A Lightweight N-gram Approach to Abbreviation Expansion in Large Corpora

Tjaša Šoltes, Marko Bajec
Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113, 1000 Ljubljana, Slovenia
tjasa.soltes@fri.uni-lj.si

Abstract

We present a lightweight, corpus-based approach to abbreviation expansion that relies solely on contextual N-gram statistics. The method models local context using two-sided and one-sided bigram and trigram counts extracted from a large domain-specific corpus. Candidate expansions are selected through linear interpolation of context-specific evidence, enhanced with reliability-based scaling to mitigate sparse data effects. The approach does not require external linguistic resources, pretrained language models, or explicit morphosyntactic analysis, making it suitable for domain-specific and resource-constrained settings. Experiments conducted on a large Slovene medical corpus demonstrate that interpolation generally outperforms strict backoff strategies, with notable improvements for medium- and low-frequency abbreviations. Despite its simplicity, the proposed framework achieves robust performance while remaining computationally efficient and scalable.

Keywords: text normalization, abbreviation expansion, N-gram model

1. Introduction

Abbreviation expansion is an essential preprocessing step in many natural language processing pipelines, particularly in applications such as text-to-speech (TTS) and automatic speech recognition (ASR), where non-standard word forms must be converted into their fully expanded equivalents. Existing approaches to abbreviation expansion often rely on predefined lexicons, rule-based systems, or semantic disambiguation methods that attempt to determine the underlying meaning of a given abbreviation.

In this paper, we propose a data-driven N-gram-based approach to abbreviation expansion in large corpora. We focused on abbreviations, which we defined as alphabetical strings ending with a dot. The method relies on contextual bigram and trigram information to select the most probable expansion candidate. Context representations are transformed into a compact binary form to enable efficient lookup and scoring, making the approach computationally lightweight. The proposed approach is fully unsupervised: all candidate expansions and contextual statistics are derived directly from raw corpus data.

Unlike many existing methods, the proposed approach does not require external linguistic resources, manually curated abbreviation dictionaries, or explicit semantic interpretation of abbreviations. It can be adapted to specific domains or languages by training directly on in-domain corpora, making it particularly suitable for low-resource or specialized settings. Furthermore, the method is computationally efficient and can be applied to large datasets without extensive preprocessing. Lightweight approaches remain valuable in clinical and domain-specific settings where computational resources, latency constraints, or data privacy concerns limit the applicability of large neural models.

For morphologically rich languages such as Slovene, abbreviation expansion presents an additional challenge: the correct expansion often depends on grammatical case and other morphosyntactic features. Traditional solutions typically require morphosyntactic analysis to determine the appropriate inflected form. [1] In contrast, our context-based N-gram approach implicitly captures such information from surface-level contextual patterns, eliminating the need for explicit morphological analysis while still producing grammatically appropriate expansions.

The contributions of this work are threefold: (1) we introduce a lightweight N-gram-based framework for context-sensitive abbreviation expansion that does not rely on external linguistic resources; (2) we propose a reliability-aware interpolation strategy that improves robustness in sparse-data conditions; and (3) we provide an evaluation on a large Slovene medical corpus with analysis across abbreviation frequency bands.

2. Related Work

Early work on abbreviation expansion was largely driven by distributional similarity and manually designed representations. Terada et al. (2004) pro-

[1] Slovene uses 6 grammatical cases and three grammatical numbers: alongside singular and plural, dual is also used. This means that a single word can have up to 18 forms.

Proceedings of the Third Workshop on Computation and Written Language (CAWL 2026) @LREC 2026, pages 95–100
12 May, 2026.

pose a method that ranks candidate expansions using cosine similarity between context vectors derived from abbreviation-rich and abbreviation-poor corpora, supplemented with character-based filtering rules that approximate human abbreviation patterns. In a similar vein, Pakhomov et al. (2005) introduces a semi-supervised approach to acronym disambiguation that automatically harvests sense-specific contexts from the Web, MEDLINE, and clinical notes. These contexts are encoded as vector-space representations, and meanings are assigned via cosine-based context similarity. Related distributional ideas also appear in text normalization work, where Cook and Stevenson (2009) propose an unsupervised method that learns mappings from non-standard SMS forms to standard words using contextual similarity in unannotated corpora, demonstrating that normalization can be addressed through corpus-based inference without labeled data or handcrafted rules.

