Peacemaker at ATE-IT: Automatic term extraction from Italian text for waste management data using encoder model

Mahdi Bakhtiyarzadeh[1,*], Hadi Bayrami Asl Tekanlou[2,*] and Jafar Razmara[1,*]

[1]Department of Computer Science, University of Tabriz, 29 Bahman Boulevard, Tabriz 51666-16471, Iran
[2]Department of Computer Science, University of Tabriz, 29 Bahman Boulevard, Tabriz 51666-16471, Iran

Abstract

The development of automatic term extraction has become increasingly important in modern technology. Automatic term extraction can be found in virtually every search engine that is currently available to users. Recent advancements have provided promising results for the extraction of automatic terms; however, accurate labeling is difficult because of several factors, such as the limited number of annotated documents available for training and the complexity of extracting multi-word expressions due to shifts in the domain. In this paper, we will present a low-cost and interpretable method of automatic term extraction, developed specifically for Task A of the ATE Shared Task. This new method utilizes fine-tuning extraction strategies that can run on a small amount of computational resources. We evaluated our automated system using both type-level and micro-level measures of precision, recall, and F1-score to measure both complementary aspects of the extraction performance. According to the experimental results, our proposed approach achieves consistent and balanced performance compared to other teams. Even though the technique itself is relatively straightforward, it serves as a good starting point for low-resource models. Overall, the findings point toward the possibility of significant future advancements (in model expansion) with higher-level performance still able to retain their ability to be interpreted.

Keywords

Automatic Term Extraction, Large Language Models, Ontology Learning, Domain-Specific Terminology, Information Extraction

1. Introduction

Terminology is the beating heart of sentences, especially in tasks such as Automatic Term Extraction (ATE), where identifying domain-specific concepts from corpora is crucial for applications like machine translation, knowledge extraction, and semantic analysis [1, 2]. The "Automatic Term Extraction" (ATE) focuses on tasks such as corpus construction, unithood (measuring word co-occurrence), termhood (assessing domain relevance), and recognizing "variants" or how a term can be expressed differently from the original way [3]. Recent research utilizing large language models for few-shot ATE demonstrated considerable progress when operated under conditions where limited resources exist. Additionally [1], systematic reviews have documented ongoing development of methods for extracting multi-word terms from specific types of text through Transformer's architecture and ability to extract terms across multiple languages and domains [2, 4]. These advances play a significant role in the field of computer-assisted translation (CAT), where having bilingual terms available improves the quality of machine translations [5] as well as supporting the evaluation of term extraction software against gold standard datasets such as ACLRD-TEC2.0 and MAGMATic [6, 7]. Monolingual and Multilingual ATE [8], extending into specialized fields such as biomedical literature [9] and media bias analysis [10], is supported by datasets created from comparable corpora. It is common practice when improving ATE to

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.
m.bakhtiyarzadeh1403@ms.tabrizu.ac.ir (M. Bakhtiyarzadeh); h.bayrami1403@ms.tabrizu.ac.ir (H. B. A. Tekanlou); razmara@tabrizu.ac.ir (J. Razmara)
https://github.com/Mahdi8424 (M. Bakhtiyarzadeh); https://github.com/HadiBayrami/ (H. B. A. Tekanlou)
0009-0002-7206-6116 (H. B. A. Tekanlou); 0000-0002-6320-8517 (J. Razmara)

integrate terminological resources [11] with well defined robust evaluation frameworks [12]. However, challenges still exist regarding the collection of semi-automated data for Systematic Reviews [13].

This study makes the following contributions:

- **Robust Data Alignment and Preparation**: We implement a custom preprocessing pipeline featuring token-level semantic alignment (BIOtagging), specifically tailored for the Italian language and the ATE-IT task. This process accurately maps and converts multiword term instances from the raw gold standard into the correct BTERM, ITERM, O sequence labels, resolving token boundary ambiguities essential for effective Token Classification.
- **High-Fidelity Feature Extraction Architecture**: We utilize a high-performance, context-aware Transformer-based Token Classification architecture, leveraging the pre-trained dbmdz/bert-base-italian-cased model. This provides deep, contextual semantic representations optimized for token-level predictions within the Italian regulatory domain.
- **Environment Stability and Reproducibility**: We ensure the stability and reproducibility of the fine-tuning process by explicitly pinning critical Python dependencies (including transformers and accelerate) to verified, compatible versions. This guarantees API compatibility across varying runtime environments, mitigating common failure modes and allowing for reliable training execution.

