ATE-IT at EVALITA 2026: Overview of the Automatic Term Extraction Italian Testbed Task

Nicola Cirillo[1,*], Giorgio Maria Di Nunzio[2,†] and Federica Vezzani[3,†]

[1]Department of Political and Communication Sciences, University of Salerno, 132 Via Giovanni Paolo II, Fisciano (SA), 84084, Italy
[2]Department of Information Engineering, University of Padova, Via Gradenigo 6/b, 35131 Padova, Italy
[3]Department of Linguistic and Literary Studies, University of Padova, Via Elisabetta Vendramini, 13 35137 Padova, Italy

Abstract

This paper presents an overview of the Automatic Term Extraction Italian Testbed (ATE-IT) shared task, organised within the EVALITA 2026 evaluation campaign. The task addresses the scarcity of benchmarks for Italian Automatic Term Extraction (ATE) by proposing a challenge focused on the domain of municipal waste management. Participants were invited to tackle two subtasks: (A) *Term Extraction*, aiming to identify domain-specific terms in institutional texts, and (B) *Term Variants Clustering*, focusing on grouping morphological and semantic term variants. Nine teams participated, submitting a total of 13 runs. The comparative analysis reveals that fine-tuned Transformer architectures generally outperform naive zero-shot Large Language Model (LLM) prompting, while hybrid approaches appear most effective for semantic clustering.

Keywords

Automatic Term Extraction, Terminology, Italian NLP, Shared Task, EVALITA 2026

1. Introduction

Automatic Term Extraction (ATE) is a foundational task in NLP and terminology work [1, 2]. Its goal is to identify domain-specific terms that designate key concepts within a specialised field of knowledge.

Although it shares some similarities with Named Entity Recognition (NER) [3], ATE differs from it. NER involves identifying and classifying mentions of named entities in running text (e.g., people, organisations, places, dates, etc.). Its focus is usually on proper names or unique references that have a clear instance-level referent, and the output is often linked to knowledge bases (e.g., "Barack Obama →Person", "Google →Organization"). ATE, by contrast, aims to extract domain-specific terms from a corpus. This means identifying both multi-word and single-word terms that are relevant to a specialised field of knowledge (e.g., "informed consent", "cryptic species", "blockchain consensus algorithm"). The terms extracted through ATE serve as essential building blocks for downstream tasks such as information retrieval, machine translation, ontology construction, knowledge graph enrichment, and domain adaptation of large language models (LLMs).

In this paper, we describe the Automatic Term Extraction Italian Testbed (ATE-IT) shared task, organised in the context of EVALITA 2026, the 9th evaluation campaign of NLP and speech tools for Italian [4]. ATE-IT is the first large-scale evaluation campaign on Italian ATE, centred on a clearly defined real-world scenario: terminology extraction from institutional texts in the domain of waste management. This domain presents a wide variety of derived terms (e.g., "ecodizionario", "biodigestore"), synonyms (e.g., "indifferenziato" and "secco residuo"), abbreviations ("TARI", "RAEE"), and multiword expressions ("mastello contenitore", "raccolta porta a porta"), making it an important testbed for assessing the robustness of different approaches.

All datasets, evaluation scripts, and baseline code are publicly available at the task repository.[1]

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT
*Corresponding author.
†These authors contributed equally.
nicirillo@unisa.it (N. Cirillo); giorgiomaria.dinunzio@unipd.it (G. M. D. Nunzio); federica.vezzani@unipd.it (F. Vezzani)
0000-0002-2107-1313 (N. Cirillo); 0000-0001-7116-9338 (G. M. D. Nunzio); 0000-0003-2240-6127 (F. Vezzani)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]https://github.com/nicolaCirillo/ate-it

The remainder of this paper is organized as follows: Section 2 and Section 3 discuss the related work and the motivation behind the task. Sections 4 through 7 detail the experimental setup, including the task definition, dataset construction, evaluation measures, and the baseline system. Section 8 introduces the participating systems, while Section 9 and Section 10 present the official results and a discussion of the key findings. Finally, Section 11 concludes the paper.

2. Related Work

Automatic Term Extraction (ATE) has evolved significantly over the last few decades, transitioning from rule-based and statistical pipelines to deep learning and, most recently, prompt-based approaches. Traditional ATE systems generally follow a three-step pipeline: candidate extraction, feature calculation, and probability estimation [5, 6]. Candidate extraction typically relies on linguistic filters [7, 8], while feature calculation and probability estimation often exploit statistical measures [9, 10]. With the advent of deep learning, ATE has increasingly been framed as a sequence labeling task, bypassing the three-step pipeline. Early neural approaches utilized LSTMs [11], but the state-of-the-art shifted rapidly toward Transformer-based models [12, 13]. Despite their effectiveness, supervised deep learning models require substantial labeled datasets, which are often unavailable for specific languages or domains. This limitation has led to the exploration of LLMs and prompting techniques. Recent studies by [14] and [15] compared fine-tuned encoders (like XLM-RoBERTa) against generative models (like GPT-3.5) in few-shot scenarios. Their findings suggest a clear trade-off: while fine-tuned models excel when data is abundant, prompt-based approaches offer a more robust solution in low-resource and few-shot settings.

