OATE at ATE-IT: A Hybrid Approach for Automatic Terminology Extraction in Italian

Omar Arab[1,*], Giorgio Maria Di Nunzio[1]

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

Abstract

This paper describes a hybrid system developed for the ATE-IT shared task at EVALITA 2026. The system combines supervised sequence labeling based on a BERT+CRF architecture with semantic post-processing driven by term embeddings. Graph-based ranking methods were also explored but performed poorly since was not suited because of the nature of the dataset. Experimental analysis shows that embedding-based semantic similarity, enhanced with large language model (LLM)–generated domain knowledge, represents the dominant signal for terminology filtering in this setting.

Keywords

Automatic Terminology Extraction, EVALITA, Italian, BERT, Conditional Random Fields, Semantic Embeddings

1. Introduction

Automatic Terminology Extraction (ATE) is a specialized Natural Language Processing task aimed at identifying and extracting domain-specific terms from textual corpora [1, 2]. Despite its apparent simplicity, ATE poses several challenges, including the detection of multi-word expressions, ambiguous term boundaries, and the distinction between domain-specific terms and generic noun phrases.

In this context, recent ATE approaches increasingly frame term identification as a span-detection problem, borrowing sequence-labelling architectures from Named Entity Recognition while retaining classical insights about unithood (whether a candidate forms a stable multi-word unit) and termhood (whether it is domain-specific) [3, 4]. Hybrid pipelines typically combine contextual encoders with linguistically informed post-processing to enforce task constraints (e.g., de-duplication, normalization, and nested-term removal) and to better control boundary errors on multi-word expressions, which remain a primary source of false positives and fragmented predictions in morphologically rich languages such as Italian.

The ATE-IT [5] shared task at EVALITA 2026 [6] focuses on Italian terminology extraction, with particular attention to technical domains such as waste management and environmental regulation. Compared to English, Italian presents additional challenges due to its rich morphology and complex syntactic structures. These characteristics require models capable of capturing both local structural constraints and global semantic context.

2. System Architecture and Methodology

The proposed system follows a hybrid pipeline composed of three main stages: (i) high-recall candidate extraction via supervised sequence labeling, (ii) semantic enrichment of candidate terms, and (iii) embedding-based ranking and filtering. While graph-based ranking techniques were initially investigated, they were ultimately found to be ineffective for the given dataset and therefore play only a marginal role in the final system.

ATE EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26–27, Bari, Italy

*Corresponding author.

omar.arab@studenti.unipd.it (O. Arab); giorgiomaria.dinunzio@unipd.it (G. M. Di Nunzio)

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

2.1. Preprocessing and Tokenization

The corpus undergoes a structured preprocessing pipeline before model training. First, all documents are normalized to UTF-8 encoding to avoid inconsistencies caused by special characters. Non-textual artifacts (e.g., HTML remnants or malformed symbols) are removed. Excess whitespace is normalized, and line breaks are standardized. No lowercasing is applied since we rely on the cased version of the Italian BERT model to preserve morphological and orthographic distinctions relevant for domain terminology.

Tokenization is performed using the WordPiece algorithm implemented in the `dbmdz/bert-base-italian-cased` model. WordPiece decomposes rare or morphologically complex words into subword units. For example, a compound technical term may be split into multiple sub-tokens prefixed with `##`.

To ensure correct supervision during training, we adopt an explicit alignment strategy between original tokens and subword units. Labels are assigned only to the first subword of each original token, while subsequent subwords are masked during loss computation. Specifically:

- If a token is labeled as `B-TERM`, its first subword receives `B-TERM` and the remaining subwords are assigned a special ignore index.
- If labeled as `I-TERM`, the same rule applies.
- Non-term tokens are labeled as `O`.

Character-level offset mappings provided by the tokenizer are used to reconstruct spans precisely after inference. This guarantees that extracted candidate terms preserve their exact surface form in the original documents.

2.2. Neural Sequence Labeling with BERT and CRF

Terminology extraction is formulated as a supervised sequence labeling task using the BIO tagging scheme. The BIO tagging scheme encodes span boundaries by assigning a label to each token in the sequence. In this framework, `B-TERM` marks the beginning of a terminology span, `I-TERM` marks tokens that continue a previously started term, and `O` marks tokens that do not belong to any terminology expression. This representation allows the model to capture multi-word domain expressions while explicitly preserving boundary information. By distinguishing between span-initial and span-internal tokens, the BIO formalism provides structural constraints that are particularly important for accurately modeling complex and morphologically rich terminology.
A pre-trained Italian BERT model (`dbmdz/bert-base-italian-cased`) is used to generate contextualized representations for each token.

To enforce global consistency in label predictions, a Conditional Random Field (CRF) layer is applied on top of BERT emissions [7]. While a token-level classifier may produce locally optimal predictions, it cannot model dependencies between adjacent labels. The CRF explicitly learns transition constraints, preventing invalid sequences such as an `I-TERM` tag following an `O` tag.

Empirically, the CRF layer significantly improved span-level consistency, particularly for multi-word terms.

