TermNinjas at ATE-IT: Overview of the Term Extraction and Term Variants Clustering Task

Zunaira Hasnain[1], Fashad Ahmed Siddque[1]

[1]Department of Computer Science and Engineering (DISI), Artificial Intelligence Programme, University of Bologna, Italy

Abstract

This paper presents a hybrid neural symbolic system submitted to the ATE-IT Shared Task at EVALITA 2026, addressing both Subtask A (Automatic Term Extraction) and Subtask B (Term Variants Clustering) on Italian municipal waste management documents. The proposed approach combines supervised fine tuning of an Italian BERT model for sequence labeling with extensive domain aware post processing, followed by a multi stage clustering pipeline for grouping term variants.

For Subtask A, the system focuses on high precision extraction of single word and multi word domain terms through BIO tagging and constraint aware reconstruction. For Subtask B, a hybrid clustering strategy is adopted, integrating lemmatization and multilingual semantic embeddings, with optional large language model based verification. Experimental results on the official test set show competitive performance, with particularly strong precision and clear improvements over the official baseline in the clustering task. The entire system relies exclusively on open source tools and models, ensuring full reproducibility without external API dependencies.

Keywords

Automatic Term Extraction, Term Variants Clustering, Italian NLP, Waste Management, BERT, Hybrid Systems

1. Introduction

Automatic Term Extraction (ATE) is a fundamental task in Natural Language Processing (NLP) aimed at identifying domain specific terminology from specialized corpora. Extracted terms represent the conceptual backbone of a domain and are widely used in downstream applications such as ontology construction, knowledge base population, terminology management, domain specific information retrieval, and technical document indexing. Unlike Named Entity Recognition (NER), which focuses on proper names and referential entities, ATE targets common nouns and multi word expressions that encode domain knowledge rather than unique identifiers.

The ATE-IT Shared Task at EVALITA 2026 [1] introduces the first large scale evaluation benchmark for Automatic Term Extraction and Term Variants Clustering in Italian. The task focuses on institutional documents related to municipal waste management, a domain characterized by formal administrative language, complex noun phrases, frequent use of acronyms, and extensive lexical variation. These properties make the task particularly challenging for purely statistical or end to end neural approaches.

The shared task is composed of two subtasks. Subtask A addresses Automatic Term Extraction at the sentence level, requiring systems to identify both single word and multi word domain terms under strict constraints. In particular, nested terms are forbidden, surface forms must match exactly, and duplicate extractions within the same sentence are penalized. Subtask B focuses on Term Variants Clustering, where systems must group extracted terms that refer to the same underlying concept, while explicitly avoiding the clustering of hypernym hyponym relations.

In this paper, we present the unified hybrid neural symbolic framework designed to address both subtasks. Our approach combines supervised transformer based sequence labeling with domain aware symbolic post processing for Subtask A, followed by a multi stage clustering pipeline integrating linguistic normalization, semantic embeddings, and strict anti lumping controls for Subtask B. A key design goal of the system is full reproducibility: all components rely exclusively on open source models and libraries and can be executed locally without external API dependencies.

2. Related Work

Early approaches to Automatic Term Extraction relied primarily on statistical association measures such as TF IDF, mutual information, log likelihood ratios, and C-value. These methods often combined frequency based scoring with shallow linguistic filters, such as part-of-speech patterns, to identify candidate terms. While effective in controlled settings, these approaches struggle with contextual ambiguity, long multi word expressions, and domain specific paraphrases.

More recent work formulates ATE as a sequence labeling problem, enabling the use of neural architectures originally developed for Named Entity Recognition. Recurrent neural networks and, more recently, transformer based models such as BERT have demonstrated strong performance due to their ability to capture contextual information. However, purely neural approaches often suffer from overgeneration, producing noisy or partially correct spans that violate task specific constraints such as non nesting and sentence level uniqueness.

Hybrid approaches combining neural models with symbolic post processing have been proposed to address these limitations. By enforcing linguistic and structural constraints after neural prediction, such systems improve precision and interpretability without sacrificing the benefits of contextualized representations.