Despite the notable performances reached by state-of-the-art ATE techniques, there is still room for improvement, especially in multilingual and domain-specific contexts [2]. Moreover, performance remains modest on complex datasets, and most approaches still struggle with domain sensitivity.

Evaluation of ATE techniques relies on benchmark corpora such as GENIA [16] and ACL RD-TEC [17], for English, or the multilingual ACTER [18], for English, French, and Dutch. Although Italian has been underrepresented in ATE benchmarks, existing bilingual corpora such as BitterCorpus [19] and MAGMATic [20] include annotated terms. However, they are primarily designed to evaluate domain-specific machine translation rather than term extraction.

Several shared tasks addressed ATE. Notably, the TermEval 2020 Shared Task on Automatic Term Extraction [21] compared a range of different techniques using the multilingual ACTER corpus. More recently, shared tasks have evolved beyond simple term recognition to emphasize deep semantic understanding and domain-specific relevance. For example, SimpleText Task 2 at CLEF 2024 [22, 23] focused on identifying and explaining complex concepts in scientific abstracts. Participants were asked not only to extract challenging terms but also to generate informative definitions or explanations. Similarly, the ongoing GutBrain Interplay Task3 [24] at BioASQ CLEF 2025 [25] targets the extraction of structured biomedical knowledge related to the gut-brain axis. Its subtasks include term span classification and relation identification, emphasizing fine-grained categorisation and concept linking in highly specialized biomedical texts.

3. Motivation

ATE systems support terminologists, translators, and technical communicators in building and maintaining controlled vocabularies and thesauri for various specialized domains. Moreover, these systems contribute to improving domain-specific language resources used by AI systems for regulatory compliance, automated indexing, and smart information retrieval.

The international interest in ATE is evidenced by the increasing number of European and global research initiatives. These range from conferences and summer schools, such as MDTT[2] and TSS,[3], to

[2]https://mdtt2026.dei.unipd.it/
[3]https://www.termnet.org/english/products_service/summer_school.php

international ATE shared tasks [21, 23], and the development of numerous gold-standard datasets in English and other languages [16, 17, 18].

Despite the growing momentum in multilingual terminology extraction, the availability of standardized evaluation benchmarks for Italian remains limited. While many shared tasks have been organized for English and select other languages, Italian still lacks a well-defined evaluation framework and publicly available datasets to foster comparative research.

Within this landscape, ATE-IT aims to advance ATE research by introducing a dedicated benchmark for the Italian language, focused on the domain of waste management.

Moreover, by combining term extraction with term variants clustering, ATE-IT aligns with ongoing efforts in terminology standardization, machine translation adaptation, and knowledge base population. Crucially, the proposed evaluation framework facilitates comparability and fosters the application of zero-shot and few-shot learning methods, which are fundamental for low-resource languages like Italian. The task will also promote methodological comparisons between rule-based systems, statistical models, and LLM-based architectures in the context of under-resourced institutional Italian.

4. Definition of the Task

The ATE-IT shared task comprises two subtasks of increasing complexity: Term Extraction and Term Variants Clustering. Both subtasks are designed to be linguistically and computationally challenging. The former requires models to generalize from sparse domain-specific training examples. The latter requires semantic comparison and abstraction over morphologically and syntactically diverse variants.

4.1. Subtask A: Term Extraction

Participants receive a set of sentences drawn from a specialized corpus related to municipal waste management. For each sentence, the goal is to identify and extract the terms that are relevant to the waste management domain. Terms may consist of single words (single-word terms) or multiword expressions (multi-word terms), including nouns, verbs, and adjectives.

4.2. Subtask B: Term Variants Clustering

From the list of unique extracted terms, participants are then required to cluster together those terms that refer to the same underlying concept. For example, “raccolta porta a porta” and “raccolta domiciliare” should be placed in the same cluster. Each cluster should represent a single, coherent concept within the waste management domain. This subtask focuses on synonymy, lexical variation, and compositional semantics.

5. Dataset

The dataset for the Term Extraction subtask comprises sentences paired with the corresponding waste management terms. This dataset is partitioned into a training set of 2,308 sentences, a development set of 577, and a test set of 1,142.

The dataset for the Term Variants Clustering subtask is derived directly from the unique terms identified in the extraction task. Within this dataset, each term is mapped to a cluster ID representing a single, coherent concept in the waste management domain. The clustering data includes 713 terms for training and 242 for development, with a test set of 378. Notably, 118 terms from the test set also appear in either the training or development sets.

Both the training and development sets are sourced from the publicly available ItaIst-TermRifiuti corpus and the ItaIst-WasteLexicon termbase [26, 27]. Conversely, the test set is specifically designed for the ATE-IT task.