2. Related Work

As presented in the previous section, Automatic Term Extraction (ATE) is critical for extracting domain-specific concepts that underpin the core tasks of natural language processing and knowledge management. The initial strategies that dealt with ATE were based on the combination of linguistic and statistical techniques to address issues such as unithood (the occurrence of words together) and termhood (the relevance of terms to domains). Examples of hybrid methods used on specialized corpora, such as in computer science and medicine, include the combination of statistical measures (mutual information and C-value/NC-value) with shallow linguistic criteria (part of speech) [14]. Various approaches have shown strong performance across domains, with similarity scores closely matching those for terms represented via traditional word-based representations in document similarity evaluations. However, the new methods will need to compensate for cases where multi-word terms have nesting problems [14]. Additionally, the comparative evaluations of term recognition systems indicate that combining several different approaches, such as using voting methods, resulted in better performance than using any one method alone when evaluated against corpus data from Wikipedia text, but performance varied from domain to domain and was not as high when using data from the Genia corpus in the life sciences [15]. Tools based on shallow grammars designed for use with noun-phrases in technical circles and combined with the application of clustering to accommodate variations in their structure (e.g., using variation of insertion structure) have generated very high levels of precision ranging from 93% to 98% when constructing an automated thesaurus [16]. PAT-tree approach also has been shown to enable the construction of adaptive keyword extraction systems, reducing dependency on rigid lexicons and word segmentation of documents written in Asian languages, such as Chinese, and facilitating their effective retrieval or classification when using these systems [17]. Multilingual and Monolingual ATE advancements take advantage of comparable corpora to develop data sets that can be used to support variant recognition and cross-lingual transfer. One particular contribution is the development of a data set specifically for both monolingual and multilingual ATE that allows terms to be extracted from comparable sources and resolve resource scarcity in low-resource languages [8, 16]. This reflects an interest in leveraging corpus-based semantic connections to create associations to MWEs by way of Information Extraction, and using these associations to compare the MWEs to thesauri within the Agriculture and Medicine domains [18]. For Agriculture, rule-based systems like RelExOnt have automated vocabulary extraction as well as relation discovery, providing ontologies with an 86.89% accuracy rate in the creation process [19]. The combination of statistical, linguistic, and hybrid extractors

has increased the accuracy of Spanish medical text extraction. It demonstrates the value of ensemble methods for different extraction methods [12]. Recent advancements in medical text extraction systems incorporating complete terminological resources and comprehensive evaluation methods allow for greater reliability of ATE systems (Automated Term Extraction). Incorporating external lexicons specific to the domain and other validated sources of information into automation models for extraction provide significant increases in recognizing multifaceted terms [18].

3. Task Description and Dataset

3.1. Task

This section describes the parameters for the ATE-IT Shared Task 2026 and its components, as well as the dataset used in the evaluation. It also describes the system architecture created for use in the evaluation; our contribution was in **Subtask A: Term Extraction**.

3.1.1. ATE-IT Shared Task 2026: Subtasks and Primary Focus

The ATE-IT Shared Task 2026 [20] has been organized around an evaluation of Automated Term Extraction (ATE) in a large scale way at the specific domain level of waste management in Italy. The challenge is divided into two distinct subtasks of increasing complexity:

1. **Subtask A: Term Extraction**. The main goal in the task of Term Extraction is to identify and extract individual domain-specific terms from the corpus sentences.
2. **Subtask B: Term Variants Clustering.** The primary objective of Term Variants Clustering is to group extracted terms that refer to the same underlying concept. This requires semantic and morphological analysis of terms so that synonymy and lexical variations can be handled.

In this study, we will be participating only in Subtask A: Term Extraction. Our main goal will be to process the sentences of the specialized corpus accurately and identify the municipal waste management domain-specific terms.

3.2. Corpus Description

Ate-it challenge organizers supplied the official textual source for the ATE-IT shared task, which contained textual examples related directly to the field of municipal waste management (an area of environmental concern). The challenge also provided guidelines regarding how to use the corpus dataset for this challenge, in order to develop models based on the provided training data and evaluate models against a pre-defined testing dataset in accordance with the established structure of the challenge. Tables 1 and 2 present an overview of the ATE-IT data statistics and representative examples from the dataset, respectively.

Table 1
ATE-IT data statistic overview.

