Enhancing Term Extraction with Larger Background Corpora and Embedding-Based Filtering

Zirui Han
New York University
zh2267@nyu.edu

Abstract

This study explores how variations in background corpus size influence the quality of term extraction using the Termolator tool, a critical step for improving domain-specific glossary creation and information retrieval. We evaluate results obtained from four different background corpus sizes (200, 500, 1000, and 2000 documents) to determine whether more extensive background data leads to improved term identification. To assess both precision and recall, we begin with a manual annotation of 100 randomly selected candidate terms from each output, recognizing the potential for bias and inconsistency introduced by non-expert labeling, which could affect the reliability of evaluation. To mitigate this, we introduce two authoritative reference sets: the Wikipedia Glossary of Artificial Intelligence and a scholarly glossary of Computer Vision terms by Haralick and Shapiro. These references enable a more objective evaluation of extracted terms.

Our findings reveal that larger background corpora enhance the richness of domain-relevant terms but concurrently increase irrelevant entries, creating a trade-off between precision and coverage. To address this, we leverage embedding-based semantic filtering—a novel application in this context—using cosine similarity scores with domain seed terms to refine the candidate list. This approach effectively balances the benefits of expanded background data against the influx of extraneous terms. After applying the threshold, we achieve enhanced precision and recall, demonstrating that embedding-based semantic filtering complements frequency-based extraction methods.

1 Introduction

Domain-specific terminology, such as 'gradient descent' in machine learning or 'polymerase chain reaction' in molecular biology, forms the foundational vocabulary for specialized fields, playing a crucial role in areas such as information extraction, machine translation, and the construction of domain-specific glossaries and ontologies (Bowker, 1998; Cabré Castellví, 1999). Extracting this terminology automatically often involves contrasting a collection of domain documents (the foreground corpus) with one or more general-purpose corpora (the background corpora), thereby highlighting terms that appear disproportionately in the domain content. Tools such as the Termolator (Meyers et al., 2016) implement this frequency-based approach, offering a practical means to identify domain-specific terms without requiring extensive manual annotation.

Despite its utility, the effectiveness of a frequency-based term extraction method can be sensitive to the choice and size of the background corpus. For instance, while a corpus with 200 documents might lack statistical robustness, one with 2000 documents could include tangentially related texts that dilute the specificity of extracted terms. Intuitively, larger background corpora should provide a more robust statistical baseline, enabling the extraction algorithm to better discriminate between domain-specific terms and more general vocabulary. However, increasing the background corpus size does not come without costs. Beyond a certain point, adding more general texts may introduce subtle noise and allow loosely related but non-essential terms to appear statistically relevant. Thus, while larger background sets often improve coverage and recall, they may also degrade precision by admitting more extraneous or tangential terms.

This study systematically examines how varying background corpus sizes (200, 500, 1000, 2000 documents) impact the balance between precision and recall in term extraction. We begin by using the Termolator (Meyers et al., 2016) with varying background corpus sizes, and we then assess the extracted terms against both a manually cre-

ated annotation set and two authoritative references: the Wikipedia Glossary of Artificial Intelligence and the Computer Vision glossary by Haralick and Shapiro (Haralick and Shapiro, 1985). While larger background corpora do capture more domain-relevant terminology, they simultaneously allow more irrelevant terms into the candidate set.

To address the precision-recall trade-off, we propose a novel integration of embedding-based semantic similarity measures, refining term extraction with a cosine similarity threshold to filter semantically irrelevant candidates. This method successfully reduces noise, thereby enhancing precision without undermining the advantages of using a larger background corpus.

2 Related Work

Term extraction research has produced a variety of methodologies aimed at identifying domain-specific vocabulary from unstructured text. Early approaches focused on statistical indicators such as term frequency, term specificity, or association measures (Bowker, 1998; Cabré Castellví, 1999). These methods were supplemented by linguistic and pattern-based techniques that leverage morpho-syntactic information or context-based filters. As computational resources and NLP methodologies matured, hybrid approaches combining statistical and linguistic clues emerged, enabling more accurate and nuanced term extraction (Jacquemin, 2001; Ananiadou, 1996).