Table 1
ATE-IT dataset composition.

Subtask	Split	Size	Items	Geographic and Temporal Coverage
Term Extraction	Training	2,308	Sentences	Campania 2005-2023
	Development	577	Sentences	Campania 2005-2023
	Test	1,142	Sentences	Italy 2025
Term Variants Clustering	Training	713	Terms	Campania 2005-2023
	Development	242	Terms	Campania 2005-2023
	Test	378	Terms	Italy 2025

5.1. Training and Development Sets

The training and development sets are derived from the ItaIst-TermRifiuti corpus and the ItaIst-WasteLexicon termbase [26, 27]. ItaIst-TermRifiuti is a stratified sample of the broader ItaIst-DdAC_GRU corpus [28], which contains institutional texts regarding municipal waste management. The corpus is carefully balanced between:

- **Administrative acts:** Ordinances, service charters, and tenders.
- **Informative texts:** Public notices, guides, and press releases.

This balance facilitates the study of terminological variation across different registers and target audiences. Four trained annotators manually identified terminological units and categorised them into domains such as waste management, law/administration, and environment.

The ItaIst-WasteLexicon termbase provides the conceptual backbone for the dataset, containing about 950 terms organised within a framework that encodes relations such as generic/specific and comprehensive/partitive. Concepts are further enriched with definitions sourced from European and Italian legislation.

In the final Term Extraction dataset, terms are associated with a sentence only if they were identified by at least one annotator and exist within ItaIst-WasteLexicon. Consequently, the dataset may exhibit certain inconsistencies. These are intentionally preserved to provide a "realistic" manually annotated environment, challenging participating systems to demonstrate robustness and generalizability to real-world data.

5.2. Test Set

The test set was specifically curated as a benchmark for the ATE-IT task. To evaluate system generalisation, it incorporates more recent documents (primarily from 2025) and covers a broader geographic scope than the training data. Specifically, while the ItaIst-TermRifiuti dataset is composed primarily of documents from the Campania region, the test set consists of documents collected from 32 municipalities that were sampled with probability proportional to their population, promoting nationwide coverage of Italy.

The initial annotation of the test set was performed by 59 students from the Translation-oriented Terminography of the Department of Linguistic and Literary Studies of the University of Padova. Each student was assigned a unique corpus segment and instructed to adhere to the annotation guidelines already established for the ItaIst-TermRifiuti dataset. Subsequently, sentences were sampled to ensure that the register balance remained consistent with the training and development sets. To finalize the corpus, the authors conducted a validation and correction phase. Through discussion, annotation principles were harmonized to resolve inconsistencies and establish a reliable gold standard.

The test set for the Term Variants Clustering subtask was generated from the list of unique terms identified within this corpus. The terms were then manually clustered by the three authors after a discussion phase.

5.3. Format

The dataset for the term extraction task is a collection of records where each entry maps a specific sentence to its corresponding domain-specific terms. Each record contains the source metadata (consisting of a unique `document_id`, a `paragraph_id`, and a `sentence_id`) alongside the raw `sentence_text`. The target annotations are encapsulated in a terms field, which lists the single-word and multi-word terms identified as domain-relevant concepts within that sentence. To maintain technical consistency, the dataset adheres to a "longest match" constraint, where nested terms are excluded in favour of the most complete expression (e.g., "impianto di trattamento rifiuti" is extracted while its constituent "trattamento rifiuti" is omitted). Furthermore, all terms are normalised to lowercase and appear without duplicates for any given sentence.

This schema is implemented in both CSV and JSON formats. Notably, in the CSV format, there is a separate row for each term, whereas in the JSON format, the terms are stored within a `term_list` array.

6. Evaluation Measures

The Term Extraction subtask is evaluated using two separate scores: Micro F1 score [29], which evaluates Precision and Recall across all term occurrences in the dataset, and Type F1 score, which assesses the ability to identify unique term types correctly. The Term Variants Clustering subtask is evaluated by using BCubed F1 [30, 31], a metric that assesses clustering quality.

6.1. Evaluation of Subtask A

To provide a comprehensive evaluation of Subtask A, we distinguished between two separate capabilities of the system: first, its ability to identify individual term mentions as they appear in running text, and second, its ability to successfully extract a unique set of terms from the corpus as a whole. Consequently, we produced two distinct rankings based on Micro F1 and Type F1 scores to reflect these separate goals.

Micro F1 is calculated by aggregating the counts of true positives, false positives, and false negatives across all sentences s in the dataset D. Specifically:

- TP_s: number of terms correctly extracted from sentence s;
- FP_s: number of terms incorrectly extracted from sentence s;
- FN_s: number of gold standard terms in s that were missed.

