48550/arXiv.2304.10495, arXiv:2304.10495 [cs].

[12] T.-J. Liu, S.-K. Hsieh, L. Prevot, Observing Features of PTT Neologisms: A Corpus-driven Study with N-gram Model, in: H.-D. Yang, W.-L. Hsu, C.-P. Chen (Eds.), Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013), The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Kaohsiung, Taiwan, 2013, pp. 250–259. URL: https://aclanthology.org/O13-1025/.

[13] Y. Pinter, C. L. Jacobs, M. Bittker, NYTWIT: A Dataset of Novel Words in the New York Times, 2020. URL: http://arxiv.org/abs/2003.03444. doi:10.48550/arXiv.2003.03444, arXiv:2003.03444 [cs].

[14] J. Zheng, A. Ritter, W. Xu, NEO-BENCH: Evaluating Robustness of Large Language Models with Neologisms, 2024. URL: http://arxiv.org/abs/2402.12261. doi:10.48550/arXiv.2402.12261, arXiv:2402.12261 [cs].

[15] J. P. McCrae, I. Wood, A. Hicks, The Colloquial WordNet: Extending Princeton WordNet with Neologisms, in: J. Gracia, F. Bond, J. P. McCrae, P. Buitelaar, C. Chiarcos, S. Hellmann (Eds.), Language, Data, and Knowledge, Springer International Publishing, Cham, 2017, pp. 194–202. doi:10.1007/978-3-319-59888-8_17.

[16] N. Sira, G. M. Di Nunzio, V. Nosilia, Towards an Automatic Recognition of Mixed Languages: The Case of Ukrainian-Russian Hybrid Language Surzhyk, Umanistica Digitale (2020) 97–116. URL: https://umanisticadigitale.unibo.it/article/view/10740. doi:10.6092/issn.2532-8816/10740, number: 9.

[17] D. Del Fante, G. M. Di Nunzio, Correzione dell'OCR per Corpus-assisted Discourse Studies: un caso di studio su vecchi quotidiani, Umanistica Digitale (2021) 99–124. URL: https://umanisticadigitale.unibo.it/article/view/13689. doi:10.6092/issn.2532-8816/13689, number: 11.