Among the tools developed for domain term extraction, the Termolator (Meyers et al., 2016) has gained prominence for its adaptable, frequency-based methodology. The Termolator compares a domain-specific (foreground) corpus against one or more general (background) corpora to highlight terms that occur disproportionately in the foreground. This differential frequency approach provides a practical and efficient means for identifying candidate terms in varied domains, from the sciences to the humanities.

In previous studies using the Termolator, careful background corpus selection has been emphasized as a key factor influencing the quality of extracted terms. A well-chosen background corpus ensures that generic language is distinguished from domain-specific terminology, improving the precision of extracted lists. Conversely, when the background corpus is too limited, the statistical baseline may be weak, resulting in the inclusion of overly general vocabulary. Conversely, enlarging the background corpus can improve the baseline and potentially capture a richer and more accurate set of terms—up to a point. Beyond that, excessively large background sets can introduce noise and dilute precision by admitting frequent but conceptually irrelevant terms.

Recent advances in embedding-based language models provide a promising avenue to address these limitations. By leveraging semantic similarity measures, candidate terms can be filtered based not only on their statistical prominence but also on their alignment with known domain concepts. For instance, referencing well-established glossaries—such as the Wikipedia Glossary of Artificial Intelligence or Haralick and Shapiro's glossary of Computer Vision terms (Haralick and Shapiro, 1985)—helps anchor candidate terms in a validated conceptual space.

My work builds on these previous efforts. By experimenting with multiple background corpus sizes in the Termolator pipeline, we illuminate the trade-off between recall and precision. Subsequently, we integrate embedding-based similarity scoring to refine the candidate lists, demonstrating that semantic filtering can selectively remove noise introduced by a large background corpus without sacrificing the enhanced recall it affords.

3 Data

To investigate the effect of background corpus size on term extraction, we focused on the foreground corpus of Artificial Intelligence and Computer Vision as two examples. The experiments involved varying the size of the background corpus used by the Termolator at four levels: 200, 500, 1000, and 2000 documents. These sizes were chosen to explore whether increasing amounts of background data would yield more accurate and comprehensive domain terminology or, conversely, introduce unwanted noise.

3.1 Manually Labeled Data

As an initial evaluation, we performed a manual annotation of 100 candidate terms randomly selected from the Termolator's output at each background corpus size. Specifically, we extracted the candidate terms and labeled them as domain-relevant terms (positive) or non-domain terms (negative) based on the domain relevance. Term: The term is directly relevant to the specified domain (e.g., Com-

puter Vision or Artificial Intelligence). It represents a concept, method, or entity commonly used or cited in that field. Non-Term: The term is unrelated or only tangentially connected to the specified domain. This step provided a preliminary assessment of how well the Termolator performed with different background corpus sizes. However, since the annotator lacked full domain expertise, human error in labeling was possible, motivating a further need for more authoritative references.

3.2 Authoritative Reference Sets

To mitigate the uncertainty inherent in a single, non-expert annotator's judgments, we introduced two authoritative reference sets (answer keys) to more objectively gauge precision and recall:

1. **Wikipedia Glossary of Artificial Intelligence:** We compiled a set of 452 terms from the official Wikipedia Glossary of Artificial Intelligence. These terms reflect a broadly accepted understanding of key concepts in AI. Since our foreground and background data are also drawn from Wikipedia sources, this glossary offers an internally consistent reference point, albeit one that may partly mirror the underlying corpus.

2. **Glossary of Computer Vision Terms by Haralick and Shapiro:** To cross-validate with a dataset external to Wikipedia, we utilized a scholarly glossary of Computer Vision terms prepared by Haralick and Shapiro (Haralick and Shapiro, 1985). This glossary comprises 457 terms and represents a more specialized and technical perspective. By using this dataset as an answer key, we aimed to ensure that our evaluation measured genuine domain relevance rather than overfitting to the Wikipedia environment.