The micro-averaged Precision, Recall, and F1 are defined as shown in equation (1).

$$P_m = \frac{\sum TP_s}{\sum(TP_s + FP_s)}, \quad R_m = \frac{\sum TP_s}{\sum(TP_s + FN_s)}, \quad F1_m = \frac{2 \cdot P_m \cdot R_m}{P_m + R_m} \tag{1}$$

The Type F1 score is computed over the set of unique term types (i.e., distinct term forms appearing at least once in the dataset). We define the following variables:

- TP_t: number of unique extracted terms that match the gold standard;
- FP_t: number of unique extracted terms that do not appear in the gold standard;
- FN_t: number of unique gold standard terms that were not extracted.

The Precision, Recall, and F1 for term types are defined as in equation (2).

$$P_t = \frac{TP_t}{TP_t + FP_t}, \quad R_t = \frac{TP_t}{TP_t + FN_t}, \quad F1_t = \frac{2 \cdot P_t \cdot R_t}{P_t + R_t} \tag{2}$$

```
You are an automatic term extraction agent. You
     will receive a list of sentences as input.
Your role is to extract waste management terms from
      the sentences. Output a list of terms for
     each sentence.

strictly adhere to the Example Output Format:
Sentence 1: [term1; term2]
Sentence 2: [term5; term6]
...

Instructions:
* Extract only terms, ignore named entities;
* Do not extract nested terms;
* Extract only terms related to waste management;
* If no terms, output an empty list [];
* You must output 20 lists of terms.
```

```
You are a term clustering agent.
You will receive a list of term clusters and a list
      of unclustered terms related to municipal
     waste management.
Your task is to cluster together exact synonyms.
     Each cluster must represent a single concept.

Output:
Return the list of clusters with the newly added
     terms. Each cluster on a new line.

Instructions:
* Group terms by meaning, not form. Use their lemma
     .
* Focus on meaning within waste management context.
* If a term does not belong to a cluster, create a
     new cluster.
```

Figure 1: System prompts used in the baseline implementation for Subtasks A and B.

6.2. Evaluation of Subtask B

To assess the quality of the clustering in Subtask B, we employed the BCubed metric. This measure is calculated by computing Precision and Recall at the item level and subsequently averaging these scores across all items. Crucially, participants were required to cluster the unique terms extracted by their own systems in Subtask A. Consequently, the set of items in the predicted clustering does not perfectly align with the gold standard. To address this issue, our evaluation framework explicitly accounts for the discrepancy between the two sets.

To compute the BCubed scores, we define the following variables:

- N_{pred}: the total number of elements in the predicted clustering;
- N_{gold}: the total number of elements in the gold clustering;
- $C(x)$: the predicted cluster containing element x (if x is not present in the predicted clustering, $C(x) = \emptyset$);
- $L(x)$: the gold cluster containing element x (if x is not present in the gold clustering, $L(x) = \emptyset$).

For each item x, the item-level Precision and Recall are calculated as in equation (3).

$$P(x) = \frac{|\{y \in C(x) \,:\, L(y) = L(x)\}|}{|C(x)|}, \qquad R(x) = \frac{|\{y \in L(x) \,:\, C(y) = C(x)\}|}{|L(x)|} \tag{3}$$

Finally, the global scores are derived by averaging these item-level values over the respective total counts, and the harmonic mean is taken to produce the final F1 score, as shown in equation (4).

$$P_{b^3} = \frac{1}{N_{pred}} \sum_x P(x), \quad R_{b^3} = \frac{1}{N_{gold}} \sum_x R(x), \quad F1_{b^3} = \frac{2 \cdot P_{b^3} \cdot R_{b^3}}{P_{b^3} + R_{b^3}} \tag{4}$$

7. Baseline

To quantify the task's complexity and provide a reference point for all participating systems, we provide a baseline system built on the `gemini-2.5-flash` model in a zero-shot setup. The choice of this model is motivated by its status as a state-of-the-art LLM with robust zero-shot capabilities. It reflects current general-purpose AI performance, allowing for a clear comparison between zero-shot LLM prompting and the specialised, domain-tuned approaches that the task seeks to promote.

Table 2
Results for Subtask A: Term Extraction.

Team	Method	P_m	R_m	$F1_m$	P_t	R_t	$F1_t$	Rank ($F1_m$)	Rank ($F1_t$)
SMTE	BERT+spaCy	**.656**	.577	**.614**	.645	.529	**.581**	1	1
TrietNLP	RoBERTa+CRF	.634	.568	.599	.599	**.545**	.571	2	2
TEXA (run 1)	BERT	.617	**.578**	.597	.576	*.460*	.512	3	5
OA-TE (run 2)	BERT+CRF	.581	*.522*	.550	.569	*.492*	.528	4	4
MinseokKIM	CRF	.569	*.476*	*.519*	.654	*.444*	.529	5	3
OA-TE (run 1)	BERT+CRF	.560	*.446*	*.497*	.595	*.415*	.489		
Juliette Tonneau	CRF	.555	*.448*	*.496*	.561	*.447*	.498	6	6
TEXA (run 2)	Gemini (zero-shot)	*.471*	*.514*	*.492*	*.425*	*.489*	*.455*		
Peacemaker	BERT	.497	*.476*	*.486*	*.430*	*.455*	*.442*	7	8
TermNinjas	BERT	*.489*	*.395*	*.437*	.528	*.404*	*.458*	8	7
Valentinitalie	Random Forest	*.364*	*.473*	*.411*	**.707**	*.262*	*.382*	9	9
baseline	Gemini (zero-shot)	.497	.559	.526	.435	.508	.469		