Dataset	Samples
Train	2308
Development	577
Test	1142

Table 2
Representative examples from the dataset

sentence_text	term_list
CONSIDERATO che la situazione in cui si trova attualmente il Comune è riconducibile all'ipotesi contemplata nelle previsioni di cui al citato art. 191, in quanto sussistono gravi condizioni e fondate ragioni di tutela della salute pubblica e dell'ambiente, che risulterebbero inevitabilmente pregiudicate in caso di mancato ricorso temporaneo ad una speciale forma di gestione del servizio di raccolta per i rifiuti provenienti da luoghi adibiti ad uso di civile abitazione in cui alloggino persone risultate positive al Covid-19, in quarantena obbligatoria ai sensi dell'articolo 1 lettera e del D.P.C.M. 8 marzo 2020	"gestione del servizio di raccolta per i rifiuti", "servizio di raccolta per i rifiuti provenienti da luoghi adibiti ad uso di civile abitazione"
Al fine di consentire la raccolta dei rifiuti nei contenitori interni agli stabili, il proprietario singolo o i condomini, in solido fra loro, hanno l'obbligo di esporre gli stessi nei giorni e nelle ore stabiliti sul tratto viario prospiciente l'immobile di competenza e di riporli all'interno dei cortili o delle pertinenze condominiali, dopo l'avvenuto servizio di raccolta.	"esporre", "raccolta dei rifiuti", "servizio di raccolta"
La richiesta di attivazione del servizio deve essere presentata dall'utente al gestore dell'attivit00e0 di gestione tariffe e rapporto con gli utenti entro novanta (90) giorni solari dalla data di inizio del possesso o della detenzione dell'immobile, a mezzo posta, via e-mail o - 5 mediante sportello fisico e online di cui all'Articolo 19, compilando 2 l'apposito modulo scaricabile dalla home page del sito internet del 01 gestore in modalit00e0 anche stampabile, disponibile presso gli sportelli fisici, laddove presenti, ovvero compilabile online art.	"gestore", "tariffe", "utente"

4. Proposed Approach

The system that has been developed to complete ATE_IT Subtask A (extraction of Italian vocabulary) uses a Sequence Labeling Framework. The identification of terms has been changed into classifying every single word in a sentence. We used a fine-tuned state-of-the-art transformer model for the extraction of Vocabulary.

4.1. Data Preprocessing and Alignment

A critical component of the methodology is the custom preprocessing pipeline, which converts raw, sentence-level term annotations into a format consumable by the sequence labeling model.

- **BIO Tagging Scheme**: The standard B-I-O (Begin, Inside, Outside) tagging scheme is employed: `B-TERM` for the first token of a term, `I-TERM` for subsequent tokens, and `O` for non-term tokens.
- **Token-Term Alignment**: Term-to-token correspondence is ensured by using offset mappings provided by the tokenizer to create a mapping between character-level terms and the subword tokens from the model. To avoid double labels and overlap between terms that are themselves "overlapping," a mask is created for each character in the input sentence as it emerges from the model for each term. The mask makes sure that for every character position, it can only be mapped to one term, thus ensuring that a single character can't be allocated to two separate terms.
- **Label Encoding**: Symbolic BIO labels are converted into numerical IDs (O=0, B-TERM=1, I-TERM=2). Special tokens (e.g., [CLS], [SEP]) and padded tokens are masked with a label ID of **-100**, which is ignored by the loss function.

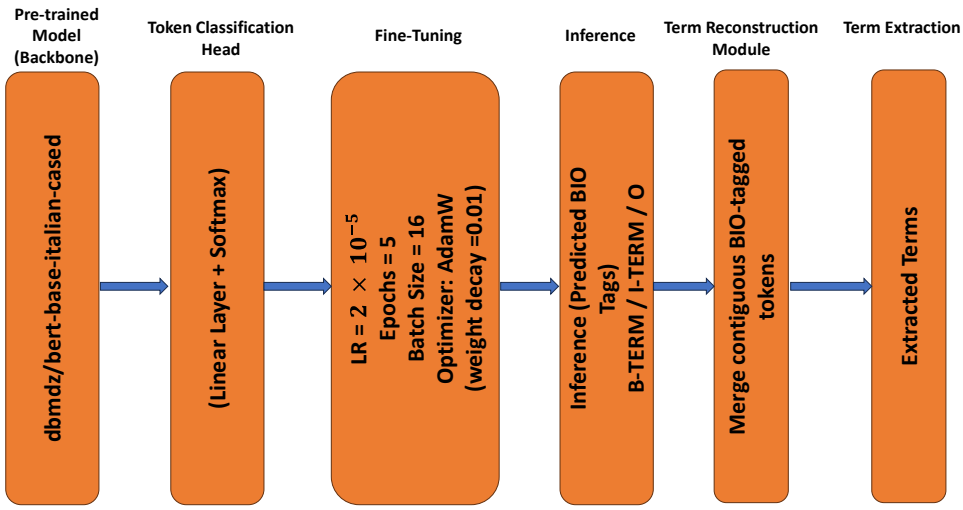

Figure 1: Overview of the system architecture used for token-level term extraction.

