Adapting TBXTools to automatic terminology extraction with BERT

Gonzalo López-Sánchez[1,*], Patricia Morales-Hurtado[1], Mercè Vázquez[1], Albert Morales-Moreno[1] and Silvia Rodríguez Vázquez[1]

[1]Universitat Oberta de Catalunya (UOC), Rambla del Poblenou, 156, 08018, Barcelona, Catalonia, Spain

Abstract

Automatic terminology extraction is a challenging task for term compilation in specialized domains, as the most implemented methods currently yield lower precision and recall results compared to manual terminology selection from a corpus. This limitation is particularly relevant when identifying terms in domain-specific corpora due to contextual complexity and language adaptation. To explore new efficient methods for terminology extraction from specialized corpora, we have adapted TBXTools, a free automatic term extraction tool, by integrating a BERT-based approach. This method models automatic terminology extraction as a token-level sequence labeling task using BIO tagging, and trains the system on compiled reference term lists. In this paper, we present the approach used for the DETECH 2026 shared task on monolingual term extraction. We report evaluation results for English across two corpora: Mental Health and Parkinson's Disease. Using the BERT model as a filtering mechanism applied to terminology extraction, we achieved a significant improvement in both precision and recall.

Keywords

Automatic Terminology Extraction, BERT, computational terminology, health domain

1. Introduction

Automatic terminology extraction (ATE) is a relevant natural language processing (NLP) task involving terminology that is used to identify domain-relevant terms applying computational methods. However, ATE is a challenging task for term compilation in specialized domains, as the most implemented methods currently yield lower precision and recall results compared to manual terminology selection from a corpus. This limitation is particularly relevant when identifying terms in domain-specific corpora due to contextual complexity and language adaptation. To explore new ATE methods for efficient term extraction from specialized corpora, the GRIAL-ATE team participated in Task A of the DETECH 2026 Definition and Term Extraction Challenge [1] using TBXTools [2], a free automatic term extraction tool that facilitates the identification of multiword terms (MWT) in specialized corpora. This task required identifying relevant single-word and multi-word terms from English texts concerning the gut–brain interplay, which allowed us to evaluate the efficiency of our ATE model. The DETECH 2026 challenge is dedicated to advancing research on explainable data-driven medical terminology at the intersection of terminology, NLP, and biomedical text analysis.

To perform terminology compilation in the domain of the gut–brain interplay —a field at the intersection of gastroenterology, neuroscience, and genetics— we adapted TBXTools by integrating a BERT-based method. By introducing BERT as a filtering mechanism, we observed a significant improvement in precision and recall.

International Workshop on Definition and Term Extraction Challenge (DETECH) 2026. June 24, 2026, Zadar, Croatia
*Corresponding author.
glopezsanch@uoc.edu (G. López-Sánchez); pmoraleshu@uoc.edu (P. Morales-Hurtado); mvazquezga@uoc.edu (M. Vázquez); amoralesmore@uoc.edu (A. Morales-Moreno); srodriguezvaz@uoc.edu (S. Rodríguez Vázquez)
0009-0006-1486-838X (G. López-Sánchez); 0009-0009-2527-516X (P. Morales-Hurtado); 0000-0002-7983-4029 (M. Vázquez); 0000-0003-3068-0767 (A. Morales-Moreno); 0000-0002-9421-8566 (S. Rodríguez Vázquez)

2. Methodology

We participated in Task A of the DETECH 2026 challenge on monolingual term extraction, aiming to provide accurate term selection in English across two domain-specific corpora (Mental Health and Parkinson's Disease). To do so, we used the DETECH dataset. The training data consists of abstracts and corresponding lists of extracted terms, while the evaluation data consists only of abstracts from which the terms must be extracted. For this task, we used TBXTools, a term extraction tool that uses statistical and linguistic methods, and expanded it by implementing BERT-based models. These models have been used extensively in NLP tasks and, more recently, in automatic terminology extraction as well [3, 4]. Building on the workflow simplifications already provided with the existing methods, our tool was extended to support transformer-based models through the Hugging Face Transformers library. This addition reduced complexity and streamlined the processes of data preprocessing, automatic data labeling, model fine-tuning, and term extraction for users to train and deploy such models.

2.1. Language resources

In order to generate the training labels to be used during our system's training phase, we compiled—besides the terms' list provided by the task organizers—a term reference list from the fields of medical science and health, with a special focus on Mental Health and Parkinson's Disease. All terms were obtained from reliable sources, including lists from renowned institutions, official databases, and published dictionaries. The description section below describes each source; then, Table 1 provides further data, such as source ID and name, authors, domain, number of terms, and URL, when available.