Together, these two glossary sets provide a stable ground truth against which we can measure both precision and recall. High alignment of extracted terms with these authoritative lists indicates strong domain specificity and relevance.

4 Methodology

Our methodology centers on adapting the Termolator's workflow to handle varying background corpus sizes and subsequently applying semantic filtering. While the Termolator's conceptual framework remains unchanged, we made specific code adjustments to streamline the extraction process, incorporate larger background sets, and enable embedding-based filtering as a post-processing step.

4.1 Term Extraction Modifications

We began by modifying the Termolator's run script to dynamically accept different background corpus sizes (200, 500, 1000, and 2000 documents). As illustrated in the provided code snippet, we introduced a loop over the predefined background corpus sizes. Inside this loop: We programmatically set the number of articles to retrieve for both the foreground and background corpora based on the chosen size. System calls to custom shell scripts were used to fetch Wikipedia articles for each subclass in the background. We ensured that the number of retrieved articles scaled appropriately with the selected background size. Each run generated a corresponding set of candidate terms for subsequent evaluation steps.

4.2 Manual Labeling and Reference-Based Evaluation Preparation

To assess precision and recall with minimal bias, we implemented code to:

1. Randomly sample 100 candidate terms from the Termolator's output and record them for manual labeling.

2. Prepare input files for cross-referencing against authoritative glossaries. We wrote a small parsing function to read in the terms from the Termolator's output and compare them against the entries in the Wikipedia AI Glossary and the Haralick-Shapiro Computer Vision Glossary. The code filtered irrelevant explanations, retaining only the core terms for straightforward matching.

4.3 Embedding-Based Semantic Filtering Implementation

For the embedding-based filtering, we used Python scripts that:

1. Loaded a sentence-transformer model (e.g., `all-MiniLM-L6-v2`) via the `sentence_transformers` library.

2. Encoded both the candidate terms and a set of domain-specific seed terms (including those derived from the glossaries) into vector embeddings.

3. Computed cosine similarities and filtered out terms falling below a chosen threshold (e.g., 0.5). The code for this step involved iterating over all candidate terms, calculating similarities, sorting them by score, and writing the filtered, ranked lists to output files.

This additional script ran after the Termolator extraction and reference-based evaluation, acting as a final refinement step. The code leveraged vectorized operations via `numpy` and minimal loops for efficiency.

5 Experiments

We conducted four experiments to evaluate how background corpus size and embedding-based filtering affected the Termolator's performance. Unlike the methodology section, which focuses on code adjustments and procedures, this section details the experimental design, evaluation criteria, and metrics.

5.1 Experiment 1: Manual Labeling

In the first experiment, we aimed to gauge the effect of background corpus size on precision by human judgment. Using the Termolator's output for Computer Vision as the foreground domain, we extracted candidate terms at each background size (200, 500, 1000, 2000). We then randomly selected 100 terms from each output list and manually labeled them as domain-relevant or not. Although subjective and potentially biased, this provided a preliminary indication of how increasing the background data influenced the quality of extracted terms.

5.2 Experiment 2: Reference-Based Evaluation

To obtain an objective measure of precision and recall, we introduced two authoritative glossaries: the Wikipedia AI Glossary and the Haralick-Shapiro Computer Vision Glossary. After extracting terms at various background sizes, we cross-referenced them against these glossaries. Precision was calculated as the proportion of extracted terms found in the glossaries, and recall as the proportion of glossary terms retrieved by our system. This step measured how effectively the Termolator captured legitimate domain terminology versus irrelevant noise as the background corpus expanded.

5.3 Experiment 3: Embedding-Based Analysis

Having observed that larger background sets often increase recall but decrease precision, we used embedding-based analysis to understand the semantic nature of the extracted terms. We computed average cosine similarities between candidate terms and core domain concepts (e.g., "computer vision") at different background sizes. This experiment aimed to reveal whether expanding the background corpus led to a dilution of semantic relevance.