The Term Extraction baseline model was instructed to identify and extract domain-specific terms from batches of 20 sentences. The system prompt, illustrated in Figure 1, establishes the extraction rules and the required output structure, while the user prompt provides 20 sentences per call.

For Term Variants Clustering, the model was instructed to group synonyms by comparing batches of 20 terms against an existing set of clusters. The system prompt, illustrated in Figure 1, provides the clustering rules while the user prompt feeds the current state of the clusters, followed by the 20 unclustered terms to be processed.

8. Participating Systems

A total of 9 teams participated in the ATE-IT shared task. For Subtask A, 11 runs were submitted, while Subtask B saw a lower participation rate with only 2 submitted runs.

The majority of runs for Subtask A (7 out of 11) leveraged Transformer-based models, specifically the Italian versions of BERT and the multilingual XLM-RoBERTa. Within this group, two teams (TrietNLP [32] and OA-TE [33]) placed a CRF (Conditional Random Field) layer on top of the model to better capture global dependencies in term spans, while the other four (SMTE [34], TEXA [35], Peacemaker [36], and TermNinjas [37]) relied on a token classification layer. The winning system, SMTE, combined BERT with a spaCy-based NER pipeline through a specific merging and vocabulary filtering strategy. In contrast, teams such as MinseokKIM [38] and Juliette Tonneau [39] relied on traditional CRF classifiers, focusing on feature engineering. A hybrid approach was proposed by Valentinitalie [40], which utilised a Random Forest classifier fed by candidates extracted through a combination of rule-based patterns and LLM prompting. Finally, zero-shot prompting was explored by the TEXA team.

Only two teams, TrietNLP and TermNinjas, participated in Subtask B. TrietNLP proposed a pipeline that integrates pre-clustering based on Levenshtein distance with prompting. Similarly, TermNinjas employed lemmatisation during pre-clustering and subsequently merged clusters based on word embedding similarity.

9. Results

The results for Subtask A (Term Extraction) and Subtask B (Term Variants Clustering) are presented in Table 2 and Table 3, respectively.

Table 3
Results for Subtask B: Term Variants Clustering.

Team	Method	P_{b^3}	R_{b^3}	$F1_{b^3}$	Rank ($F1_{b^3}$)
TrietNLP	levenshtein+Gemini	**.528**	*.378*	**.441**	1
TermNinjas	lemmatization+embedding	.390	.333	*.359*	2
baseline	Gemini (zero-shot)	.177	**.396**	.245	

9.1. Results for Subtask A

In Subtask A, overall performance was robust, with several systems exceeding an $F1_m$ score of .55. The system submitted by SMTE achieved the highest performance across both micro- and type-based metrics ($F1_m = .614$, $F1_t = .581$). This result highlights the efficacy of integrating BERT-based architectures with specialised NLP pipelines, such as spaCy, for refined terminology extraction.

While TrietNLP and TEXA recorded comparable Micro F1 scores (.599 and .597, respectively), their performance on unique term types ($F1_t$) diverged by a significant margin of .059. This discrepancy suggests that while standard BERT models are effective at identifying frequent terminological mentions, the RoBERTa+CRF architecture employed by TrietNLP is better suited for identifying the "long tail" of rare variants and unseen terms.

The performance of Valentinitalie is noteworthy; despite recording the lowest overall F1 scores ($F1_m = .411, F1_t = .382$), it achieved the highest Type Precision across all participants ($P_t = .707$). However, subsequent analysis reveals that the extracted terms were almost exclusively present in the training set. This indicates a conservative extraction strategy that prioritized precision over the ability to generalise to novel terminology in the test set.

General trends across the submissions indicate that supervised approaches typically achieved higher Precision than Recall. Conversely, zero-shot models favoured Recall, which may suggest a superior ability to identify term boundaries and adhere to the strict "longest match" constraint. Finally, the consistent disparity between Micro F1 and Type F1 scores across all teams confirms that identifying frequent term mentions remains significantly easier than discovering the full diversity of unique terms within a specialised corpus.

9.2. Results for Subtask B

Participation in the Term Variants Clustering subtask was significantly lower than in Subtask A, with only two teams submitting valid runs, reflecting the higher complexity of the challenge.

The best performance was achieved by TrietNLP, which recorded a BCubed F1 score of .441. Their hybrid approach, combining Levenshtein distance for morphological pre-clustering with a Gemini-based component for semantic aggregation, proved effective in maintaining high purity within clusters, as evidenced by the highest Precision score ($P_{b^3} = .528$).

