Benchmarking Hindi Term Extraction in Education: A Dataset and Analysis

Shubhanker Banerjee[‡], Bharathi Raja Chakravarthi, John P. McCrae[‡],
[‡]Research Ireland ADAPT Centre
University of Galway
Ireland
shubhanker.banerjee@adaptcentre.ie

Abstract

This paper introduces the HTEC Hindi Term Extraction Dataset 2.0, a resource designed to support terminology extraction and classification tasks within the education domain. HTEC 2.0 has been developed with the objective of providing a high-quality benchmark dataset for the evaluation of term recognition and classification methodologies in Hindi educational discourse. The dataset consists of 97 documents sourced from Hindi Wikipedia, covering a diverse range of topics relevant to the education sector. Within these documents, 1,702 terms have been manually annotated where each term is defined as a single-word or multi-word expression that conveys a domain-specific meaning. The annotated terms in HTEC 2.0 are systematically categorized into seven distinct classes. Furthermore, this paper outlines the development of annotation guidelines, detailing the criteria used to determine term boundaries and category assignments. By offering a structured dataset with clearly defined term classifications, HTEC 2.0 serves as a valuable resource for researchers working on terminology extraction, domain-specific named entity recognition, and text classification in Hindi. We release the dataset publicly for the research community[1].

1 Introduction

Terminology extraction techniques are essential in various computational applications that involve processing domain-specific language. These techniques focus on identifying and extracting specialized lexical units from text, which can be useful for structuring information (Leonardi et al., 2009; Wozniak-Kasperek, 2014), improving knowledge organization (Golub et al., 2014), and supporting automated text analysis (Musacchio et al., 2001). The extracted terms serve as key components in various natural language processing tasks, including text classification (Liu and Chen, 2019), information retrieval (Zeng et al., 2002), and domain-specific knowledge modeling (Agt and Kutsche, 2013).

A term is defined as a lexical unit that conveys a precise meaning within a specific field (Cabré, 2012). Various approaches have been proposed for extracting terms, differing in methodology, scope, and intended application. Count-based methods such as TF-IDF (Salton and Buckley, 1988) and CValue (Lossio-Ventura et al., 2013) have traditionally been used to extract terms. Although these methods are computationally efficient, they have been outperformed by data-driven term extraction techniques. Particularly, deep learning based methods based on language models have established state-of-the-art benchmarks on this task (Rigouts Terryn et al., 2022; Lang et al., 2021).

The development of term extraction systems for low-resource languages has remained an open challenge due to the lack of high-quality annotated datasets and standardized evaluation frameworks. In this paper, we introduce a dataset specifically designed for term extraction in Hindi, aiming to address this gap. Additionally, with the increasing availability of synthetic data generated using generative language models, this dataset can also serve as a gold standard for evaluating term extraction systems. As discussed by QasemiZadeh and Schumann (2016) evaluation frameworks for term extraction typically consist of two essential components. The first component is a gold-standard dataset, which is a collection of manually annotated texts that serve as a benchmark for comparison. The second component involves performance metrics such as precision, recall, and the F1-score. These metrics allow for a systematic assessment by comparing the outputs of extraction methods against the annotations in the gold standard. By providing a reliable benchmark, this resource facilitates the development and assessment of extraction methodologies tailored for Hindi and other low-resource languages.

[1] https://tinyurl.com/6jcr5umc

To support this goal, the dataset has been carefully curated with enhanced annotation quality. Two annotators were engaged during the initial rounds, allowing for iterative refinement of the guidelines to improve consistency and reliability before proceeding with the final annotation process. Additionally, terms are classified into fine-grained semantic categories, enabling detailed analysis and supporting a range of terminology extraction and classification tasks. Furthermore, detailed annotation guidelines were developed to standardize the annotation process. These guidelines evolved over multiple annotation rounds, incorporating feedback and refinements to enhance clarity and consistency. This iterative approach ensured that the annotated terms adhered to a well-defined framework, reducing subjectivity and improving overall dataset quality.

The structure of this paper is as follows: Section 2 discusses related work on term extraction and the development of term-annotated datasets. Section 3 presents the dataset statistics and details the process of creating the annotation guidelines as well as the inter-annotator agreement. Section 4 discusses the Experimental setup and the experiments. Section 5 discussed the results. Finally, the paper concludes in Section 6.

2 Related Work

2.1 Term Annotated Datasets

2.1.1 Monolingual

Several term-annotated datasets have been developed to support terminology extraction across different domains. In the biomedical domain, the Colorado Richly Annotated Full Text Corpus (CRAFT) (Bada et al., 2012) and the GENIA corpus (Kim et al., 2003) provide extensive term annotations, while the Gene Ontology (GO) corpus (DBL, 2004) structures biological terminology into three sub-ontologies.

