VVTE at ATE-IT: From Candidates to Terms: Hybrid Italian ATE with Dependency Heuristics, Gemini, and Random Forest Filtering

Valentine G. L. Vandervoort[1,*], Giorgio Maria Di Nunzio[1]

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

Abstract

Automatic Term Extraction (ATE) aims to identify domain-specific lexical units, including single- and multi-word expressions, that encode the conceptual structure of specialized fields. This paper presents our submission to Subtask A (Term Extraction) of the ATE-IT shared task [1] at EVALITA 2026 [2], the first large-scale evaluation campaign dedicated to Italian ATE, focused on institutional texts in the municipal waste-management domain. To cope with sparse terminology, rich Italian morphology, and complex administrative multi-word terms, we adopt a hybrid pipeline that prioritizes recall and then progressively filters candidates. We first generate a broad pool of candidate terms using dependency-based heuristics with spaCy and zero-shot term identification via Google Gemini. We then apply normalization and task-specific constraints and refine predictions with a Random Forest classifier trained on a custom feature set capturing morphological patterns, domain-keyword density, and semantic similarity signals. We report official results on the test set and provide an analysis of typical errors, highlighting the benefits and limitations of combining linguistic heuristics, generative models, and supervised learning for Italian institutional automatic term extraction.

Keywords

Automatic Term Extraction, ATE-IT, Linguistic features, Waste management

1. Introduction

Automatic Term Extraction (ATE) is a core task in Natural Language Processing (NLP) and computational linguistics, designed to identify domain-specific terms that represent key concepts within a specialized field of knowledge [3, 4, 5]. Unlike Named Entity Recognition (NER), which focuses on identifying unique referents such as specific individuals, organizations, or geographic locations, ATE aims to extract lexical units, both single-word and multi-word expressions, that constitute the conceptual framework of a domain. The outputs of ATE systems are vital for numerous downstream applications, including the construction of ontologies, enhancement of knowledge graphs, machine translation, and the domain adaptation of Large Language Models (LLMs).

This paper describes our participation in the ATE-IT [1] (Automatic Term Extraction – Italian Testbed) shared task, organized within the framework of EVALITA 2026 [2]. EVALITA is the main evaluation campaign for NLP and speech technologies for the Italian language [**?**]. The ATE-IT task represents an important step toward systematic evaluation for Italian ATE, with a focus on the institutional domain of waste management. This domain presents significant linguistic challenges, including a high density of derived terms (e.g., *"ecodizionario"*), technical abbreviations (e.g., *"TARI", "RAEE"*), and complex, multi-word administrative terms (e.g., *"raccolta porta a porta"*).

The ATE-IT campaign is structured into two distinct subtasks: Term Extraction (Subtask A) and Term Variants Clustering (Subtask B). Our contribution focuses exclusively on Subtask A, which requires the identification of domain-relevant terms, including nouns, verbs, and adjectives, from a specialized corpus.

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

*Corresponding author.

valentinegabrielleLucile.vandervoort@studenti.unipd.it (V. G. L. Vandervoort); giorgiomaria.dinunzio@unipd.it (G. M. Di Nunzio)

https://valouvalouvalou.github.io/ (V. G. L. Vandervoort)

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

To address the inherent difficulty of identifying sparse, domain-specific terms in Italian, we propose a hybrid pipeline that combines linguistic heuristics, generative artificial intelligence, and supervised machine learning. Our approach first maximizes recall by extracting a broad set of candidates using spaCy for dependency parsing [**?**] and Google Gemini for zero-shot term identification. This candidate list is subsequently refined through a rigorous filtering process and a Random Forest classifier [**?**]. The classifier is trained on a custom feature set including morphological patterns, domain-keyword density, and semantic similarity to distinguish valid domain terms from general language.

In the following sections, we provide a comprehensive breakdown of our system architecture, followed by an analysis of our performance on the official test set and a discussion of the implications of our hybrid approach.

2. Description of the System

The proposed system follows a hybrid architecture designed to achieve high recall in the initial phase, followed by a precision-oriented filtering stage. The pipeline integrates rule-based linguistic analysis, generative AI, and supervised machine learning.

2.1. Preprocessing and Candidate Extraction

The first stage of the pipeline aims to identify a broad set of potential terms (candidates) from the raw text without immediate domain-specific filtering. This ensures that rare or complex terms are not overlooked.

Linguistic Parsing. I utilize the `it_core_news_lg` model from *spaCy* to perform tokenization, part-of-speech (POS) tagging, and dependency parsing. To handle complex sentence structures, each sentence is segmented using punctuation anchors (e.g., semicolons, question marks, and commas) while preserving the integrity of segments that lack these separators.