Subsequent research increasingly framed abbreviation expansion as a structured inference problem, often within noisy-channel architectures. Roark and Sproat (2014) introduce a conservative, largely unsupervised methodology that mines billions of words of unannotated text to extract candidate abbreviation–expansion pairs and trains a joint 7-gram character model to score their plausibility. This pair model is combined with two context-sensitive components—a pruned trigram language model and an SVM classifier using contextual likelihood and token-level features—with expansions applied only when both models independently agree. Gorman et al. (2021) further formalize this paradigm by introducing a large, openly available dataset of English ad hoc abbreviations and evaluating two supervised noisy-channel models: a finite-state pipeline combining a trigram language model with a character-level pair model, and a neural pipeline pairing an LSTM language model with a subsequence-based abbreviation model. Their results show that both approaches achieve near-human accuracy, with the neural variant performing best.

More recent work has shifted toward neural architectures that integrate morphological and contextual modeling more tightly. Chopard and Spasić (2019) propose a modular neural framework that combines pattern-based abbreviation detection, a character-level RNN to distinguish abbreviations from acronyms, and a Siamese network to model morphological formation and generate candidate expansions, followed by context-based ranking without reliance on external lexical resources. At a larger scale, contemporary approaches increasingly leverage deep neural networks and transformer-based language models, enabling end-to-end contextual modeling. Privacy-preserving training strategies, such as reverse substitution on public corpora, show that high-quality clinical expansion models can be developed without direct access to protected health records (Rajkomar et al., 2022). Large language models have also been evaluated in zero-shot and fine-tuned settings for clinical acronym expansion, demonstrating strong adaptability alongside persistent domain-specific limitations (Kugic et al., 2024; Nezhad et al., 2025). In parallel, lightweight and task-oriented systems emerging from shared tasks emphasize efficiency, domain adaptation, and standardized evaluation protocols (Kugic et al., 2024).

3. Corpus and data

The experiment was conducted on a proprietary corpus of real Slovenian medical documents collected from various fields of medicine and various institutions. The corpus includes 3,893,434 sentences and 59,083,266 tokens (137,423 types).

We identified 237 abbreviations types in the corpus (with the total frequency of 87,610). We defined an abbreviation as any alphabetic string that ends with a dot. We divided the abbreviations into five frequency categories, that is: very frequent (more than 5,000 occurrences), frequent (1,000–5,000), medium (100–1,000), rare (10–100), and very rare (fewer than 10). The median value of expansions per abbreviation is 4.

4. Methodology

4.1. Data

To ensure scalability and efficient lookup, all contextual N-gram statistics were stored in compact binary files (*.bin)*. As per standard practice, each unique token was mapped to an integer identifier, allowing N-gram contexts to be represented as fixed-size integer tuples rather than string sequences. This representation significantly reduces memory usage and accelerates access during candidate retrieval and scoring. Separate binary count tables were constructed for each context type (center trigrams, left and right trigrams, and left and right bigrams), which can be observed in 1, enabling fast aggregation of contextual evidence at inference time. The use of binary storage allows the method to operate efficiently even on large corpora without requiring database systems or external indexing frameworks.

4.2. Detecting abbreviations and their possible expansions

We defined an abbreviation as any token that appears in the sentence and ends with a dot which is preceded by letters only. Due to some mistakes in

N-gram name	N-gram content
center trigram (C3)	(t-1, _, t+1)
left trigram (L3)	(t-2, t-1, _)
right trigram (R3)	(_, t+1, t+2)
left bigram (L2)	(t-1, _)
right bigram (R2)	(_, t+1)

Table 1: Representation of the five N-gram context types.

sentence tokenization, we identified 18 cases (out of 237) where a sentence was not split which led to the last word in the sentence being incorrectly identified as an abbreviation.