For computational linguistics, the ACL RD-TEC dataset, built from the ACL Anthology Reference Corpus, consists of two versions: ACL RD-TEC v1.0 (QasemiZadeh and Handschuh, 2014), which contains 82,000 annotated terms, and ACL RD-TEC v2.0 (QasemiZadeh and Schumann, 2016), which annotates 300 abstracts. Other domain-specific resources include the JPED corpus for pediatric texts (Coulthard et al., 2005), the ECO corpus for ecology (Zavaglia et al., 2005), and the N&N corpus for nanoscience (Coleti et al., 2009).

Efforts in low-resource languages have also contributed to terminology extraction. The RSDO5 corpus[2] provides Slovenian term annotations, while an Irish Wikipedia dataset (McCrae and Doyle, 2019) contains 864 manually annotated terms. The CoastTerm Dataset (Delaunay et al., 2024) offers over 12,000 annotated terms in coastal sciences.

In the context of Hindi terminology extraction, the Hindi Term Extraction in Education Corpus (HTEC 1.0) (Banerjee et al., 2022) was introduced as a manually annotated resource for terminology extraction. The dataset was constructed using Hindi Wikipedia's API, retrieving 71 documents (11,960 words) from pages categorized under शिक्षा (shiksha, "education"). Terms were annotated following the surface representation of concepts approach (Pazienza, 1998), with no syntactic constraints to ensure broad coverage. Given the subjective nature of term identification, annotation relied on the annotators' judgment. However, the first dataset release (HTEC 1.0) was annotated by a single annotator, which posed challenges in terms of annotation consistency and reliability.

Building upon HTEC 1.0, our new release addresses these limitations by introducing multi-annotator agreement, refined annotation guidelines, and fine-grained semantic term classification. This extension enhances both the dataset's quality and its applicability to a wider range of terminology extraction and classification tasks.

2.1.2 Multilingual

Multilingual term-annotated datasets facilitate cross-linguistic terminology extraction. The ACTER dataset (Rigouts Terryn et al., 2020) provides English, French, and Dutch corpora across four domains. The TTC project (Daille, 2012) supports Wind Energy and Mobile Technology term extraction in seven languages. Other multilingual resources include the KAS-biterm dataset (Ljubešić et al., 2018) for Slovene academic writing, BitterCorpus (Arcan et al., 2014), an English-Italian IT domain corpus, and TermFrame v1.0 (Pollak et al., 2019), which focuses on karstology in Slovene, Croatian, and English.

These datasets establish benchmarks for term extraction across languages, emphasizing support for low-resource languages through annotated corpora.

[2] https://www.clarin.si/repository/xmlui/handle/11356/1400

2.2 Automatic Term Extraction

2.2.1 Unsupervised Term Extraction

Unsupervised Automatic Term Extraction (UATE) methods extract domain-specific terms without requiring annotated corpora. Frequency-based methods such as TF-IDF (Salton and Buckley, 1988) and CValue (Lossio-Ventura et al., 2013) prioritize terms based on statistical occurrence patterns, while reference corpus-based methods like domain pertinence (Meijer et al., 2014) contrast domain-specificity against general corpora. More advanced techniques integrate semantic information, such as Normalized Pointwise Mutual Information (NPMI) (Bordea et al., 2013), topic modeling (Nugumanova et al., 2022), and graph-based ranking (Zhang et al., 2018). Despite their scalability, these methods struggle with ambiguity and domain adaptation.

2.2.2 Supervised Term Extraction

Supervised ATE methods leverage labeled datasets and machine learning models for term classification. Traditional approaches use linguistic and statistical features with classifiers such as SVMs (Ljubešić et al., 2018) and random forests (Yuan et al., 2017). More recent deep learning methods employ embeddings like Word2Vec (Mikolov et al., 2013) and BERT (Rokas et al., 2020) for improved contextual representation. End-to-end neural architectures, including BiLSTM-CRF (Rokas et al., 2020) and XLM-R (Lang et al., 2021), achieve state-of-the-art performance. However, supervised methods require large annotated corpora, making them less practical for low-resource languages.

Recent systematic reviews confirm that while supervised approaches significantly outperform unsupervised methods, even state-of-the-art systems rarely exceed 60% F1-score on benchmark datasets (Di Nunzio et al., 2023).

3 Dataset

This section outlines the annotation guidelines established to ensure consistency in the annotation process and provides an overview of the dataset statistics.

3.1 Data Collection

The dataset was collected from Hindi Wikipedia[3] by extracting an initial pool of 186 pages categorized under relevant educational topics. The search parameters included the categories शिक्षा (translation: Education), शैक्षणिक संस्थान (translation: Educational Institution), शिक्षण (translation: Teaching), and शिक्षक (translation: Educator), ensuring coverage of terminology related to education. From this corpus, 67 pages (36.0%) were removed due to duplication, 33 pages (17.7%) were excluded for containing fewer than 100 words, and 21 pages (11.3%) were eliminated due to excessive Latin characters (>15% of content). The remaining 65 articles underwent segmentation due to their length, resulting in the final 97 documents selected for annotation. This process prioritized comprehensive educational content with domain-specific terminology. The dataset statistics have been illustrated in Table 1.