Multi-source Extraction. To capture a diverse range of terminological units, we employ several concurrent extraction strategies:

- **Syntactic Subtrees:** We extract subtrees anchored in nouns, subjects, verbs, and adjectival phrases.
- **Noun Chunks:** We leverage *spaCy*'s native noun chunking to identify base nominal groups.
- **Technical Abbreviations:** A specialized module extracts technical acronyms, filtering out common Italian stopwords and purely numerical strings.
- **Linguistic Patterns:** We specifically target prepositional phrases, compound nouns, appositions, and conjunction chains to capture complex multi-word terms.

LLM-Augmented Extraction. In addition to rule-based methods, we incorporate Google Gemini to identify candidates. By processing each sentence individually, the model leverages its internal knowledge to suggest terms that might escape rigid syntactic patterns, providing a robust complement to the heuristic-based extraction.

Initial Refinement. All extracted candidates undergo a cleaning process where leading and trailing stopwords are removed. This ensures that the boundaries of the candidates are linguistically sound before entering the validation phase.

2.2. Validation and Feature Engineering

Once the initial set of candidates is gathered, we apply a series of filters and transformation steps to prepare the data for the classification model. This stage is designed to eliminate noise and quantify the characteristics of each term.

Structural Filtering. We first perform a preliminary validation of the candidate's form. Any term containing digits, special characters (with the exception of hyphens and commas), or patterns resembling

dates is automatically rejected. Furthermore, we discard single-token candidates that appear in the Italian stopword list to ensure that only potentially meaningful units remain.

Feature Extraction. For each remaining candidate, we compute a comprehensive set of features that capture its morphological, lexical, and contextual properties:

- **Morphological features:** We calculate the word count, character length, and the ratio of vowels to consonants. We also check for the presence of typical Italian suffixes and prefixes that often characterize technical terminology.
- **Domain-specific features:** Using a pre-compiled list of domain keywords, we determine the number and percentage of keywords within the term. We also verify whether the candidate or its components appeared in the training dataset.
- **Contextual and Semantic features:** We record the term's relative position in the sentence and its internal structure (e.g., whether it is a compound noun). Additionally, We calculate the ratio of internal stopwords and the semantic similarity of the candidate to validated terms from the training set.

Normalization. Before passing these features to the classifier, we normalize the terms to ensure consistency across the dataset. This step is vital for the model to generalize effectively across different morphological variants of the same concept.

2.3. Supervised Classification and Final Selection

To distinguish genuine waste management terminology from general language candidates, we implemented a supervised learning approach designed to prioritize domain relevance.

Model Training. We trained a Random Forest classifier using the provided training dataset. To do this, we constructed a balanced dataframe consisting of positive examples (validated terms from the corpus) and negative examples (general language words and irrelevant extractions). To address the inherent sparsity of technical terms, we configured the model with a specific weight ratio, assigning three times more importance to the positive class (`class_weight=1: 3`). This weighting strategy is crucial for ensuring the model remains sensitive to domain-specific terminology despite the prevalence of common vocabulary.

Decision Process. During the inference phase, we submit each candidate to the trained model, along with its vector of morphological and contextual features. To maximize precision and minimize "noise" from false positives, we apply a strict dual-condition filter. A candidate is only accepted as a valid term if:

1. The model predicts it as a domain-relevant term (Class 1).
2. The prediction probability exceeds a confidence threshold of 0.7.

Output Generation. Finally, we reconstruct the definitive dictionary of identified terms. We take particular care to ensure the output data structure matches the input format exactly, maintaining the required alignment for the EVALITA evaluation scripts. The final results are exported into a structured JSON file, formatted for official submission.

3. Results

In this section, we present the official results obtained on the ATE-IT test set after processing the data through my hybrid pipeline. The performance is evaluated using the official metrics provided by the EVALITA 2026 organizers, focusing on both Micro (instance-level) and Type (unique term-level) scores.

Quantitative Performance. Table 1 summarizes the comparison between the official ATE-IT baseline and our system's output.

Analysis of the Precision-Recall Trade-off. The most striking result is the significant leap in Type Precision, where our system achieved 0.707, substantially outperforming the baseline's 0.435. This

System	M-Prec	M-Rec	M-F1	T-Prec	T-Rec	T-F1
Baseline	0.497	0.559	0.526	0.435	0.508	0.469
My System	0.364	0.473	0.411	0.707	0.262	0.382

Table 1
Comparison of official baseline results and my system performance on the test set.

outcome reflects a deliberate architectural choice: we prioritized high-confidence extractions over broad coverage. By setting a strict probability threshold of 0.7 and utilizing multi-stage filtering, we ensured that the majority of the unique terms identified were indeed valid and highly relevant to the waste management domain.