Term Variants Clustering has traditionally combined linguistic normalization techniques, such as lemmatization and stemming, with vector space similarity measures. With the advent of sentence and phrase embeddings, semantic similarity has become a dominant signal for clustering. Evaluation is commonly performed using BCubed Precision, Recall, and F1, which assess both cluster purity and completeness. Recent baselines explore zero shot large language models for clustering, but these approaches raise concerns regarding reproducibility, computational cost, and controllability.

3. System Description

The system is designed as a modular pipeline addressing both subtasks of the ATE-IT Shared Task. The system follows a hybrid neural symbolic paradigm, where neural models are responsible for capturing contextual and semantic information, while symbolic components enforce task specific constraints and domain knowledge.

For Subtask A, Automatic Term Extraction is modeled as a BIO sequence labeling problem using a fine tuned Italian BERT model. Neural predictions are refined through constraint aware decoding and extensive post processing to enforce non nesting, sentence level uniqueness, and domain relevance.

For Subtask B, extracted terms are clustered using a hybrid approach that combines lemmatization based grouping with semantic similarity computed from multilingual sentence embeddings. A strict clustering configuration introduces additional rule based modules to prevent overly broad or semantically incoherent clusters.

The overall pipeline is fully reproducible and relies exclusively on open source tools, including HuggingFace Transformers, SpaCy, and sentence transformers. No proprietary APIs or external services are required.

3.1. Design Principles

This system is designed around three core principles: (i) precision oriented extraction suitable for downstream terminology management tasks, (ii) modularity, enabling independent experimentation across subtasks, and (iii) full reproducibility using exclusively open source components. Rather than relying on end-to-end generative approaches, the system explicitly separates linguistic modeling from symbolic constraints, allowing finer control over error sources and interpretability.

3.2. Overall System Architecture

This system follows a two stage hybrid neural symbolic pipeline. In the first stage, raw text is processed by a supervised transformer based sequence labeling model to identify candidate term spans. These predictions are subsequently refined through constraint aware decoding and domain specific symbolic filtering to ensure task compliance and terminological precision.

In the second stage, the set of unique extracted terms is passed to a modular clustering pipeline. Linguistic normalization and lemmatization are first applied to capture inflectional variants. Semantic similarity is then computed using multilingual sentence embeddings, and clusters are refined through threshold based merging and explicit anti lumping mechanisms. Optional large language model verification can be applied to ambiguous cases, though it is disabled in the official submission to preserve efficiency and reproducibility.

3.3. Subtask A: Automatic Term Extraction

Subtask A addresses Automatic Term Extraction (ATE) from Italian institutional documents related to municipal waste management. The objective is to identify domain specific terminology, including both single word terms and multi word expressions, under strict task constraints prohibiting nested terms and duplicate extractions within the same sentence. Unlike Named Entity Recognition, the target expressions do not follow canonical entity patterns and frequently involve compositional noun phrases, technical jargon, and administrative formulations.

3.3.1. Data Analysis and Task Characteristics

The training data exhibits substantial sparsity and imbalance. Only a subset of sentences contains annotated terms, and term tokens are heavily outnumbered by non term tokens. The average term length exceeds two tokens, with a long tail of complex multi word expressions reaching up to twenty tokens. These properties motivate a sequence labeling formulation combined with explicit constraint enforcement to control overgeneration.

3.3.2. Preprocessing and Linguistic Normalization

All sentences are lowercased to ensure case insensitive matching, as required by the evaluation protocol. Bracketed content is preserved while removing enclosing characters, and excessive whitespace is normalized. Tokenization is performed using the `it_core_news_sm` SpaCy model, which provides robust handling of Italian morphology, contractions, and punctuation. No lemmatization or stemming is applied at this stage to avoid altering surface forms required for exact match evaluation.