3.1.1 Dataset Structure and Format

The dataset is provided as a collection of documents, with each document accompanied by a JSON file containing detailed annotations of extracted terms and their corresponding semantic categories.

Each JSON annotation follows a hierarchical structure and consists of the following components:

- **Document ID**: A unique identifier for each document.
- **Annotated Terms**: A list of terms extracted from the document.
- **Category Labels**: The predefined semantic category assigned to each term.
- **Term Position**: The start and end character positions of the annotated term within the document.

The example below illustrates the JSON annotation format with Hindi terms:

[3] https://hi.wikipedia.org/wiki

Category	Unique Terms Count
Ambiguous	409
Educational Institutions, Governing bodies, Think Tanks and Research Institutes	508
Degrees, Disciplines and different stages of education	274
Educationists, Learners and Researchers	253
Education Related Policy and Regulatory Frameworks	133
Mode of Dissemination	106
Education Technology and Equipment	19
Total Terms	1702

Table 1: Count of Unique Terms in Each Category


```
{
"document_id": "doc_001",
"terms": [
{
"term": "शिक्षा नीति",
"category": "Education-Related Policy and Regulatory Frameworks",
"start": 35,
"end": 45
},
{
"term": "शिक्षण संस्थान",
"category": "Educational Institutions, Governing Bodies, Think Tanks, and Research Institutes",
"start": 92,
"end": 108
},
{
"term": "ऑनलाइन शिक्षा",
"category": "Mode of Dissemination",
"start": 150,
"end": 165
}
]
}
```


This structured annotation format enables straightforward integration into various NLP frameworks for tasks such as supervised and unsupervised term extraction, named entity recognition, and domain adaptation. By providing precise term boundaries and categorization, the dataset supports both rule-based and machine learning-based approaches for automatic term extraction.

3.2 Annotation Guidelines

The annotation guidelines were developed based on insights from HTEC 1.0, acknowledging ISO 5078:2025(en)[4] terminology principles that differentiate between "candidate terms" and "validated terms." While the ISO standard prescribes a sequential approach where candidate terms undergo a discrete validation phase, our methodology adopted a more integrated, iterative refinement process due to project-specific constraints. Rather than separating initial identification from formal validation, we implemented a progressive improvement cycle where terms underwent concurrent identification and validation across multiple annotation rounds, effectively addressing ISO objectives through alternative means. This approach maintained classification quality while accommodating practical resource limitations inherent in specialized linguistic annotation projects.

The annotation process spanned four rounds with two annotators (PhD and Masters students in NLP with prior experience in lexical annotation tasks) independently annotating 10 documents per round. Before commencing, annotators underwent a two-day training on educational terminology and domain concepts. Their annotations were compared using the Highlight Tool[5], a Google Docs add-on that visualized discrepancies. Disagreements were resolved through moderated consensus meetings. Key revisions included: (1) adding explicit criteria for minimum term length requirements, (2) refining category definitions with boundary cases, and (3) developing decision trees for handling terms with multiple potential classifications. The termination criterion was a Jaccard Index exceeding 65%, balancing annotation quality with budgetary constraints. This refinement continued until Round 4, where agreement reached 66.2%, indicating sufficient consistency for reliable annotation.

3.2.1 Inter-annotator Agreement

To evaluate the consistency of the annotation process, inter-annotator agreement was measured using the Jaccard Index (Jaccard, 1901), a widely used metric for assessing set similarity. The agreement calculation considered both the overlap in anno-

[4] https://www.iso.org/standard/81917.html

[5] https://jsonchin.github.io/highlight_tool/

tated terms and the semantic categories assigned to them, ensuring a comprehensive evaluation of annotation consistency. The Jaccard Index for two sets of annotated terms, A_1 and A_2, is defined as:

$$J(A_1, A_2) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} \quad (1)$$

where $|A_1 \cap A_2|$ represents common annotations between annotators, and $|A_1 \cup A_2|$ represents total unique annotated terms.

To compute overall inter-annotator agreement, the Jaccard Index was averaged across all annotated documents:

$$IAA = \frac{1}{N} \sum_{i=1}^{N} J(A_1^{(i)}, A_2^{(i)}) \quad (2)$$

where $J(A_1^{(i)}, A_2^{(i)})$ is the Jaccard similarity score for the i^{th} document.

While the annotators lacked formal education backgrounds, they acquired domain knowledge through studying educational terminology resources. After achieving satisfactory agreement, the main annotation task was completed by a single annotator with regular validation checks. The inter-annotator agreement scores, presented in Table 2, indicate progressive improvement, reflecting increasing consistency. The process of iterative refinement in each annotation round has been discussed in Appendix A.