TermNinjas ranked second with an F1 score of .359. Their strategy, which relied on lemmatization and word embedding similarity, struggled to match the Precision of the top system, achieving a P_{b^3} of .390.

Notably, the zero-shot baseline achieved the highest Recall ($R_{b^3} = .396$) but the lowest Precision ($P_{b^3} = .177$). This inverse relationship suggests that the baseline tended to over-cluster, aggressively grouping terms together. This behavior underscores the primary challenge of the task: effectively distinguishing exact synonyms from other semantic relations, thereby ensuring that hyponyms and hypernyms are not erroneously merged into the same concept cluster.

10. Discussion

The results of the ATE-IT shared task provide a snapshot of the current capabilities in Italian ATE. Beyond the individual rankings, several major trends emerged: the reliability of deep learning, the necessity of structural decoding for accurate term boundary detection, the challenge of generalisation over diatopic and diachronic variations, and the emergence of hybrid approaches for term variants clustering.

10.1. Deep learning vs. LLMs prompting

The campaign confirms the resilience of standard supervised deep learning architectures in the era of LLMs, aligning with findings already established in the broader ATE literature [5]. While the naive zero-shot baseline achieved a respectable performance, it was consistently outperformed by more lightweight systems based on BERT and RoBERTa architectures.

This result reinforces that for ATE, where a training set exists, fine-tuned models offer a better alternative to prompting LLMs, making them suited for real-world applications, where speed and cost are key factors.

10.2. Term boundary detection

Moreover, results suggest that Transformer architectures alone are often insufficient for accurate term boundary detection. In fact, the top-performing systems augmented it with structural decoding mechanisms like the spaCy-based pipeline of SMTE, or the CRF layer integrated by TrietNLP. These choices proved decisive in enforcing the "longest match" constraint.

Terminology in the waste management sector is highly compositional (e.g., *"raccolta"* vs. *"raccolta differenziata"* vs. *"raccolta differenziata porta a porta"*). Therefore, purely statistical models may fragment these multi-word terms, whereas systems with explicit boundary modelling or post-processing heuristics successfully captured the full syntactic structure of these terminological units.

10.3. Diatopic and diachronic variations

Another crucial insight from this evaluation is the significant impact of temporal and geographic shifts on model performance. During the development phase, several systems reportedly achieved F1 scores exceeding .70; however, performance dropped noticeably on the test set ($F1_m \approx .60$ for the best system). This decline highlights the difficulty of the realistic scenario proposed by ATE-IT.

This suggests that even robust deep learning models tend to overfit to specific regional denominations or temporal periods. The challenge for future Italian ATE lies in improving its generalization capabilities, since terminology may exhibit significant diatopic variation and evolve diachronically.

10.4. The "long tail" challenge

The consistent disparity between Micro F1 and Type F1 scores across all teams indicates that identifying frequent term mentions is significantly easier than discovering the full diversity of unique terms.

This "long tail" problem is particularly relevant for the intended application of ATE systems. For a terminologist or a translator, high-frequency terms (e.g., *"rifiuto"*, *"organico"*) are often already known and documented. The real value of an extraction system lies in its ability to surface rare, highly specific, or emerging terms (e.g., *"presidi sanitari monouso"*, *"pseudo-edili"*).

10.5. Hybrid approaches for semantic clustering

Finally, the results of Subtask B shed light on the complexity of Term Variants Clustering. The baseline performance was low ($F1_{b^3} = .245$), confirming that simple zero-shot prompting struggles to effectively group domain-specific synonyms without guidance. As observed in the results, the main challenge lies

in the strict definition of synonymy required by the task; systems must navigate the subtle semantic boundary that separates synonyms from hypernyms and hyponyms, a distinction where the baseline frequently faltered.

The winning team, TrietNLP, achieved a significantly higher score by employing a hybrid approach. The result suggests that an effective strategy for addressing this task appears to be a composite pipeline: using symbolic methods to handle surface-level morphological variations and leveraging the reasoning capabilities of LLMs to enforce semantic equivalence.

11. Conclusion

The ATE-IT shared task addressed the need for a dedicated benchmark for Automatic Term Extraction in Italian. By introducing a dataset characterized by marked terminological variation, the campaign allowed for a comparative analysis of diverse computational approaches.

The results from the participation of 9 teams and the evaluation of 13 runs point to three main findings. First, regarding Term Extraction, fine-tuned Transformer architectures outperformed the naive zero-shot baseline provided for the task. This indicates that while simple prompting strategies are insufficient for high-precision extraction, supervised models remain a reliable standard for this specific workload. However, given the rapid evolution of generative AI, these results should not discount the potential of LLMs; rather, they suggest that future research should investigate more advanced prompting techniques to bridge the gap with dedicated supervised systems.