3.3.3. BIO Encoding and Gold Alignment

Gold standard terms are converted into a BIO tagging scheme with labels `B-TERM`, `I-TERM`, and `O`. To correctly handle overlapping and nested annotations in the training data, terms are processed using a longest first strategy: longer multi word expressions are encoded before shorter ones, preventing fragmentation of maximal spans. Token level BIO labels are aligned with transformer subword tokenization using word index mappings, where only the first subword of each token contributes to the loss and subsequent subwords are masked.

3.3.4. Neural Model and Training Configuration

The extraction model is based on `dbmdz/bert-base-italian-uncased`, fine tuned for token classification. The model is trained for five epochs with a batch size of 16 and a learning rate of 2×10^{-5} using the Adam optimizer. Training is performed in a fully supervised manner without any prompting, generative inference, or external language model interaction. Random seeds are fixed to ensure reproducibility.

Due to the extreme dominance of the O label, token level predictions tend to be conservative. While class weighted loss is supported, the primary strategy for mitigating this imbalance is applied at inference time through probability aware decoding rather than hard argmax selection. This design choice allows borderline B-TERM predictions to be retained, improving recall without retraining the model.

3.3.5. Inference and Probability Based Decoding

At prediction time, token level logits are converted into label assignments with careful alignment back to SpaCy tokens. Only one label per original token is retained. Instead of relying exclusively on the most probable label, probability thresholds are explored to capture low confidence but semantically plausible term boundaries. This approach is particularly beneficial for long multi word expressions, where uncertainty accumulates across tokens.

3.3.6. Constraint Aware Term Reconstruction

Predicted BIO sequences are reconstructed into surface form terms by merging contiguous B-TERM and I-TERM spans. A multi stage post processing pipeline is then applied to enforce ATE-IT constraints and remove systematic noise. Duplicate terms within the same sentence are eliminated. Nested terms are removed unless the shorter term occurs independently outside the span of a longer term in the same sentence. This check is implemented using span level matching over the original sentence text.

3.3.7. Domain Specific Filtering

To further improve precision, extracted candidates are filtered using a set of domain aware rules. These include the removal of Italian stopwords, generic verbs, administrative headers, days of the week, isolated prepositions, incomplete fragments, and non domain English terms. Acronyms are retained only if they belong to a curated list of waste management related abbreviations (e.g., RAEE, TARI). Formatting normalization corrects spacing around punctuation and Italian contractions, ensuring consistent surface forms.

3.3.8. Evaluation Protocol

Evaluation follows the official ATE-IT specification. Micro F1 is computed at the sentence level by comparing sets of extracted terms against gold annotations, aggregating true positives, false positives, and false negatives across the corpus. Type F1 evaluates unique term types globally, measuring the system's ability to recover the domain vocabulary independently of sentence context. All comparisons are performed in a case-insensitive manner without lemmatization.

3.4. Subtask B: Term Variants Clustering

Subtask B focuses on clustering term variants that refer to the same underlying concept within the Italian municipal waste management domain. Given a set of unique terms extracted in Subtask A, the goal is to group together inflectional variants, acronyms and their expansions, and true synonyms, while explicitly avoiding the clustering of hypernym–hyponym relations (e.g., *rifiuti* vs. *rifiuti indifferenziati*). The task is evaluated using BCubed Precision, Recall, and F1, which measure cluster coherence and completeness at the item level.

3.4.1. Overview of the Clustering Pipeline

We implement a modular and configurable production pipeline that supports multiple clustering strategies, including lemma based, embedding based, LLM based, and hybrid approaches. The final submission adopts a hybrid strategy that combines linguistic normalization with semantic similarity,

while maintaining strict control over cluster growth and noise. All components are implemented using open source tools and can be executed locally without external API dependencies.

3.4.2. Data Handling and Preprocessing

Input data consists of a list of unique terms derived from the output of Subtask A. Terms are normalized by lowercasing and whitespace stripping. When enabled, linguistic preprocessing is performed using the `it_core_news_sm` SpaCy model to obtain Italian lemmas. Lemmatization is applied conservatively to avoid altering domain specific surface forms while still capturing inflectional variation.