3.2.2 Fine-grained Term Classification

Each annotated term is assigned to a predefined semantic category to maintain consistency and support structured analysis. The classification scheme covers key aspects of the education domain, including individuals, institutions, policies, technology, and knowledge dissemination.

Educationists, Administrators, Learners, and Researchers encompasses teachers, students, professors, and education officials, including deans and policymakers. Example terms in this category include *principal, teacher,* and *researcher*.

Education-Related Policy and Regulatory Frameworks covers government policies and regulations governing education, such as *National Education Policy (2022)* and *Education for All Scheme*. General terms like *education policy* and *exam system* are included here.

Educational Institutions, Governing Bodies, Think Tanks, and Research Institutes consists of institutions involved in education and policymaking, including *schools, colleges, universities, and research organizations*. Examples include the *Ministry of Education* and *University of Amsterdam*.

Education Technology and Equipment includes digital platforms and hardware used in education, such as *Blackboard, Piazza, student information systems*, and classroom tools like *chalk, writing boards*.

Mode of Dissemination refers to teaching methods and educational resources, covering *video lectures, tutorials, books, research papers*, and other instructional materials.

Degrees, Disciplines, and Stages of Education consists of academic subjects (*physics, mathematics*), degrees (*Bachelor of Technology, Bachelor of Education*), and education levels (*primary, secondary, and higher education*).

Ambiguous Terms are those that do not fit any specific category or belong to multiple categories.

To maintain consistency, annotators use a color-coded system to distinguish different semantic classes.

3.2.3 Why Ambiguous Terms?

The inclusion of an ambiguous category remains essential despite predefined categories. Terms often exhibit context-dependent meanings or interdisciplinary overlap, complicating classification. For example, कोर्स (course) shows true domain ambiguity across education, culinary contexts, and navigation. Some cases represent polysemy rather than ambiguity—like नामांकन संख्या (enrollment number) referring to both student registration and administrative processes. We chose this category over separate terms or multi-label annotation to ensure consistency and simplify evaluation. This approach prevents subjective decisions that could introduce errors, as uncertain terms are marked for expert review rather than forced into inappropriate categories. Ambiguous terms also enhance model robustness by exposing multiple meanings and context-dependent variations, enabling machine learning models to learn real-world usage patterns while allowing for future refinement as classification standards evolve.

3.2.4 Term Length Distribution Across Categories

To analyze the structural characteristics of annotated terms, we examined the length of terms across different semantic categories. The term length is

Iteration	Inter-Annotator Agreement (IA)
Round 1	25.5
Round 2	19.1
Round 3	41.3
Round 4	66.2

Table 2: Inter-Annotator Agreement (IA) measured using the Jaccard Index across four annotation rounds. Two annotators independently annotated a set of 10 documents per round.

defined by the number of words forming a single annotated term. Figure 1 provides a summary of the percentage distribution of term lengths across the predefined categories.

The distribution of term lengths across categories reveals notable variations in the structural composition of domain-specific terminology. As shown in Figure 1, Education Technology and Equipment has the highest proportion of single-word terms (60%), followed by Educationists, Administrators, Learners, and Researchers (41%) and Mode of Dissemination (39.5%). In contrast, Education-Related Policy and Regulatory Frameworks has the lowest percentage of single-word terms (16.3%), indicating that policy terminology rarely takes the form of individual words.

Two-word terms are most prevalent in Degrees, Disciplines, and Stages of Education (48.6%), while constituting only 17% of Education-Related Policy terminology. This suggests that academic disciplines and educational stages are frequently characterized by concise, two-word descriptors.

Longer terms (3+ words) dominate the Education-Related Policy and Regulatory Frameworks category (66.7%) and Educational Institutions category (50.9%), reflecting the complex and descriptive nature of policy frameworks and institutional designations. Conversely, Education Technology and Equipment has the lowest proportion of longer terms (12%), indicating a preference for concise, well-established terminology in this category.

These percentage distributions highlight significant structural variations across semantic categories, emphasizing the need for classification strategies that account for these inherent differences in term length. Categories dominated by longer, multi-word terms (such as Policy and Institutions) present different challenges for terminology extraction and classification compared to categories with predominantly shorter terms (such as Technology and Educational Roles).

For detailed annotation guidelines, readers may refer to Annotation Guidelines (Anonymous, 2024).

4 Experimental Setup

To establish benchmark performance on this dataset, we conducted experiments using both unsupervised and supervised term extraction methods. The objective of these experiments is to evaluate the effectiveness of various methodologies in extracting domain-specific terms and to provide a baseline for future research.

4.1 Unsupervised Term Extraction

For unsupervised term extraction, we implemented four widely used methods: Basic, ComboBasic, CValue, and non-negative matrix factorization term extraction. We utilized the TermXtract library[6] to perform experiments with these unsupervised approaches.