Second, the performance gap observed between the development and test sets indicates that geographic and temporal generalisation remains a complex issue. Current models show a tendency to overfit to the specific term variants of the training data, resulting in lower F1 scores when applied to documents from different municipalities.

Finally, the low baseline results in Term Variants Clustering suggest that synonym clustering requires more than simple semantic similarity. The most effective system adopted a hybrid strategy, combining symbolic methods for morphological matching with LLMs for semantic disambiguation.

We hope the open release of the ATE-IT dataset will support the community in developing more robust Automatic Term Extraction methodologies specifically tailored to the complexities of the Italian language.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini in order to: Drafting content, Improve writing style, and Abstract generation. Further, the authors used Gemini for equations 1 to 4 in order to: Formatting assistance. Finally, the authors used Gemini and Grammarly in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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.

[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] B. Jehangir, S. Radhakrishnan, R. Agarwal, A survey on Named Entity Recognition — datasets, tools, and methodologies, Natural Language Processing Journal 3 (2023) 100017. doi:10.1016/j.nlp.2023.100017.

[4] 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.

[5] J. C. Blandón Andrade, C. M. Medina Otálvaro, C. M. Zapata Jaramillo, A. Morales Ríos, Approaches, Tools, Algorithms, and Methods for Automatic Term Extraction: A Systematic Literature Mapping, Journal of Intelligent & Fuzzy Systems (2025) 18758967251392652. doi:10.1177/18758967251392652.

[6] N. A. Astrakhantsev, D. G. Fedorenko, D. Yu. Turdakov, Methods for automatic term recognition in domain-specific text collections: A survey, Programming and Computer Software 41 (2015) 336–349. doi:10.1134/S036176881506002X.

[7] J. S. Justeson, S. M. Katz, Technical terminology: Some linguistic properties and an algorithm for identification in text, Natural Language Engineering 1 (1995) 9–27. doi:10.1017/S1351324900000048.

[8] A. L. Meyers, Y. He, Z. Glass, J. Ortega, S. Liao, A. Grieve-Smith, R. Grishman, O. Babko-Malaya, The Termolator: Terminology Recognition Based on Chunking, Statistical and Search-Based Scores, Frontiers in Research Metrics and Analytics 3 (2018). doi:10.3389/frma.2018.00019.

[9] K. Ahmad, L. Gillam, L. Tostevin, University of surrey participation in TREC8: weirdness indexing for logical document extrapolation and retrieval (WILDER), in: E. M. Voorhees, D. K. Harman (Eds.), Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999, volume 500-246 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1999. URL: http://trec.nist.gov/pubs/trec8/papers/surrey2.pdf.

[10] P. Drouin, Term extraction using non-technical corpora as a point of leverage, Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 9 (2003) 99–115. doi:10.1075/term.9.1.06dro.

[11] M. Kucza, J. Niehues, T. Zenkel, A. Waibel, S. Stüker, Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks, in: Proc. Interspeech 2018, 2018, pp. 2072–2076. doi:10.21437/Interspeech.2018-2017.

[12] A. Hazem, M. Bouhandi, F. Boudin, B. Daille, TermEval 2020: TALN-LS2N System for Automatic Term Extraction, in: B. Daille, K. Kageura, A. R. Terryn (Eds.), Proceedings of the 6th International Workshop on Computational Terminology, European Language Resources Association, Marseille, France, 2020, pp. 95–100.

[13] C. Lang, L. Wachowiak, B. Heinisch, D. Gromann, Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 3607–3620. doi:10.18653/v1/2021.findings-acl.316.

[14] S. Banerjee, B. R. Chakravarthi, J. P. McCrae, Large Language Models for Few-Shot Automatic Term Extraction, in: A. Rapp, L. Di Caro, F. Meziane, V. Sugumaran (Eds.), Natural Language Processing and Information Systems, Springer Nature Switzerland, Cham, 2024, pp. 137–150. doi:10.1007/978-3-031-70239-6_10.

[15] H. T. H. Tran, C.-E. González-Gallardo, J. Delaunay, A. Doucet, S. Pollak, Is Prompting What Term Extraction Needs?, in: E. Nöth, A. Horák, P. Sojka (Eds.), Text, Speech, and Dialogue, Springer Nature Switzerland, Cham, 2024, pp. 17–29. doi:10.1007/978-3-031-70563-2_2.

[16] J. D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, Genia corpus - a semantically annotated corpus for bio-textmining, in: Bioinformatics, volume 19, Oxford University Press, 2003. doi:10.1093/bioinformatics/btg1023.

[17] B. Qasemizadeh, A.-K. Schumann, The acl rd-tec 2.0: A language resource for evaluating term extraction and entity recognition methods, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), 2016, pp. 1862–1868.

[18] A. Rigouts Terryn, V. Hoste, E. Lefever, In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora, Language Resources and Evaluation 54 (2020) 385–418. doi:10.1007/s10579-019-09453-9.