5.4 Experiment 4: Embedding-Based Filtering

Finally, we applied the embedding-based filtering method. Here, we took the largest background corpus scenario and filtered candidate terms below a certain similarity threshold. We then re-evaluated precision and recall. This experiment tested whether semantic filtering could restore or improve precision and recall, striking a more favorable balance between coverage and relevance.

In all four experiments, the emphasis was on assessing the trade-offs between background corpus size, domain term coverage, and the introduction of extraneous terms. While the methodology section described how we modified and ran the code to generate these outputs, the experiments described here focus on the evaluation design, comparative analyses, and metrics that reveal the effectiveness of our approach.

6 Results

6.1 Experiment 1: Manual Labeling (Computer Vision)

Table 1 presents the precision estimates derived from manual annotation for varying background corpus sizes.

Table 1: Manual precision estimates for Computer Vision foreground (100 sampled terms per set).

Background Size	Precision Estimate
200	62.00% (62/100)
500	54.00% (54/100)
1000	40.00% (40/100)
2000	44.00% (44/100)

6.2 Experiment 2: Reference-Based Evaluation

Tables 2 and 3 show the precision and recall for the Computer Vision and Artificial Intelligence domains, respectively, using authoritative glossaries as answer key(wikipedia's glossary of artificial intelligence and glossary of computer vision) .

Table 2: Precision and recall for Computer Vision terms (Haralick and Shapiro glossary, 457 terms).

Bg Size	#Terms Extracted	#Matched Terms	Precision	Recall
200	1,844	36	1.95%	7.88%
500	4,401	48	1.09%	10.50%
1000	4,991	45	0.90%	9.85%
2000	4,993	51	1.02%	11.16%

Table 3: Precision and recall for AI terms (Wikipedia AI glossary, 444 terms).

Bg Size	#Terms Extracted	#Matched Terms	Precision	Recall
200	1,507	20	1.33%	4.50%
500	2,304	51	2.21%	11.49%
1000	3,630	64	1.76%	14.41%

6.3 Experiment 3: Embedding-Based Analysis

Table 4 shows how the average embedding similarity between candidate terms and a core domain concept (e.g., "computer vision") decreases as the background corpus grows, indicating that non-domain noise increases with corpus size.

Table 4: Average embedding similarity to the core domain concept for varying background sizes.

Foreground	Bg Size	Avg. Similarity
Computer Vision	200	0.1723
	500	0.1635
	1000	0.1561
	2000	0.1525
AI	200	0.2810
	500	0.2511
	1000	0.2459

As background size increases, the average similarity declines, reflecting diminished semantic coherence in the candidate terms.

6.4 Experiment 4: Embedding-Based Filtering

Applying a 0.5 similarity threshold to the 2000-background-size Computer Vision terms significantly improved the relevance of extracted terms. Table 5 shows that this filtering step increased both precision and recall relative to the unfiltered set.

Table 5: Impact of embedding-based filtering (threshold=0.5) on Computer Vision terms with a 2000-article background. Evaluated against the Haralick and Shapiro glossary (457 terms).

Method	#Terms	Precision	Recall
Unfiltered (2000)	4993	1.02%	11.16%
Filtered (2000)	1999	4.50%	19.69%

The filtering substantially raised precision (from 1.02% to 4.50%) and improved recall (from 11.16% to 19.69%).

7 Discussion

The results from our four experiments reveal a complex interplay between background corpus size and the quality of extracted domain terms.

From the manual labeling results in Table 1, we see a clear decline in precision as the background corpus grows. This suggests that although the Termolator identifies more candidate terms, it also introduces a proportionally greater number of irrelevant ones. Since the random sampling for manual evaluation selects from an increasingly noisy candidate pool, the chance of selecting a non-term rises, reducing observed precision.

Experiment 2's reference-based evaluation (Tables 2 and 3) confirms that larger background corpora raise recall while lowering precision. Thus, while more domain terms are captured, more extraneous terms are introduced as well.

The embedding-based analysis in Experiment 3 (Table 4) clarifies the nature of this noise. The declining average cosine similarity scores highlight that as the background corpus expands, the growth of semantically irrelevant terms outpaces the gain in relevant domain terms. This indicates that the semantic coherence of the candidate list diminishes at larger scales.