3.4.3. Lemma Based Initial Grouping

As a first step, terms are grouped by their lemmatized forms. This provides an efficient baseline that captures morphological variants such as singular/plural alternations (e.g., *isola ecologica* vs. *isole ecologiche*). Lemma based grouping produces high precision but limited recall, as it does not account for lexical or semantic variation beyond morphology. For this reason, it is primarily used as an initialization step in the hybrid pipeline.

3.4.4. Embedding Based Semantic Clustering

To capture semantic similarity between non identical surface forms, we employ multilingual sentence embeddings computed with the `paraphrase-multilingual-MiniLM-L12-v2` model. Embeddings are computed for all terms and compared using cosine similarity. Several clustering algorithms are supported, including greedy threshold based clustering, DBSCAN, and hierarchical agglomerative clustering.

In the hybrid configuration, embedding similarity is used to merge lemma based clusters whose centroid embeddings exceed a similarity threshold. The similarity threshold is treated as a tunable hyperparameter and is optimized on the training set by maximizing BCubed F1 over a predefined range. The optimal threshold is stored and reused during development and test inference to ensure consistency.

3.4.5. Hybrid Clustering Strategy

The core clustering strategy combines lemma based grouping and embedding based refinement in a two stage pipeline. First, lemma based clusters are created to handle inflectional variants deterministically. Second, semantic similarity between cluster representatives is computed, and clusters are merged when their similarity exceeds the learned threshold. This approach balances precision and recall by exploiting complementary linguistic and semantic signals.

3.4.6. Strict Hybrid Pipeline and Anti Monster Cluster Controls

To further improve precision and prevent the formation of overly large and semantically incoherent clusters, we introduce a strict hybrid pipeline composed of four deterministic modules:

Acronym and Substring Solver A rule based preprocessing module identifies and locks obvious matches before semantic clustering. This includes acronyms and their expansions (e.g., *CCR* vs. *centro comunale di raccolta*) and strict substring matches after removing domain stopwords. Locked clusters are excluded from later merging steps to preserve high confidence relations.

Refined Embedding Clustering with Anti Lumping Logic Embedding based merging is performed using a stricter similarity threshold (0.90) and additional guardrails. Similarity scores are penalized when overlap is driven primarily by generic domain stopwords (e.g., *rifiuti*, *raccolta*). Explicit antonym checks prevent merging terms that differ along critical semantic dimensions (e.g., *differenziato* vs. *indifferenziato*).

Component	Configuration
Embedding Model	paraphrase-multilingual-MiniLM-L12-v2
Similarity Metric	Cosine similarity
Clustering Strategy	Hybrid (lemma + embeddings)
Similarity Threshold	Optimized on training set
Strict Threshold (Anti lumping)	0.90
Monster Cluster Size Limit	Empirical (configurable)
Clustering Algorithms	Greedy, DBSCAN, HAC
LLM Verification	Disabled (official run)

Table 1
Hyperparameters and configuration for Subtask B clustering.

Subtask	Metric	System	Precision	Recall	F1
Subtask A	Micro	Baseline	0.497	0.559	0.526
		Proposed system	0.489	0.395	0.437
	Type	Baseline	0.435	0.508	0.469
		Proposed system	0.528	0.404	0.458
Subtask B	BCubed	Baseline	0.177	0.396	0.245
		Proposed system	0.390	0.333	0.359

Table 2
Official test-set results comparing Proposed system with the baseline for Subtask A (Automatic Term Extraction) and Subtask B (Term Variants Clustering).

Monster Cluster Breaker Clusters exceeding a predefined size threshold are identified as potential "monster clusters". These clusters are recursively split using agglomerative clustering with cosine distance and average linkage. Internal cluster coherence is evaluated via average pairwise similarity, and low coherence groups are further decomposed into smaller clusters or singletons.

3.4.7. Training and Threshold Optimization

The cosine similarity threshold used for embedding based cluster merging was optimized on the training set by maximizing the BCubed F1 score over a predefined range of candidate values. The optimal threshold was found to be 0.85 and was subsequently fixed for all development and test experiments. This value is persisted and reused across runs to ensure consistent clustering behavior and full reproducibility. Such explicit threshold optimization allows the system to adapt to the domain specific distribution of term variants without relying on manual tuning.