- Basic (Bordea et al., 2013): A frequency-based approach that identifies multi-word term candidates using substring occurrence patterns.
- ComboBasic (Astrakhantsev, 2015): An extension of Basic that introduces parameters to adjust term specificity, refining term selection.
- CValue (Lossio-Ventura et al., 2013): A statistical method that enhances multi-word term extraction by penalizing nested term occurrences.
- NMF-based Term Extraction (Nugumanova et al., 2022): A topic modeling approach that applies Non-negative Matrix Factorization (NMF) to extract domain-specific terms by identifying high-weighted words in topic-term distributions.

Each of these methods was evaluated in an unsupervised setting to establish baseline performance on the dataset.

[6] https://github.com/TeangaNLP/TermXtract

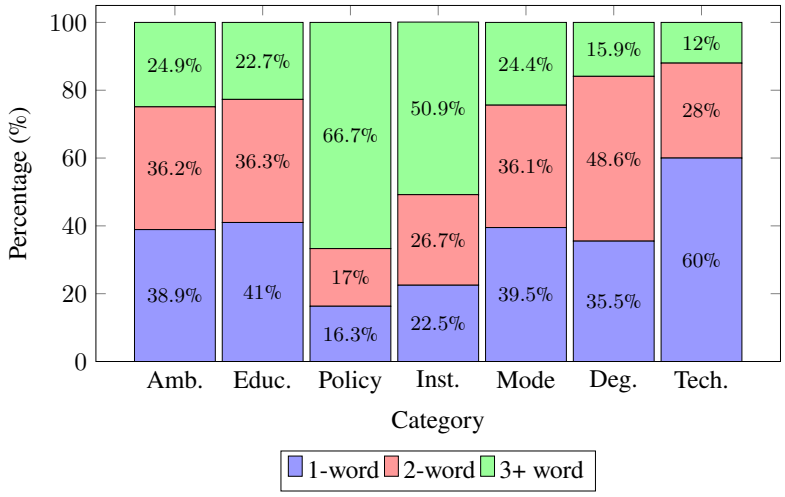

Figure 1: Term length distribution across categories (Amb.=Ambiguous, Educ.=Educationists, Policy=Education-Related Policy, Inst.=Educational Institutions, Mode=Mode of Dissemination, Deg.=Degrees and Disciplines, Tech.=Education Technology).

Unsupervised Methods	
Method	**P / R / F1**
ComboBasic	0.12 / 0.04 / 0.06
Basic	0.15 / 0.06 / 0.09
NMFExtractor	0.08 / 0.02 / 0.03
CValue	0.14 / 0.05 / 0.07
Supervised Methods	
XLM-RoBERTa (Token Classifier)	72.1 / 67.8 / 69.8
XLM-RoBERTa (Sequence Classifier)	47.3 / 43.5 / 45.3
mBART (NMT-based ATE)	58.9 / 52.1 / 55.3

Table 3: Performance comparison of different term extraction methods on the dataset, categorized into unsupervised and supervised approaches. Each cell in the second column reports Precision (P), Recall (R), and F1-score (F1) in that order.

4.2 Supervised Term Extraction

For supervised term extraction, we conducted experiments using three transformer-based approaches following the methodology proposed by Lang et al. (2021): (1) a token classifier, (2) a sequence classifier, and (3) a Neural Machine Translation (NMT)-based approach. Each method was implemented and evaluated using the Hugging Face Transformers[7] library.

We utilized XLM-RoBERTa (XLM-R) (Conneau et al., 2020), a state-of-the-art multilingual transformer model, due to its strong generalization capabilities and effectiveness in domain adaptation (Lang et al., 2021; Hazem et al., 2022). The dataset was split into training (70%), validation (10%), and testing (20%) sets to ensure a balanced evaluation.

- **Token Classifier**: A NER-style model that classifies each token as part of a term or not, achieving state-of-the-art results in ATE (Lang et al., 2021).

- **Sequence Classifier**: An n-gram-based model that classifies term candidates, serving as a strong comparative baseline.

- **NMT-based ATE**: An mBART-based (Liu et al., 2020) model that transforms sentences into comma-separated term sequences, excelling in multi-word term extraction .

5 Results

The results demonstrated in Table 3 emphasize the substantial performance gap between supervised and unsupervised methods in Automated Term Extraction (ATE). Though unsurprising, this underscores the necessity of annotated datasets for improving term extraction accuracy.

[7] https://huggingface.co/

5.1 Unsupervised Methods

The unsupervised approaches namely ComboBasic, Basic, NMFExtractor, and CValue demonstrate consistently poor performance. The highest F1-score among them (0.09 for Basic) is an order of magnitude lower than that of supervised models. This discrepancy underscores the inherent limitations of rule-based and statistical heuristics in capturing nuanced term structures.