We compiled a list of abbreviations to be excluded from the expansion process. This decision was motivated by the observation that most conventionalized (highly frequent and generally non-field-specific) abbreviations rarely occur in their expanded form. This category primarily comprises titles (e.g., *dr.* 'doktor' (doctor), *ga.* 'gospa' (Mrs.)), as well as common non-domain abbreviations (e.g., *npr.* 'na primer' (e.g.)). Additionally, it includes selected domain-specific abbreviations that are conventionally not written out in full, such as, for example, various forms of *b. p.* 'brez posebnosti' (without specifics). The final exclusion list comprised a total of 34 abbreviations.

Candidate expansions are generated directly from contextual N-gram statistics extracted from the corpus. For each occurrence of an abbreviation, its surrounding context within a five-token window is matched against precomputed bigram and trigram tables. These tables store all tokens observed in identical context. The union of all tokens retrieved from the matching context tables forms the candidate set.

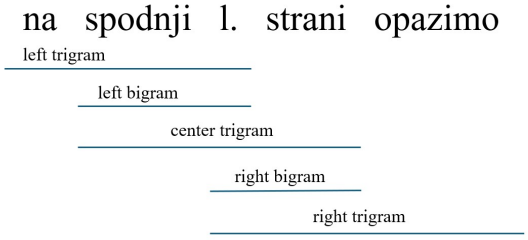

Figure 1: Representation of N-grams in an example sentence *na spodnji l. strani opazimo* 'at bottom l. side we_notice'.

As can be observed in 1s, in the example sentence: *na spodnji l. strani opazimo* (literally: 'at bottom l. side we_notice'), the target abbreviation to expand is 'l.' (to be expanded into *levi* 'left' in singular feminine locative form). The candidates are retireved from the following contexts: left trigram (*na, spodnji, _*), left bigram (*spodnji, _*), center trigram (*spodnji, _, strani*), right bigram (*_, strani*) and right trigram (*_, strani, opazimo*).

The contexts store all tokens appearing in these N-grams in the position of the target abbreviation *l.* and the frequency of the N-gram. Out of these, the candidates are selected. A candidate is always a single token. It must satisfy several constraints: they must share the prefix of the abbreviation (excluding the trailing dot), consist only of alphabetic characters, and not themselves be abbreviations. This procedure ensures that only contextually plausible and orthographically compatible expansions are considered during scoring.

4.3. Scoring Function

Once a list of candidates is generated, the candidates are scored. We compared two scoring approaches: backoff and linear interpolation.

We used a simple backoff algorithm where we used the result of a single context. We defined the order of contexts, and started from most determinative. For example: if the center trigram produces any candidates, the abbreviation is expanded into the candidate with the highest frequency (in the context of this trigram). If no candidates are produced in the center trigram, the same process is repeated with the right trigram, then left trigram, and so on.

The second approach we used for scoring is linear interpolation. First, the candidates were produced in all contexts. For each candidate w, the final score is computed using linear interpolation with predefined fixed weights across contextual models:

$$S(w) = \sum_i \lambda_i \cdot R_i \cdot \frac{c_i(w)}{\sum_{w'} c_i(w')}$$

where λ_i denotes the interpolation weight assigned to context type i, $c_i(w)$ represents the count of candidate w in that context, and R_i is a reliability factor defined as:

$$R_i = \frac{N_i}{N_i + k}$$

with N_i being the total number of observations for context i and k a smoothing constant. This reliability scaling reduces the impact of sparse contextual evidence.

It is important to note that the candidate w represents a single token (i.e., a possible expansion of the abbreviation) rather than a sequence of tokens. The proposed method ranks individual candidate expansions conditioned on observed contextual patterns.

The interpolation weights and smoothing constant were selected using grid search on a held-out validation subset.

5. Results

The test set was manually constructed to ensure balanced evaluation across abbreviation frequency bands. Abbreviations were grouped into five categories based on their corpus frequency (very frequent, frequent, medium, rare, and very rare). Sentences were randomly sampled for each abbreviation within each category. Only the selected abbreviation and its expansion were evaluated in each sentence, even if additional abbreviations appeared in the sentence. Sentences with insufficient contextual information, ambiguous expansions, or predominantly non-alphabetic content (e.g., laboratory reports) were excluded from the test set. The final test set contains 1030 instances and provides a stratified evaluation across varying levels of data sparsity. We included 151 cases of very frequent abbreviations, frequent abbreviations contribute to 252 cases, medium abbreviations appear 432 times, rare abbreviations contribute to 90 cases and very rare abbreviations are included in all 105 cases.