3.4.8. Evaluation Protocol

Clustering performance is evaluated using BCubed Precision, Recall, and F1, following the official ATE-IT guidelines. Precision measures cluster purity, while recall reflects the completeness of concept coverage. Evaluation is performed on the development set during model selection and on the test set for final submission. All terms are guaranteed to be assigned to a cluster, with unmatched terms defaulting to singleton clusters.

4. Results

This section reports the official evaluation results of the system on the ATE-IT Shared Task test set. We compare our system against the official baseline provided by the task organizers for both subtasks.

Sentence	Gold Term(s)	System Output	Category
NUOVO CALENDARIO RACCOLTA RIFIUTI	calendario raccolta rifiuti	calendario raccolta rifiuti	True Positive
COMUNE DI AMATO	—	—	True Negative
Raccolte solo per i commercianti	—	raccolte	False Positive
libri, quaderni e album disegno	imballaggi in cartone	—	False Negative

Table 3
Representative qualitative examples for Subtask A.

4.1. Subtask A: Automatic Term Extraction

Table 2 reports the official test-set performance of the proposed system compared with the baseline. The results show that the system achieves competitive precision while trading off recall due to strict constraint-aware decoding and domain-specific filtering.

In Micro evaluation, the proposed system achieves strong precision but lower recall, reflecting a conservative extraction strategy. In Type evaluation, the system improves over the baseline in precision, indicating the ability to produce a cleaner and more coherent terminology inventory at the vocabulary level.

4.1.1. Observations

The extraction component is intentionally precision-oriented. Constraint-aware reconstruction, duplicate removal, and domain-specific filtering significantly reduce noise but also remove borderline valid terms, which lowers recall. This behavior is particularly visible for long multi-word expressions and syntactically complex phrases, where uncertainty accumulates across tokens.

Despite lower recall, the system produces highly consistent and terminologically coherent outputs. Administrative headers, institutional metadata, and non-domain fragments are consistently filtered out. This design choice favors high-quality term inventories suitable for downstream terminology management rather than exhaustive coverage.

Error analysis suggests that most remaining false negatives correspond to very long expressions or implicit domain terms that are difficult to detect through token-level prediction alone.

4.1.2. Qualitative Analysis

To complement the quantitative evaluation, we analyze representative predictions from the test set. Table 3 shows one example for each outcome category: true positive, true negative, false positive, and false negative.

These examples highlight the strengths and limitations of the proposed hybrid neural–symbolic approach. Domain-relevant expressions such as *calendario raccolta rifiuti* are correctly identified even in uppercase titles and non-sentential contexts. At the same time, occasional false positives emerge from generic administrative phrases, while false negatives mainly involve long or structurally complex expressions.

4.2. Subtask B: Term Variants Clustering

4.2.1. Observations

The hybrid and strict hybrid configurations consistently outperform single method baselines. Lemma based clustering yields very high precision but low recall, while embedding based clustering improves recall at the cost of increased noise. The proposed hybrid pipeline achieves a favorable balance by combining deterministic linguistic rules with semantic similarity and explicit anti lumping mechanisms. Error analysis reveals that most remaining errors arise from borderline cases involving abstract admin-

Concept	Gold Cluster Terms	Predicted Cluster Terms	Outcome
Abbandono dei rifiuti	abbandono; abbandono dei rifiuti; abbandono di materiali; abbandono irregolare dei rifiuti	abbandono; abbandono dei rifiuti; abbandono di materiali; abbandono irregolare dei rifiuti	Correct cluster
Calendario di raccolta	calendario; calendario raccolta rifiuti; calendario raccolta differenziata; calendario porta a porta	calendario; calendario raccolta; calendario raccolta rifiuti; calendario raccolta differenziata; calendario porta-a-porta	Correct cluster
Centri di raccolta	centro comunale di raccolta; centro di raccolta; isola ecologica; ecocentro; stazione ecologica	centro comunale di raccolta; centro di raccolta; isola ecologica; ecocentro; stazione ecologica; ccr; c.c.r.	Correct cluster
Servizi di gestione rifiuti	servizio di gestione dei rifiuti urbani; servizio di raccolta; servizi di raccolta	attivazione del servizio; attivazione del servizio di gestione dei rifiuti urbani; avvio del servizio integrato di gestione dei rifiuti urbani	Over grouping