Experiment 4 demonstrates that embedding-based filtering can effectively restore and even improve term quality. By applying a similarity threshold, irrelevant terms are pruned, leading to substantial gains in both precision and recall (Table 5). The results show that a larger background corpus is indeed beneficial for coverage, as long as a semantic filtering step is applied to maintain precision.

These findings underscore the importance of bal-

ancing the advantages of a large background corpus with strategies to manage the influx of irrelevant terms. Embedding-based filtering, in particular, proves crucial for refining the extracted lists, ensuring that the benefits of expanded coverage are not overshadowed by a flood of semantically distant candidates. Additionally, the choice of threshold and domain seed terms can be fine-tuned to optimize performance further, suggesting a flexible approach adaptable to various domains and use cases.

8 Future Work

This study highlights several promising directions for future research:

1. **Exploring Larger and More Diverse Background Corpora:** While we examined background sizes up to 2000 documents, future work could investigate even larger and more diverse collections. Testing at larger scales and with different domains will determine whether the observed improvements hold and whether further gains in coverage can be achieved without diminishing term quality.

2. **Integrating Embedding-Based Filtering into the Termolator Pipeline:** Currently, we applied the embedding-based filtering as a post-processing step. A natural next step is to incorporate this semantic similarity measurement directly into the Termolator's workflow. By automatically identifying relevant superclasses and subclasses of the foreground term and using them as seed terms, the Termolator could calculate embeddings and filter out irrelevant candidates during the extraction process itself. This integrated approach has the potential to streamline term extraction and improve precision from the outset.

3. **Adaptive Thresholding and Domain-Specific Models:** Another avenue for exploration involves dynamically adjusting the similarity threshold based on domain characteristics or using domain-specific embedding models.

By pursuing these directions, future studies can further enhance the Termolator's accuracy, scalability, and adaptability, ultimately supporting more robust domain-specific terminology extraction.

9 Conclusion

In this study, we explored how varying the size of the background corpus influences the quality of domain term extraction using the Termolator. Our experiments demonstrated a clear trade-off: while larger background corpora improved recall by capturing more true domain terms, they also introduced an increasing number of irrelevant candidates, thereby reducing precision. Embedding-based semantic similarity analysis further revealed that as the background corpus grows, semantically unrelated terms proliferate faster than relevant ones, leading to a decline in overall coherence.

By applying embedding-based filtering, we were able to significantly enhance both precision and recall. This result suggests that large background corpora can indeed be leveraged to improve coverage, provided that a semantic post-processing step refines the candidate list. The synergy between broad statistical extraction and targeted semantic filtering proved effective in reducing noise and improving overall term quality.

Our findings highlight a pathway to more effective term extraction such as integrating embedding-based measures directly into the Termolator's extraction pipeline, exploring more extensive background corpora, and experimenting with domain-specific embedding models.

References

Sophia Ananiadou. 1996. Automatic term recognition—a review focusing on the computational and linguistic aspects. *Terminology*, 3(2):259–289.

Lynne Bowker. 1998. *Terminology: Theories, methods and applications*. John Benjamins Publishing, Amsterdam, The Netherlands.

M. Teresa Cabré Castellví. 1999. *Terminology: Theory, Methods and Applications*. John Benjamins Publishing, Amsterdam, The Netherlands.

Robert M. Haralick and Linda G. Shapiro. 1985. Glossary of computer vision terms. In *Proceedings of the Workshop on Computer Vision: Representation and Control*, Annapolis, MD, USA. Dept. of Electrical Engineering, University of Washington.

Christian Jacquemin. 2001. *Spotting and Discovering Terms through Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Adam Meyers, Michelle Zhu, Tian Shi, Rui Jia, and Ralph Grishman. 2016. The termolator: Terminology recognition based on chunked sequences of ats. In *Proceedings of the Tenth International Conference on*

Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. European Language Resources Association (ELRA).