[19] M. Arcan, M. Turchi, S. Topelli, P. Buitelaar, Enhancing statistical machine translation with bilingual terminology in a cat environment, in: Y. Al-Onaizan, M. Simard (Eds.), Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track, Association for Machine Translation in the Americas, 2014, pp. 54–68.

[20] R. Scansani, L. Bentivogli, S. Bernardini, A. Ferraresi, Magmatic: A multi-domain academic gold standard with manual annotation of terminology for machine translation evaluation, in: M. Forcada, A. Way, B. Haddow, R. Sennrich (Eds.), Proceedings of Machine Translation Summit XVII: Research Track, European Association for Machine Translation, 2019, pp. 78–86.

[21] A. Rigouts Terryn, V. Hoste, P. Drouin, E. Lefever, TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset, in: B. Daille, K. Kageura, A. R. Terryn (Eds.), Proceedings of the 6th International Workshop on Computational Terminology, European Language Resources Association, Marseille, France, 2020, pp. 85–94.

[22] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, G. M. Di Nunzio, F. Vezzani, J. D'Souza, J. Kamps, Overview of the CLEF 2024 SimpleText Track, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2024, pp. 283–307. doi:10.1007/978-3-031-71908-0_13.

[23] G. M. Di Nunzio, F. Vezzani, V. Bonato, H. Azarbonyad, J. Kamps, L. Ermakova, Overview of the CLEF 2024 SimpleText Task 2: Identify and Explain Difficult Concepts, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740 of *CEUR Workshop Proceedings*, CEUR, Grenoble, France, 2024, pp. 3129–3146.

[24] M. Martinelli, G. Silvello, V. Bonato, G. M. Di Nunzio, N. Ferro, O. Irrera, S. Marchesin, L. Menotti, F. Vezzani, Overview of GutBrainIE@CLEF 2025: Gut-Brain Interplay Information Extraction, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, CEUR, Madrid, Spain, 2025, pp. 65–98.

[25] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. R. Ortega, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, L. Menotti, G. Silvello, G. Paliouras, BioASQ at CLEF2025: The Thirteenth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 407–415. doi:10.1007/978-3-031-88720-8_61.

[26] N. Cirillo, D. Vellutino, D. Nicoletti, E. Sabarese, B. Rubino, Itaist_gru, 2025. URL: https://doi.org/10.5281/zenodo.15173712. doi:10.5281/zenodo.15173712.

[27] N. Cirillo, Risorse Linguistiche Digitali per la Transizione Verde, Phd thesis, University of Salerno, 2025.

[28] D. Vellutino, N. Cirillo, Corpus «itaist»: Note per lo sviluppo di una risorsa linguistica per lo studio dell'italiano istituzionale per il diritto di accesso civico, Italiano LinguaDue 16 (2024) 238–250.

[29] R. Verborgh, M. Röder, R. Usbeck, A.-C. Ngonga Ngomo, Gerbil – benchmarking named entity recognition and linking consistently, Semant. Web 9 (2018) 605–625. URL: https://doi.org/10.3233/SW-170286. doi:10.3233/SW-170286.

[30] A. Bagga, B. Baldwin, Entity-based cross-document coreferencing using the vector space model, in: COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics, 1998. URL: https://aclanthology.org/C98-1012/.

[31] E. Amigó, J. Gonzalo, J. Artiles, F. Verdejo, A comparison of extrinsic clustering evaluation metrics based on formal constraints, Information Retrieval 12 (2009) 461–486. doi:10.1007/

s10791-008-9066-8.
[32] N. M. Triet, D. Van Thin, Trietnlp at ate-it: A hybrid pipeline for italian waste management terminology analysis, 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.
[33] O. Arab, G. M. Di Nunzio, Oate at ate-it: A hybrid approach for automatic terminology extraction in 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.
[34] S. Maule, G. M. Di Nunzio, Smte at ate-it: Ensemble term extraction with italian bert, spacy, and vocabulary-based filtering, 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.
[35] D. Smirnov, B. Khashechian, G. M. Di Nunzio, Dsbkte at ate-it: From token classification to zero-shot generation: Two approaches to italian ate at evalita 2026, 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.
[36] M. Bakhtiyarzadeh, H. Bayrami Asl Tekanlou, J. Razmara1, Peacemaker at ate-it: Automatic term extraction from italian text for waste management data using encoder model, 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.
[37] Z. Hasnain, F. Ahmed Siddque, Termninjas at ate-it: Overview of the term extraction and term variants clustering 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.
[38] M. Kim, G. M. Di Nunzio, Mkte at ate-it: Crf-based term extraction for italian waste management documents, 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.
[39] J. Tonneau, G. M. Di Nunzio, Jtte at ate-it: A crf model with contextual embeddings, 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.
[40] V. G. L. Vandervoort, G. M. Di Nunzio, Vvte at ate-it: From candidates to terms: Hybrid italian ate with dependency heuristics, gemini, and random forest filtering, 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.