Table 4
Qualitative comparison of gold and predicted clusters for Subtask B.

istrative phrases or subtle hypernymic relations, which remain challenging without deeper semantic modeling.

4.2.2. Qualitative Error Analysis

To complement BCubed based quantitative evaluation, we perform a qualitative analysis by comparing representative gold standard concept groupings with the clusters produced by our system. Table 4 reports illustrative examples highlighting correct cluster formation as well as remaining semantic over grouping phenomena.

The qualitative analysis confirms that the proposed hybrid clustering strategy is effective at capturing strong semantic equivalence relations, particularly for morphologically and lexically related variants. Concepts such as *abbandono dei rifiuti*, *calendario di raccolta*, and *centro di raccolta* are clustered correctly, demonstrating the effectiveness of lemma based initialization combined with embedding based refinement.

At the same time, the analysis reveals instances of semantic over grouping. In particular, expressions related to service activation (e.g., *attivazione del servizio*, *avvio del servizio integrato*) are grouped together with operational waste management services, despite representing procedural or administrative events rather than the service concepts themselves. This behavior highlights the difficulty of distinguishing closely related but conceptually distinct terms in an institutional domain characterized by dense lexical overlap.

Overall, remaining errors primarily involve subtle distinctions between administrative actions and operational concepts, suggesting that future improvements should focus on incorporating ontology aware or role sensitive semantic constraints.

5. Discussion

The experiments highlight several insights. First, neural sequence labeling alone is insufficient for high quality ATE in institutional domains; symbolic post processing is essential for controlling noise. Second, probability aware decoding provides a lightweight alternative to retraining for mitigating class imbalance effects. Third, clustering benefits from explicit mechanisms to prevent over merging, particularly in domains with dense lexical overlap.

Limitations include reduced recall due to conservative filtering and sensitivity of clustering performance to similarity thresholds. These trade offs are deliberate, favoring terminological precision over

coverage.

6. Conclusion

This paper presents the proposed system for the ATE-IT Shared Task at EVALITA 2026. By combining supervised transformer based modeling with domain-aware symbolic constraints and a hybrid clustering pipeline, the system achieves competitive and reproducible performance across both subtasks. Future work will explore class weighted training objectives, CRF-based decoding, and domain adapted embeddings to further improve recall and semantic resolution.

Acknowledgments

We thank the organizers of the ATE-IT Shared Task for providing the dataset and evaluation framework. We are also grateful to Prof. Paolo Torroni (University of Bologna) for his guidance and valuable discussions in the context of the Natural Language Processing course, which contributed to the development and refinement of this work. Finally, we acknowledge the open source NLP community for the tools and models that made this research possible.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools (ChatGPT and Gemini) exclusively for writing assistance, grammar checking, and stylistic refinement. These tools were not used to generate scientific content, design the system, develop models, conduct experiments, or derive results or conclusions. All system implementations, experimental results, analyses, and interpretations are based on original work by the authors, who carefully reviewed and edited all AI-assisted text and take full responsibility for the content of this paper.

References

[1] N. Cirillo, G. M. Di Nunzio, and 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, February 2026, Bari, Italy.

[2] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, 2009.

[3] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[4] A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 79–85, 1998.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 2019.

[6] R. Wang, W. Liu, and C. McDonald. Neural attention models for sequence classification: Analysis and application to key term extraction. *arXiv preprint arXiv:1604.00077*, 2016.

[7] M. Honnibal and I. Montani. spaCy: Industrial-strength Natural Language Processing in Python. Zenodo, 2017.

[8] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of EMNLP-IJCNLP*, 2019.