Description of the compiled reference term lists:

- 001. The Genomic reference terms have been downloaded from the *Terminology of health sciences* published by the Centre for Terminology TERMCAT [5].
- 002. The Health reference terms have been selected with the highest reliability from the European IATE database.
- 003. The Medical terms were manually selected from a corpus we compiled based on the information published by the European Medicines Evaluation Agency.
- 004. We downloaded mental health terminology from [6]. The database is maintained and updated by the Translation Centre for the Bodies of the European Union, the official translation agency of the European Union. Its main goal is to provide translation and related language services to other decentralized EU agencies. It may also assist EU institutions and bodies with their own translation services, particularly during workload peaks or for specific projects [7]. To obtain the data, we created an account and accessed the "Download IATE" section. We then submitted a request specifying CSV as output format and selected the domain Social Questions (28) > health (2841) > health policy > organization of health care > mental health. In total, we obtained 20 entries used by the EU as reliable terms.
- 005. We expanded mental health terminology using [8], published by the Public Mental Health Implementation Centre (PMHIC), which is part of the Royal College of Psychiatrists. As the source was provided in PDF format, we converted it into a machine-readable format (XLSX) for processing. We extracted 35 terms from this publication for use in our system.
- 006. We manually selected Parkinson-related reference terms from a corpus compiled using information published by the National Library of Medicine.
- 007-009. We incorporated terminology from three glossaries provided by TERMIUM, the terminology and linguistic data bank of the Government of Canada, widely used as a reliable source of standardized terminology across multiple domains and languages. First, we extracted 333 mental health terms from the Workplace Mental Health Glossary. We also included terminology from the Glossary of Health Services, which required additional preprocessing due to its PDF format and complex double-column layout, in order to convert it into a machine-readable format. Finally, we incorporated 406 terms from the Glossary on Climate Change and Public Health.

- 010. We incorporated terminology from the Minimal Standard Terminology (MST) developed by the World Endoscopy Organization (WEO). This resource focuses on standardized terminology for gastrointestinal endoscopic procedures and findings. The dataset is freely available for use, provided that the copyright conditions are respected. We extracted 770 terms for our compilation.
- 011. We included terminology from the Talking Glossary of Genomic and Genetic Terms by the US National Human Genome Research Institute. This resource provides accessible definitions of key concepts in genomics and genetics.
- 012. We incorporated terminology from the Neuroscience Book Glossary by the US National Library of Medicine. This glossary provides domain-specific terminology in neuroscience.
- 013. We included terminology from the Gastrointestinal Glossary of Terms by the American Society for Gastrointestinal Endoscopy. This resource provides specialized terminology related to gastrointestinal procedures and clinical practice.
- 014. We incorporated the DISEASES database, a large-scale resource containing disease–gene associations. This dataset provides extensive terminology in the biomedical domain and was used to enrich the coverage of our terminology compilation.

ID	Name of the source	Author	Domain	Number of terms	Access
001	Genomics	TERMCAT, Centre of Terminology	Genomics	1570	Part of [5]
002	Health	Translation Centre for the Bodies of the European Union	Health	3208	Downloadable at [6]
003	Medical terms EMEA	European Medicines Evaluation Agency	Medicine	201	
004	Mental Health	Translation Centre for the Bodies of the European Union	Mental Health	20	Downloadable at [6]
005	List of terminology. Supplement to Public Mental Health. Leadership Certification course	Public Mental Health Implementation Centre	Mental Health	35	https://www.rcpsych.ac.uk/docs/default-source/improving-care/pmhic/list-of-terminology---pmhl-course_final.pdf?sfvrsn=4fb819a5_5
006	Parkinson		Parkinson	407	
007	Workplace Mental Health Glossary	TERMIUM	Mental Health	333	https://www.btb.termiumplus.gc.ca/publications/sante-mentale-mental-health-eng.html
008	Glossary of Health Services	TERMIUM	Health		https://publications.gc.ca/site/eng/9.674489/marcXml.html
009	Glossary on Climate Change and Public Health	TERMIUM	Health	406	
010	Minimal Standard Terminology (MST)	World Endoscopy Organization (WEO)	Gastroenterology	770	https://www.worldendo.org/resources/minimal-standard-terminology
011	Talking Glossary of Genomic and Genetic Terms	US National Human Genome Research Institute	Genomics	227	https://www.genome.gov/genetics-glossary
012	Neuroscience Book - Glossary	US National Library of Medicine	Neuroscience	715	https://www.ncbi.nlm.nih.gov/books/NBK10981/
013	Gastrointestinal Glossary of Terms	American Society for Gastrointestinal Endoscopy	Gastroenterology	224	https://www.asge.org/home/about-asge/gastrointestinal-glossary-of-terms
014	DISEASES	Sune Frankild, Alexander Junge, Albert Pallejà, Dhouha Grissa, Kalliopi Tsafou, and Lars Juhl Jensen	Medicine	9829	https://diseases.jensenlab.org/Downloads