2.3. Training Strategy

The model is trained using the AdamW optimizer [8] with discriminative learning rates. A lower learning rate is applied to the BERT encoder to preserve pre-trained linguistic knowledge and to avoid catastrophic forgetting, while higher rates are used for the CRF layer to speed up learning. This strategy allows effective adaptation to the terminology extraction task. In our experiments we have kept 5 epochs since this number has shown better performances in terms of loss.

2.4. Candidate Reconstruction

Predicted BIO sequences are converted into candidate terms by extracting maximal contiguous spans of `B-TERM` and `I-TERM` labels. To avoid reconstruction artifacts introduced by subword tokenization, term spans are mapped back to the original text using character-level offsets. This guarantees that extracted terms preserve the exact surface form found in the source documents.

2.5. Graph-Based Ranking: Empirical Limitations

We experimented with graph-based ranking methods, including PageRank-style centrality and TextRank, to estimate candidate-term salience from co-occurrence and (where available) syntactic relations [9, 10].

Each candidate term is represented as a node, and edges connect terms that are observed as related within a fixed context window and/or within the same dependency neighbourhood. Edge weights correspond to the strength of association, operationalised here as (weighted) co-occurrence frequency. Once the graph is constructed, we compute node centrality scores and use them as an additional signal in post-processing (e.g., re-ranking or filtering low-salience candidates).

Empirically, however, this signal had limited impact on the ATE-IT dataset: documents are relatively short and specialised terms exhibit limited repetition, leading to sparse, weakly connected graphs. As a consequence, centrality tended to privilege generic, high-frequency expressions that act as hubs across many contexts, rather than genuinely domain-specific terminology.

This behaviour is consistent with known failure modes of frequency-driven graph centrality when domain terms have low recurrence; addressing it would likely require stronger association measures, domain-adaptive priors, or document level aggregation,which we leave for future work.

2.6. Semantic Augmentation via Large Language Models

To enhance domain coverage, we leveraged a large language model accessed via the Groq API. The model employed was `llama-3.1-8b-instant`, accessed in November 2025. The model was prompted in zero-shot mode to extract domain-specific terminology from Italian environmental legislation, in particular *D.Lgs. 152/2006*. The prompting instructions were written in Italian to maximize domain and linguistic alignment.

The model was instructed to return only technical terms relevant to waste management and environmental regulation, one per line, without additional explanations. The extracted terminology was aggregated across document chunks and subsequently deduplicated and filtered to remove noise.

While LLM-generated candidates cannot be considered fully reliable due to possible variability across runs, they provide broader conceptual coverage and introduce domain signals that may not be sufficiently represented in the supervised training data. Rather than being used directly as predictions, these terms serve as a semantic reference set in the embedding validation phase.

2.7. Embedding-Based Semantic Ranking

The dominant ranking signal in the final system is based on semantic similarity in an embedding space. Each extracted candidate term is represented as a dense embedding vector, together with a set of domain-relevant reference terms obtained from model extractions and LLM-generated terminology.

For each candidate term, cosine similarity is computed against all reference terms in the semantic space. A candidate is retained if at least one reference term exceeds a predefined similarity threshold τ. This nearest-neighbor semantic validation strategy allows the system to identify candidates that are semantically aligned with the domain, even when they do not exhibit strong frequency or co-occurrence signals.

In practice, this approach proved effective at filtering generic expressions while preserving domain-specific terminology, particularly in sparse corpora where traditional graph-based or frequency-based signals are unreliable.

2.8. Final Ranking Strategy

The final ranking relies primarily on embedding-based semantic similarity, with extraction confidence from the BERT+CRF model used as a secondary signal. Graph-based centrality was evaluated but its effect on the final scoring was practically nil.

3. Results

The system was evaluated using the official ATE-IT evaluation script, reporting both micro-averaged and type-based Precision, Recall, and F1-score. Micro-level metrics measure performance over all term occurrences, while type-level evaluation considers unique term instances, providing complementary perspectives on extraction quality.

The OA-TE system (run 2) achieved a micro Precision of 0.581, a micro Recall of 0.522, and a micro F1-score of 0.550, outperforming the official baseline (micro F1 = 0.526). At the type level, OA-TE obtained a Precision of 0.569, a Recall of 0.492, and a type F1-score of 0.528, again showing a clear improvement over the baseline (type F1 = 0.469). The observed performance indicates a clear precision–recall trade-off introduced by the semantic filtering stage. While the embedding-based validation slightly reduces recall by discarding candidates that are weakly aligned with the target domain, it contributes to a modest but consistent gain in precision.

Although graph-based centrality was experimentally integrated into the scoring function, it did not influence the final evaluation metrics. The observed Precision, Recall, and F1-scores remained unchanged within rounding tolerance across tested configurations. Consequently, the final submitted system does not depend on PageRank-based refinement, and the reported results correspond to the embedding-driven filtering strategy.