However, this conservative strategy resulted in a significant drop in Type Recall (0.262). Upon internal review of the output, it appears that while the extracted terms are highly accurate, they consist mostly of single-word or short multi-word units (maximum 3 words). The pipeline struggled to capture longer, more complex administrative expressions, leading to a high number of false negatives. This suggests that while my Random Forest classifier is highly effective at validating core domain vocabulary, the initial candidate extraction or the subsequent filtering was perhaps too restrictive for complex compositional terms.

Micro-Level Performance Analysis. Regarding instance-level performance, our system achieved a Micro Precision of 0.364 and a Micro Recall of 0.473. While these figures remain below the official baseline, the discrepancy between the Micro and Type metrics offers a deeper insight into the pipeline's behavior. The gap between the high Type Precision (0.707) and the lower Micro Precision (0.364) suggests that while we successfully identify a correct unique concept once, the system lacks consistency in extracting every occurrence of that concept across the entire corpus. A term might be correctly validated in a favorable syntactic context but rejected elsewhere if the local features or the Gemini confidence score fall below the 0.7 threshold.

Furthermore, the fact that Micro Recall (0.473) is significantly higher than Type Recall (0.262) indicates that our system tends to reliably capture a core set of frequent, well-defined terms while missing the "long tail" of less frequent terminology. This confirms that the Random Forest classifier, coupled with the strict filtering rules, favored statistical safety. By prioritizing terms with strong morphological and keyword-based signatures the system ensured that the extracted output was qualitatively relevant to the waste management domain, even at the cost of exhaustive coverage and instance-level consistency.

4. Discussion

The evaluation results provide a nuanced view of the hybrid pipeline's capabilities. While the system achieved an exceptional leap in Type Precision compared to the baseline, the trade-off in recall and the difficulty in capturing complex expressions offer important insights into the nature of the task and the limitations of the chosen architecture.

4.1. Successes and Model Selection

During the development phase, we experimented with several classification architectures. The Random Forest Classifier consistently proved to be the most stable and effective model on the training dataset, leading to its selection for the final system. The inclusion of *class_weight* adjustments was particularly effective in forcing the model to recognize domain-specific vocabulary that might otherwise have been overshadowed by general language tokens. The primary success of this approach is its ability to identify high-confidence, unique terms within the waste management domain. Our final system achieved a Type Precision of 0.707, which indicates a very high rate of true positives among the unique terms extracted.

However, this result must be interpreted with caution. While it demonstrates a high level of accuracy in identifying valid terminology, the score is also a mathematical consequence of a very low Type Recall (0.262). Because the system was highly selective and extracted a relatively small number of

unique terms, the pool of candidates for which precision was calculated was limited. Nevertheless, the score remains a positive indicator of the system's ability to maintain a clean output, largely free from non-terminological noise.

4.2. Experiments with LLM Filtering

In one iteration of the pipeline, we integrated the Grok LLM via API to act as an initial filter immediately following the first extraction phase. The goal was to use the LLM to prune irrelevant candidates before they reached the feature engineering and classification stages. Interestingly, this additional layer of generative filtering resulted in negligible changes to the final evaluation scores. This suggests that the heuristic-based extraction and the subsequent Random Forest classifier were already capturing the most salient features of the data.

4.3. Analysis of Limitations and Short-term Bias

The main challenge encountered was a significant "short-term bias," with the output consisting mostly of single words or very short multi-word units. We attribute this to two factors:

- **Classifier Rigidity:** The Random Forest model appears to have been the primary bottleneck for longer expressions. It likely penalized complex strings whose statistical features, such as stopword ratios or character lengths, deviated from the more frequent, shorter terms found in the training data.
- **Structural Fragmentation:** The aggressive segmentation strategy used to handle complex sentences may have inadvertently severed the continuity of longer institutional phrases, preventing the model from ever seeing the full multi-word candidate.

4.4. Future Work

To bridge the gap between precision and recall, future work will shift focus toward finding a more balanced equilibrium. We intend to move away from a precision-only focus and instead prioritize improving the Recall score. This will involve relaxing the current strict filtering rules and lowering the probability threshold to capture a wider variety of terms that were previously overlooked. Additionally, we plan to refine the extraction phase to better preserve the integrity of complex noun phrases and experiment with transformer-based embeddings to help the system generalize better to rare, long-tail terminology that the current model failed to capture.

Declaration on Generative AI

During the preparation of this work, the author used OpenAI GPT-5.2 and Google Gemini in order to perform grammar and spelling checks, as well as to assist with text reformulation for clarity and flow. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] 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.

[2] 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.

[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] Y. Chun, M. Kim, D. Kim, C. Park, H. Lim, Enhancing Automatic Term Extraction with Large Language Models via Syntactic Retrieval, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 9916–9926. doi:10.18653/v1/2025.findings-acl.516.