Table 1
Summary of reference term lists

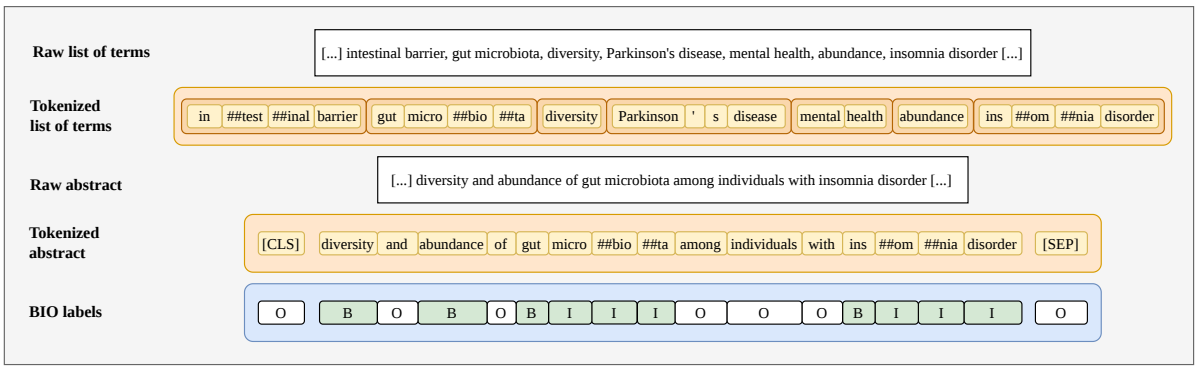

Figure 1: Tokenization and BIO labeling of the data.

2.2. Model

To identify biomedical terminology—both single-word and multi-word terms—from English texts related to the gut–brain interplay (specifically Mental Health and Parkinson's disease), we modelled terminology extraction as a token-level sequence labeling task using BIO tagging. BIO labels are primarily used in named-entity recognition (NER) tasks, but have also been applied to terminology extraction to some extent [4].

We fine-tuned BioBERT [9], a domain-specific variant of BERT pre-trained on biomedical corpora, using the Hugging Face Transformers library. Specifically, we used the `BertForTokenClassification` architecture, which adds a classification layer on top of the encoder to predict BIO labels for each token.

To obtain training labels, we automatically annotated the data given by the task organizers, i.e., the abstracts and the related terms. We tokenized and matched terms, alongside our own list of terms (see Table 1), against the tokenized abstracts and assigned BIO tags accordingly, i.e., B for 'beginning', I for 'inside' and O for 'outside'. In other words, our algorithm simply iterates through the tokenized terms one by one in descending order and, using a window of same length as each tokenized term, scans the tokenized abstract. When a match is found, it assigns the corresponding tags.

Since BERT tokenizers may split words into subword units, BIO labels were aligned to subword tokens. The first element in a term was always tagged as B, while the rest of the elements, whether subwords of the first one or different elements, were tagged as I, as shown in Figure 1, and non-terms were tagged as O. This tagging strategy may be considered contentious. For instance, it assigns the B label only to the first subword token and I to the rest, so it does not propagate labels consistently across all subwords. Nevertheless, the model maintained a good performance, and therefore the labeling system was preserved. Future research could further investigate the impact of different labeling strategies on model performance.

Likewise, BERT tokenizers automatically prepend a `[CLS]` token at the beginning of the sequence and append a `[SEP]` token at the end. These tokens are essential as the model relies on them for pre-training and specific downstream tasks, such as sentiment classification. In our study, these tokens were present at the beginning and end of each abstract but were disregarded during the prediction stage. Furthermore, it is important to acknowledge that any process involving the automatic generation of training labels may introduce noise. Notably, that we did not preprocess the terms provided by the task organizers associated with each abstract or those from our compiled term list, beyond the removal of duplicates. Consequently, any inconsistencies inherent in the source data were inevitably propagated into the training labels.

Dataset	System	Micro Precision	Micro Recall	Micro-F1	Type Precision	Type Recall	Type-F1
Mental Health	Q1	0.74	0.25	0.39	0.78	0.35	0.48
	Median	0.80	0.42	0.45	0.80	0.55	0.50
	Q3	**0.82**	0.64	0.56	**0.82**	0.58	0.67
	GRIAL-ATE	0.77	**0.66**	**0.71**	0.79	**0.62**	**0.69**
Parkinson	Q1	0.53	0.21	0.31	0.56	0.27	0.37
	Median	0.63	0.43	0.46	0.63	0.53	0.47
	Q3	**0.74**	0.61	0.57	**0.80**	0.55	0.65
	GRIAL-ATE	**0.74**	**0.62**	**0.68**	0.78	**0.57**	**0.66**

Table 2
Evaluation results comparing our approach (GRIAL-ATE) to the median, Q1, and Q3 system performance on the Mental Health and Parkinson datasets.