4.2. Modeling and Training

The core of the system is the *dbmdz/bert-base-italian-cased* model, a cased **BERT** (**Bidirectional Encoder Representations from Transformers**) variant pre-trained on an extensive Italian corpus. This model is adapted for **Token Classification** by appending a linear layer for transfer learning. The system architecture is illustrated in Figure 1.

5. Results

In this section, we describe the results obtained in the Automatic Term Extraction 2026 shared task. A total of 9 teams participated in the competition. The present paper presents results based on micro and type based evaluation metrics. Micro measures consider the total count of each type of term (e.g., phone numbers, names) that appear in the data set and provide a measure of accuracy of the entire extraction process. Type measures provide insight into how well the system is able to find different types of terms within the dataset.

5.1. Baselines

As a reference point, Table 3 presents the baseline results for both type-based and micro-based evaluation metrics. The following is our first baseline, which should be used to compare all participating systems. This baseline was created using the latest version (gemini-2.5-flash) of a closed-source large language model, in a zero-shot mode. The input text was processed in 20 instance batches according to the predefined prompt.

Table 3
Baseline model results using type-based and micro-based evaluation metrics

Type Precision	Type Recall	Type F1	Micro Precision	Micro Recall	Micro F1
0.435	0.508	0.469	0.497	0.559	0.526

5.2. Evaluation Metrics

The criteria used to evaluate the submitted work included a comparison between the automated term extraction and the corresponding gold standard annotations (as previously outlined). At the same time, type-level and micro-level metrics provide separate initial views of system performance. Type-level metrics provide an aggregate measure of how many unique domain terms have been identified by the system regardless of their frequency of occurrence; micro-level metrics provide detailed performance data on a sentence-by-sentence basis by accumulating all "True," "False Positive," and "False negative" results from all sentences to arrive at totals. Additionally, the precision, recall and F1 score calculations were derived from both type-level and micro-level evaluations allowing for assessment between extracted term correctness and completeness from both perspectives.

Type Metrics

- **TP = number of extracted unique terms that appear in the gold set**
- **FP = number of extracted unique terms that do not appear in the gold set**
- **FN = number of gold unique terms that were not extracteds**

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Micro Metrics [21]

- **TP$_s$ = number of terms extracted from sentence s that match the gold standard**
- **FP$_s$ = number of terms extracted from sentence s that do not match the gold standard**
- **FN$_s$ = number of gold standard terms in sentence s that were not extracted**

$$Precision_{\textbf{micro}} = \frac{\sum_{s \in D} TP_s}{\sum_{s \in D} (TP_s + FP_s)}$$

$$Recall_{\textbf{micro}} = \frac{\sum_{s \in D} TP_s}{\sum_{s \in D} (TP_s + FN_s)}$$

$$F1_{\textbf{micro}} = \frac{2 \cdot Precision_{\textbf{micro}} \cdot Recall_{\textbf{micro}}}{Precision_{\textbf{micro}} + Recall_{\textbf{micro}}}$$

5.3. Team Results

The outcomes from our application of the method we've described will now be discussed in detail. The performance of our system is assessed with the metrics shown in Section 5.2, and as for how our performance measures up against the results reported by each of the other teams that participated in this study, it is important to first compare these results so that we can accurately identify where the submitted systems rank with respect to one another and any associated strengths/weaknesses. For reference, Table 4 presents the detailed type-level and micro-level evaluation metrics achieved by the Peacemaker team, which are used as a baseline comparison in our analysis. In order to compare our model's performance against all evaluated dimensions, we provide a multi-dimensional view of our model's performance along with those of its competitors through the use of a radar plot (Figure 2), which allows us to quickly compare Precision, Recall and F1-score evaluations for the two evaluation types (type-level and micro-level) evaluated.