A key observation is the trade-off between precision and recall. Precision remains relatively low across all unsupervised methods, suggesting a tendency to misclassify non-terms as terms, while recall is even lower, reflecting the failure to capture many valid terms. Notably, NMFExtractor performs the worst (F1 = 0.03), indicating that matrix factorization-based approaches fail to discern term boundaries effectively. This is likely due to their reliance on latent topic distributions, which may not align with term granularity.

More fundamentally, these methods lack the ability to account for semantic context. They rely heavily on frequency-based patterns, statistical co-occurrence, or fixed linguistic rules, making them brittle and domain-dependent. As a result, their applicability to real-world datasets is extremely limited, particularly for specialized terminology that does not conform to simple statistical regularities.

5.2 Supervised Methods

In contrast, the supervised models XLM-RoBERTa (Token Classifier and Sequence Classifier) and mBART demonstrate better performance, leveraging deep learning's capacity for contextual understanding. The best-performing method, XLM-RoBERTa (Token Classifier), achieves an F1-score of 69.8, with balanced precision (72.1) and recall (67.8), indicating strong generalization.

A particularly striking observation is the difference in performance between token classification and sequence classification. The sequence classifier model achieves an F1-score of 45.3 far lower than its token classification counterpart. This suggests that the n-gram-based sequence classification approach struggles to delineate term boundaries effectively. Unlike token classification, which identifies terms at the individual token level, sequence-level classification processes entire text spans at once. This can lead to errors, especially when terms are embedded within longer sequences, making it harder to precisely delineate term boundaries.

mBART (NMT-based ATE) achieves a moderate F1-score (55.3), performing better than sequence classification but worse than token classification. This suggests that sequence-to-sequence models can be effective for term extraction but still struggle with precise boundary detection. The relatively lower recall (52.1) suggests that mBART may be omitting relevant terms, possibly due to its reliance on translation-style decoding rather than direct classification.

6 Conclusion

We introduced HTEC 2.0, a Hindi Term Extraction dataset for education, supporting term extraction and classification. It features annotated terms with improved consistency and a category for ambiguous cases. Evaluations show XLM-RoBERTa outperforms statistical methods, demonstrating the need for context-aware models. Results highlight limitations of unsupervised approaches in low-resource languages like Hindi.

7 Acknowledgement

Author Shubhanker Banerjee was supported by Research Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT Centre at University Of Galway.
Author John McCrae was supported by Research Ireland under Grant Agreement No. 13/RC/2106_P2 at the SFI ADAPT Centre at University Of Galway.

References

2004. Gene ontology consortium: The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.*, 32(Database-Issue):258–261.

Henning Agt and Ralf-Detlef Kutsche. 2013. Automated construction of a large semantic network of related terms for domain-specific modeling. In *Advanced Information Systems Engineering: 25th International Conference, CAiSE 2013, Valencia, Spain, June 17-21, 2013. Proceedings 25*, pages 610–625. Springer.

Anonymous. 2024. Annotation guidelines. Available at https://shorturl.at/Ao18j.

Mihael Arcan, Marco Turchi, Sara Topelli, and Paul Buitelaar. 2014. Enhancing statistical machine translation with bilingual terminology in a cat environment. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 54–68.

Nikolay Astrakhantsev. 2015. *Methods and Software for Terminology Extraction from Domain-Specific Text Collection*. Ph.d. thesis, Institute for System Programming of Russian Academy of Sciences.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner Jr., K. Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC Bioinform.*, 13:161.

Shubhanker Banerjee, Bharathi Raja Chakravarthi, and John Philip McCrae. 2022. A dataset for term extraction in Hindi. In *Proceedings of the Workshop on Terminology in the 21st century: many faces, many places*, pages 19–25, Marseille, France. European Language Resources Association.

Georgeta Bordea, Paul Buitelaar, and Tamara Polajnar. 2013. Domain-independent term extraction through domain modelling. In *TIA 2013-10th International Conference on Terminology and Artificial Intelligence*.

Teresa Cabré. 2012. Terminology and translation. In *Handbook of Translation Studies: Volume 1*, pages 356–365. John Benjamins Publishing Company.

J. S. Coleti, D. F. Mattos, and G. M. B. Almeida. 2009. Primeiro dicionário de nanociência e nanotecnologia em língua portuguesa. In *II Encontro Acadêmico de Letras (EALE)*, pages 1–10.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Robert James Coulthard et al. 2005. *The Application of Corpus Methodology to Translation: The JPED Parallel Corpus and the Pediatrics Comparable Corpus*. Ph.D. thesis, Universidade Federal de Santa Catarina, Centro de Comunicação e Expressão. Programa de Pós-Graduação em Estudos da Tradução. Master's Thesis.