The model was trained using the default optimization setup of the Hugging Face Trainer API, with a learning rate of 5×10^{-5}, a batch size of 16, and a weight decay of 0.03 for 6 epochs. Hyperparameters were chosen based on the highest F1-score on a validation split of the training data.

The fine-tuned model was then applied to the evaluation data, predicting a BIO label for each token in the input sequence. Tokens assigned B or I labels were considered part of a term, while tokens labelled as O were treated as non-terms. Consecutive sequences of B and I labels were then grouped to form term spans. These spans of tokens were subsequently reconstructed using a custom regex function that merged subword units and accounted for dataset-specific characteristics, resulting in a list of extracted terms.

Finally, the aforementioned list underwent post-processing to remove stopwords and punctuation marks, given that they sometimes appeared individually in the predictions. This issue likely stems from their occurrence within correct entities, which introduces ambiguity during classification. For instance, the period in 'E. coli' or 'a' in tokenized 'aberrations' (`a ##ber ##rations`). Furthermore, other errors may be attributable to patterns inherent in the training abstracts, as these were not pre-processed prior to tokenization. Consequently, any noise present in the raw text may have propagated to the model, resulting in certain incorrect predictions.

3. Results and evaluation

The extracted terms were evaluated by the task organizers in both datasets using Micro-F1, which measures how consistently individual term instances are detected, and Type-F1, which measures performance over unique term types. Table 2 reports the evaluation results of our best-performing run, alongside the median scores of all task participants. Our model achieves strong performance across both Type-F1 and Micro-F1, exceeding the median F1 results in both datasets. Notably, the model obtains significantly higher precision and recall in the Parkinson dataset than the median. In contrast, in the Mental Health dataset, precision remains closer to the median, although recall still surpasses it.

4. Conclusions and future work

In Task A of the DETECH 2026 challenge on monolingual term extraction, we adapted TBXTools, a free automatic term extraction tool, by integrating a BERT-based approach to maximize terminology extraction from the Mental Health and Parkinson's Disease corpora of the DETECH dataset. This integration models terminology extraction as a token-level sequence labeling task using BIO tagging and trains the system on compiled reference term lists.

The results obtained in the DETECH 2026 challenge compared to the median system performance confirm that the BERT-based adaptation of TBXTools is a suitable approach for terminology extraction

in specialized domains. In future work, we plan to further improve the results by refining the applied methodology.

Acknowledgments

This work was supported by project TamTAS PCI2025-167063-2, funded by MICIU/AEI/10.13039/501100011033 and European Union in the Chist-era call 2025 Science in your own language.
This work was supported by the Industrial Doctorates Plan from the Department of Research and Universities of the Government of Catalonia (reference number: 2025 DI 00112).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] F. Vezzani, G. M. Di Nunzio, V. Bonato, G. Silvello, Overview of the international workshop on definition and term extraction challenge (detech) 2026, in: Proceedings of the International Workshop on Definition and Term Extraction Challenge (DETECH) 2026, CEUR.org, Zadar, Croatia, 2026.
[2] A. Oliver, M. Vàzquez, TBXTools: A free, fast and flexible tool for automatic terminology extraction, in: R. Mitkov, G. Angelova, K. Bontcheva (Eds.), Proceedings of the International Conference Recent Advances in Natural Language Processing, INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, 2015, pp. 473–479. URL: https://aclanthology.org/R15-1062/.
[3] K. Xu, Y. Feng, Q. Li, Z. Dong, J. Wei, Survey on terminology extraction from texts 12 (2025) 29. doi:10.1186/s40537-025-01077-x.
[4] H. T. H. Tran, M. Martinc, J. Caporusso, J. Delaunay, A. Doucet, S. Pollak, Recent advances in automatic term extraction: A comprehensive survey 58 (2026) 1–35. doi:10.1145/3787584.
[5] TERMCAT, Centre de Terminologia, Terminologia de ciències de la salut, https://www.termcat.cat/en/diccionaris-en-linia/198, 2015–2026.
[6] Translation Centre for the Bodies of the European Union, IATE, 2018. URL: https://iate.europa.eu/home.
[7] Translation Centre For the Bodies of the EU, Homepage, 2026. URL: https://cdt.europa.eu/en.
[8] Public Mental Health Implementation Centre, List of terminology: Supplement to Public Mental Health Leadership Certification course, Public Mental Health Implementation Centre, London, 2024.
[9] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining 36 (2020) 1234–1240. doi:10.1093/bioinformatics/btz682.