Table 4
Peacemaker team results using type-based and micro-based evaluation metrics

Type Precision	Type Recall	Type F1	Micro Precision	Micro Recall	Micro F1
0.430	0.455	0.442	0.497	0.476	0.486

6. Discussion

The results of the proposed method provide valid insight into how lightweight and low-engineering techniques behave in the domain of ontology term extraction even though it is not producing highest performance level. Evidence to support this claim is captured in the radar plot, which displays balanced precision and recall sections between the type-level and micro-level metrics. Therefore, this suggests that the proposed method produces stable extraction as opposed to maximally optimizing precision scores through repetitive extraction cycles.

While our experimental results show that the proposed method ranks lower than others, we have observed that for both the Type and Micro levels the Precision and Recall values are similar, indicating that this method does not rely heavily on just using high-frequency terms. We conclude that our proposed method has consistent extraction performance for both frequent and less frequent terms from many different evaluation granularities. Thus, our proposed method provides a strong baseline for future researchers in this area with a reasonable level of computational and implementation complexity.

It should be noted that our system was created using limited resources for both computation and modeling, which prevented us from being able to take full advantage of many of the larger or more

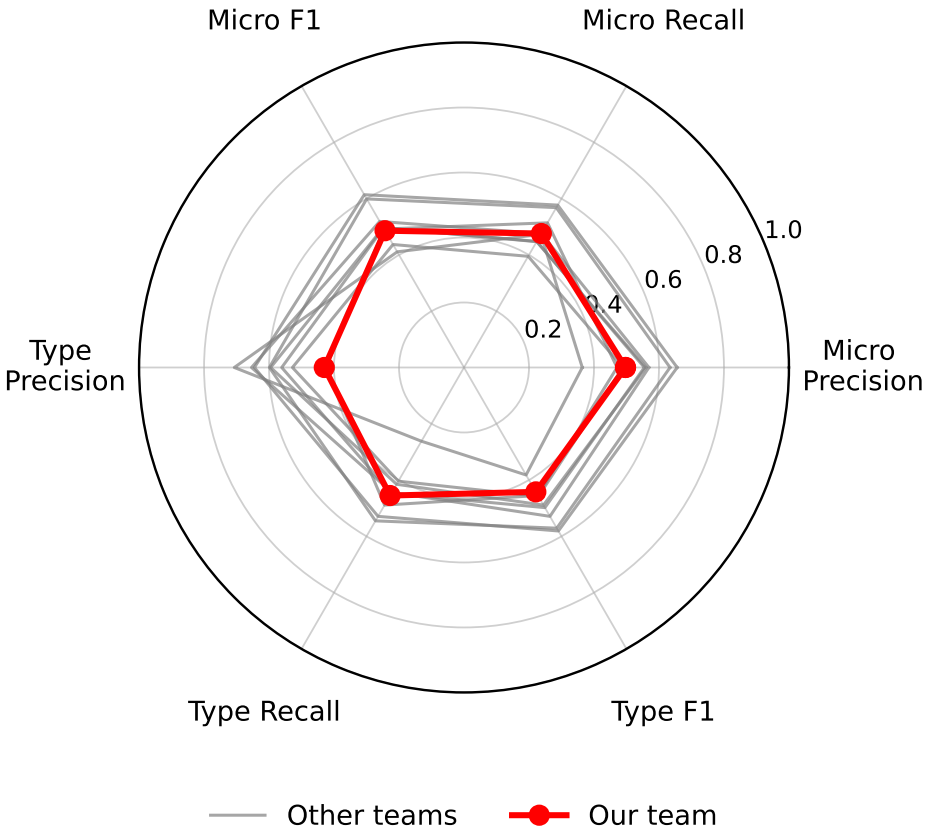

Figure 2: Radar plot comparing the performance of our approach with other participating systems across type-level and micro-level Precision, Recall, and F1-score.

specialized models that are available to some teams that have a higher ranking. If the teams were able to use better models and were able to increase their computational capabilities, it is likely that their overall performance—especially Recall—would also improve greatly. However, the current results show that the proposed method is effective and competitive and has the advantage of being lightweight.

7. Conclusions

In this research, we proposed an easy-to-understand, lightweight process for extracting ontology terms using data captured during the shared task evaluation phase and applying type-level and micro-level metrics to evaluate the performance of our solution. Our reduced resources and choice of more constrained models did not hinder our application from demonstrating steady and well-balanced extraction results across all models and evaluations, which implies that the extraction capabilities of our modeling process are not based solely on selecting terms that appear very frequently. The fact that our team came in 7th place out of the 9 teams that competed shows how well we stack up against other teams, even though we had far fewer resources than most of the other teams that took part in the competition; therefore, our system remains highly competitive based on what we have demonstrated.