Béatrice Daille. 2012. Building bilingual terminologies from comparable corpora: The TTC TermSuite. In *5th Workshop on Building and Using Comparable Corpora with special topic" Language Resources for Machine Translation in Less-Resourced Languages and Domains", co-located with LREC 2012*.

Julien Delaunay, Tran Thi Hong Hanh, Carlos-Emiliano González-Gallardo, Georgeta Bordea, Mathilde Ducos, Nicolas Sidere, Antoine Doucet, Senja Pollak, and Olivier de Viron. 2024. Coastterm: A corpus for multidisciplinary term extraction in coastal scientific literature. In *Text, Speech, and Dialogue - 27th International Conference, TSD 2024, Brno, Czech Republic, September 9-13, 2024, Proceedings, Part I*, volume 15048 of *Lecture Notes in Computer Science*, pages 97–109. Springer.

Giorgio Maria Di Nunzio, Stefano Marchesin, and Gianmaria Silvello. 2023. A systematic review of automatic term extraction: What happened in 2022? *Digital Scholarship in the Humanities*, 38(Supplement_1):i41–i47.

Koraljka Golub, Douglas Tudhope, Marcia Lei Zeng, and Maja Žumer. 2014. Terminology registries for knowledge organization systems: Functionality, use, and attributes. *Journal of the association for information science and technology*, 65(9):1901–1916.

Amir Hazem, Mérième Bouhandi, Florian Boudin, and Béatrice Daille. 2022. Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 648–662. European Language Resources Association.

Paul Jaccard. 1901. Comparative study of floral distribution in a portion of the Alps and Jura. *The Company Vaudoise Bulletin of Natural Sciences*, 37(5):547–579.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. In *Proceedings of the Eleventh International Conference on Intelligent Systems for Molecular Biology, June 29 - July 3, 2003, Brisbane, Australia*, pages 180–182.

Christian Lang, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann. 2021. Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3607–3620. Association for Computational Linguistics.

Natascia Leonardi et al. 2009. Terminology as a system of knowledge representation: an overview. *La ricerca nella comunicazione interlinguistica: modelli teorici e metodologici*, pages 37–52.

Kan Liu and Lu Chen. 2019. Medical social media text classification integrating consumer health terminology. *IEEE Access*, 7:78185–78193.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. 2018. KAS-term and KAS-biterm: Datasets and baselines for monolingual and bilingual terminology extraction from academic writing. *Digital Humanities*, 7.

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. 2013. Combining c-value and keyword extraction methods for biomedical terms extraction. In *LBM: Languages in Biology and Medicine*.

John Philip McCrae and Adrian Doyle. 2019. Adapting term recognition to an under-resourced language: The case of Irish. In *Proceedings of the Celtic Language Technology Workshop*, pages 48–57.

Kevin Meijer, Flavius Frasincar, and Frederik Hogenboom. 2014. A semantic approach for extracting domain taxonomies from text. *Decision Support Systems*, 62:78–93.

Tomás Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 746–751. The Association for Computational Linguistics.

Maria Teresa Musacchio et al. 2001. The contribution of terminology to text analysis in specialised translation. *RIVISTA INTERNAZIONALE DI TECNICA DELLA TRADUZIONE*, 5:29–40.

Aliya Nugumanova, Darkhan Akhmed-Zaki, Madina Mansurova, Yerzhan Baiburin, and Almasbek Maulit. 2022. Nmf-based approach to automatic term extraction. *Expert Syst. Appl.*, 199:117179.

Maria Teresa Pazienza. 1998. A domain-specific terminology-extraction system. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 5(2):183–201.

Senja Pollak, Andraz Repar, Matej Martinc, and Vid Podpecan. 2019. Karst exploration: Extracting terms and definitions from karst domain corpus. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal*, pages 934–956. Lexical Computing.

Behrang QasemiZadeh and Siegfried Handschuh. 2014. The ACL RD-TEC: A dataset for benchmarking terminology extraction and classification in computational linguistics. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, pages 52–63.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. 2020. Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. In *6th International Workshop on Computational Terminology (COMPUTERM 2020)*, pages 85–94. European Language Resources Association (ELRA).

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2022. Tagging terms in text: A supervised sequential labelling approach to automatic term extraction. *Terminology*, 28(1):157–189.

Aivaras Rokas, Sigita Rackeviciene, and Andrius Utka. 2020. Automatic extraction of lithuanian cybersecurity terms using deep learning approaches. In *Human Language Technologies - The Baltic Perspective - Proceedings of the Ninth International Conference Baltic HLT 2020, Kaunas, Lithuania, September 22-23, 2020*, volume 328 of *Frontiers in Artificial Intelligence and Applications*, pages 39–46. IOS Press.

Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.*, 24(5):513–523.

Jadwiga Wozniak-Kasperek. 2014. Terminology as a picture of knowledge organization in a scientific discipline. In *Knowledge Organization in the 21st Century: Between Historical Patterns and Future Prospects*, pages 305–311. Ergon-Verlag.