4. Discussion

The experimental findings show how strongly Automatic Terminology Extraction methods depend on the characteristics of the dataset. In the ATE-IT task, documents are relatively short and many domain-specific terms appear only a few times. This has important consequences for the effectiveness of different ranking strategies. Methods that rely mainly on frequency or global connectivity signals naturally struggle when repetition is limited.

Graph-based ranking techniques such as PageRank were explored as a complementary signal to supervised extraction. The idea was that terms that frequently co-occur with other relevant terms would become more central in a co-occurrence graph and therefore more representative of the domain. We built weighted graphs based on sentence-level co-occurrence and also enriched them using syntactic proximity information obtained from dependency parsing. However, due to the sparsity of the dataset, the resulting graph was weakly connected and centrality scores were not very discriminative. In practice, PageRank tended to favor terms that were structurally well connected rather than truly domain-specific. When integrated into a hybrid scoring function together with embedding similarity, the graph signal did not produce consistent improvements in Precision, Recall, or F1-score. This suggests that in small and specialized corpora, graph connectivity alone may not be sufficient to model domain relevance.

The supervised extraction component based on BERT with a CRF decoder proved much more reliable. The CRF layer helps enforce valid transitions within the BIO tagging scheme and reduces inconsistent label sequences. This is especially important for Italian, where technical terms often consist of multi-word expressions with morphological agreement. By modeling dependencies between adjacent labels, the CRF improves boundary consistency and reduces span fragmentation. Nevertheless, the sequence labeling approach also has limitations. The BIO scheme cannot represent nested or overlapping terminology, and the model remains dependent on the coverage and quality of the annotated training data.

The embedding-based semantic validation stage was the most impactful part of the system. By representing candidate terms in a contextual embedding space and comparing them to a domain-specific reference set, the model filters candidates based on semantic consistency rather than frequency. This approach is particularly effective in sparse datasets, where repetition is not a reliable indicator of termhood. Embedding similarity allows the system to retain rare but semantically aligned terms while discarding generic noun phrases that are not truly domain-specific.

At the same time, this strategy introduces some sensitivity. The similarity threshold directly affects the precision–recall trade-off. A lower threshold increases recall but may allow more noise, while a higher threshold improves precision but risks filtering out valid terms. Although the threshold was tuned empirically, it remains a manually chosen hyperparameter and may not generalize perfectly to different domains.

Another important aspect concerns the use of LLM-derived terminology for semantic augmentation. While the prompting strategy was carefully designed to restrict outputs to relevant technical terms, large language models can show variability across runs and may introduce subtle bias. The inclusion of externally generated terms strengthens domain coverage, but it also introduces dependence on model behavior at a specific time and version. A more integrated solution, such as domain-adaptive fine-tuning instead of post-hoc filtering, could potentially provide more stable improvements.

More generally, the current pipeline treats extraction and semantic validation as two separate stages. While this modular design simplifies experimentation, it may limit the interaction between structural boundary detection and semantic alignment. Future work could explore joint optimization strategies or contrastive training objectives that directly encourage domain-consistent representations during fine-tuning.

Overall, the experiments suggest that for the ATE-IT dataset, semantic alignment in contextual embedding space is more effective than graph-based structural connectivity. In sparse and highly specialized corpora, embedding similarity provides a stronger and more stable signal for domain consistency. However, the system remains sensitive to threshold selection, training data coverage, and external semantic augmentation, indicating room for further methodological refinement.

Declaration on Generative AI

During the preparation of this work, generative AI tools were used to assist with language refinement. All generated content was reviewed and edited by the author, who takes full responsibility for the final manuscript.

References

[1] K. Kageura, B. Umino, A survey of automatic term extraction, Terminology 16 (2010) 259–289.

[2] G. M. Di Nunzio, S. Marchesin, G. Silvello, A systematic review of Automatic Term Extraction: What happened in 2022?, Digital Scholarship in the Humanities 38 (2023) i41–i47. doi:10.1093/llc/fqad030.

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

[4] K. Frantzi, S. Ananiadou, H. Mima, Automatic recognition of multi-word terms:. the C-value/NC-value method, International Journal on Digital Libraries 3 (2000) 115–130. doi:10.1007/s007999900023.

[5] N. Cirillo, G. M. Di Nunzio, F. Vezzani, Ate-it at evalita 2026: Overview of the automatic term extraction italian testbed task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[6] F. Cutugno, A. Miaschi, A. P. Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, Evalita 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.
[7] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, Proceedings of the 18th International Conference on Machine Learning (2001) 282–289.
[8] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations (ICLR), 2019.
[9] L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking : Bringing order to the web, in: The Web Conference, 1999. URL: https://api.semanticscholar.org/CorpusID:1508503.
[10] R. Mihalcea, P. Tarau, TextRank: Bringing Order into Text, in: D. Lin, D. Wu (Eds.), Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 404–411.
[11] X. Wan, J. Xiao, Singlerank: A graph-based keyword extraction method, Proceedings of the 23rd International Conference on Computational Linguistics (COLING) (2010) 465–473.
[12] T. B. Brown, B. Mann, N. Ryder, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020).