Declaration on Generative AI

Generative AI tools played a supplemental and shallow role in our work. We used generative AI for brainstorming and developing ideas on how to tackle the Automatic Term Extraction (ATE) task; they were also helpful for enhancing the clarity and organization of notes and drafts. However, we did not rely on generative AI to create any part of the dataset, including the annotations, experimental results, or final models. All authors made the decisions about method choice, implementation, and analysis. After using these tools/services, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] 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.

[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. URL: https://doi.org/10.1093/llc/fqad030. doi:10.1093/llc/fqad030.

[3] K. Heylen, D. D. Hertog, Automatic term extraction, 2014. URL: https://api.semanticscholar.org/CorpusID:184348418.

[4] 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. URL: https://aclanthology.org/2021.findings-acl.316/. doi:10.18653/v1/2021.findings-acl.316.

[5] 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, Vancouver, Canada, 2014, pp. 54–68. URL: https://aclanthology.org/2014.amta-researchers.5/.

[6] 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), Portorož, Slovenia, 2016, pp. 1862–1868. URL: https://aclanthology.org/L16-1294/.

[7] 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, Dublin, Ireland, 2019, pp. 78–86. URL: https://aclanthology.org/W19-6608/.

[8] 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. URL: https://doi.org/10.1007/s10579-019-09453-9. doi:10.1007/s10579-019-09453-9.

[9] P. Tiwari, S. Uprety, S. Dehdashti, M. S. Hossain, TermInformer: unsupervised term mining and analysis in biomedical literature, Neural Comput. Appl. 37 (2020) 1–14.

[10] T. Spinde, L. Lin, S. Hinterreiter, I. Echizen, Leveraging large language models for automated definition extraction with taxomatic a case study on media bias, 2025. URL: https://arxiv.org/abs/2504.00343. arXiv:2504.00343.

[11] S. Aubin, T. Hamon, Improving term extraction with terminological resources, 2006. URL: https://arxiv.org/abs/cs/0609019. arXiv:cs/0609019.

[12] J. Vivaldi, H. Rodríguez, Evaluation of terms and term extraction systems: A practical approach, Terminology 13 (2007) 225–248. doi:10.1075/term.13.2.06viv.

[13] L. Schmidt, A. N. Finnerty Mutlu, R. Elmore, B. K. Olorisade, J. Thomas, J. P. T. Higgins, Data extraction methods for systematic review (semi)automation: Update of a living systematic review, F1000Res. 10 (2021) 401.

[14] E. E. Milios, Y. Zhang, B. He, L. Dong, Automatic term extraction and document similarity in special text corpora, 2003. URL: https://api.semanticscholar.org/CorpusID:3128871.

[15] Z. Zhang, J. Iria, C. Brewster, F. Ciravegna, A comparative evaluation of term recognition algorithms, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, D. Tapias (Eds.), Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008. URL: https://aclanthology.org/L08-1281/.

[16] D. Bourigault, C. Jacquemin, Term extraction + term clustering: an integrated platform for computer-aided terminology, in: Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics, EACL '99, Association for Computational Linguistics, USA, 1999, p. 15–22. URL: https://doi.org/10.3115/977035.977039. doi:10.3115/977035.977039.

[17] L.-F. Chien, Pat-tree-based keyword extraction for chinese information retrieval, SIGIR Forum 31 (1997) 50–58. URL: https://doi.org/10.1145/278459.258534. doi:10.1145/278459.258534.

[18] E. Morin, C. Jacquemin, Projecting corpus-based semantic links on a thesaurus, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, ACL '99, Association for Computational Linguistics, USA, 1999, p. 389–396. URL: https://doi.org/10.3115/1034678.1034739. doi:10.3115/1034678.1034739.

[19] N. Kaushik, N. Chatterjee, Automatic relationship extraction from agricultural text for ontology construction, Information Processing in Agriculture 5 (2018) 60–73. URL: https://www.sciencedirect.com/science/article/pii/S2214317317300227. doi:https://doi.org/10.1016/j.inpa.2017.11.003.

[20] 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.

[21] M. Röder, R. Usbeck, A.-C. N. Ngomo, Gerbil – benchmarking named entity recognition and linking consistently, Semantic Web 9 (2018) 605–625. URL: https://journals.sagepub.com/doi/abs/10.3233/SW-170286. doi:10.3233/SW-170286. arXiv:https://journals.sagepub.com/doi/pdf/10.3233/SW-170286.