Yu Yuan, Jie Gao, and Yue Zhang. 2017. Supervised learning for robust term extraction. In *2017 International Conference on Asian Language Processing, IALP 2017, Singapore, December 5-7, 2017*, pages 302–305. IEEE.

Claudia Zavaglia, Leandro Henrique Mendonça de Oliveira, Maria das Graças Volpe Nunes, Maria Fernanda Teline, Sandra Maria Aluisio, et al. 2005. Avaliação de métodos de extração automática de termos para a construção de ontologias. (2005).

Qing Zeng, Sandra Kogan, Nachman Ash, Robert A Greenes, and Aziz A Boxwala. 2002. Characteristics of consumer terminology for health information retrieval. *Methods of information in medicine*, 41(04):289–298.

Ziqi Zhang, Johann Petrak, and Diana Maynard. 2018. Adapted textrank for term extraction: A generic method of improving automatic term extraction algorithms. In *Proceedings of the 14th International Conference on Semantic Systems, SEMANTiCS 2018, Vienna, Austria, September 10-13, 2018*, volume 137 of *Procedia Computer Science*, pages 102–108. Elsevier.

A Appendix A

The annotation of the Hindi Term Extraction Dataset (HTEC 2.0) was conducted in four iterative rounds. Each phase introduced refinements to improve term selection, classification, and annotation consistency. The following sections describe the progressive improvements made in each round.

Round 1: Initial Term Identification and Broad Classification

- Annotators identified domain-specific terms in the education corpus, including both single-word and multi-word expressions. Examples include शिक्षक (teacher), शिक्षा नीति (education policy), and विश्वविद्यालय (university).
- Named entities were annotated, including institutions (such as एम्स्टर्डम विश्वविद्यालय (University of Amsterdam)), organizations (such as शिक्षा मंत्रालय (Ministry of Education)), and individuals (such as रिचर्ड फाइनमैन (Richard Feynman)).
- Acronyms were not initially annotated, leading to inconsistencies in their treatment.
- Several challenges were identified:
  - Multi-word boundaries were often unclear. Certain terms could be either standalone entities or components of larger phrases. For instance, शिक्षा प्रणाली and उच्च शिक्षा प्रणाली.
  - Some commonly used words had domain relevance but were also part of general discourse, leading to inconsistencies. Examples include पंजीकरण (registration) and परीक्षा (examination).
  - Ambiguous terms were not handled systematically, resulting in variation in annotation decisions.
  - The inclusion of foreign-origin terms such as STEM and MOOC lacked clear guidelines.
- In response to these challenges, the following refinements were introduced:
  - A longest valid term selection rule was implemented to standardize the treatment of multi-word terms.
  - An ambiguous category was introduced for terms with unclear domain specificity.
  - A rule was established to ensure acronyms and their full forms were annotated separately but assigned the same category.

Round 2: Refinement of Term Selection Rules and Handling Ambiguity

- The selection criteria for multi-word terms were refined to ensure annotators consistently selected the longest meaningful phrase.
- Acronyms and their full forms were explicitly annotated as distinct entities while maintaining the same semantic classification.
- Guidelines for the treatment of foreign-origin terms were introduced. Commonly used terms such as STEM and MOOC were annotated, whereas highly specialized foreign terms outside the education domain were not.
- Following challenges were identified:
  - Disagreements in compound term boundaries continued to affect annotation consistency.
  - Some terms exhibited overlap between categories. For example, शिक्षा प्रणाली could be classified under both शिक्षा नीति and शिक्षा के प्रसार के माध्यम.
- To address these issues, the following refinements were introduced:
  - A semantic classification scheme was implemented to improve structured categorization.
  - Overlapping terms were discussed on a case-by-case basis and assigned to the most appropriate category.

Round 3: Introduction of Semantic Classification and Color Coding

- Annotators classified terms into predefined semantic categories, improving clarity in classification.
- Color coding was introduced, assigning distinct colors to each category to enhance visualization.
- Overlapping terms were systematically discussed and assigned to a single category based on contextual usage.
- Following challenges were identified:
  - Certain ambiguous terms continued to lack clear classification criteria.
  - Some categories overlapped, requiring additional clarification.
- In response, the following refinements were made:

  - The ambiguous category rules were further refined to ensure consistency in annotation.
  - Final validation checks were introduced to improve annotation agreement.

Round 4: Final Validation and Quality Check

- A final validation process was conducted, involving cross-review by annotators to resolve inconsistencies and improve inter-annotator agreement.
- Overlapping terms were systematically assigned after discussions among annotators.
- The Jaccard Index evaluation was conducted to measure annotation consistency before finalizing the dataset.
- Following challenges were addressed:
  - Inter-annotator agreement was improved through refined classification rules.
  - The final validation process removed inconsistencies, ensuring a high-quality dataset.