The abbreviations in the test set exhibited an average of 3.3 expansions, with a median of 3, a maximum of 11, and a minimum of 1.

Table 2 presents the performance of different scoring strategies across abbreviation frequency bands. The weakest configuration is the fixed-weight interpolation[2] model that includes all bigrams during candidate production, achieving an overall accuracy of 77.4%. This setting performs consistently worse across all frequency bands, with particularly low accuracy for very frequent (69.5%) and frequent abbreviations (77.0%). These results indicate that incorporating left bigram context during candidate retrieval substantially increases noise and negatively affects ranking performance.

Excluding left bigrams from candidate production while retaining fixed-weight interpolation leads to a marked improvement, raising overall accuracy to 87.1%. Gains are observed across all frequency bands, including very frequent (88.1%), frequent (90.1%), medium (89.6%), rare (78.9%), and very rare abbreviations (75.2%). This demonstrates that separating candidate generation from scoring is crucial for maintaining precision.

Introducing reliability-weighted interpolation further improves performance. With $k=5$ and left bigrams excluded from candidate production, but used in scoring, the model achieves the best overall accuracy of 89.2%. The most pronounced improvements are observed for very frequent abbreviations (95.4%–96.0%) and medium-frequency abbreviations (up to 91.9%). Rare abbreviations also benefit, improving from 78.9% under fixed interpolation to 81.1%. Performance for very rare abbreviations remains stable at approximately 75%, suggesting that extremely sparse contextual evidence continues to limit predictive accuracy regardless of scoring strategy.

Backoff strategies perform competitively, particularly for rare and very rare abbreviations, where the best backoff configuration achieves 83.3% and 76.2%, respectively. However, backoff does not surpass reliability-weighted interpolation in overall accuracy, as it underutilizes partial contextual evidence when multiple informative signals are available.

Overall, the results confirm that excluding left bigrams from candidate production is essential for stable performance, and that reliability-aware interpolation provides the most robust scoring strategy across frequency bands.

5.1. Discussion

The experimental results provide several important insights into the behavior of context-based N-gram models for abbreviation expansion. The lowest-performing configuration—fixed-weight interpolation with all bigrams included in candidate production—achieves only 77.4% overall accuracy, with particularly poor performance for very frequent abbreviations (69.5%). This indicates that incorporating left bigram context during candidate retrieval substantially increases noise. These results show that overly permissive candidate generation can harm performance.

Excluding left bigrams from candidate production raises overall accuracy to 87.1%, with consistent improvements across all frequency bands. This demonstrates the importance of separating candidate retrieval from scoring: restricting candidate generation to more informative contexts reduces noise while preserving useful contextual evidence for ranking.

Reliability-aware interpolation further improves performance, reaching up to 89.2% overall accuracy. The gains are most pronounced for very frequent and medium-frequency abbreviations, where contextual evidence is sufficient but unevenly distributed. Reliability scaling downweights sparse contexts and leads to consistent improvements for rare abbreviations as well (from 78.9% to 81.1%). Performance for very rare abbreviations remains stable at approximately 75%, suggesting that interpolation alone cannot overcome extreme sparsity.

In such sparse conditions, strict backoff performs slightly better, achieving 83.3% for rare and 76.2%

[2]The weight in all interpolation approaches used were: center trigram: 0.45, left trigram: 0.20, right trigram: 0.20, lef bigram: 0.075, right bigram: 0.075

Model	All	V. Freq.	Freq.	Medium	Rare	V. Rare
Interpolation (no reliability)						
all context for candidate production	77.4	69.5	77.0	82.2	74.4	72.4
no L2 for candidate prod.	87.1	88.1	90.1	89.6	78.9	75.2
Interpolation (reliability-weighted)						
$k = 5$	89.1	96.0	89.3	91.7	81.1	75.2
$k = 5$, no L2 for candidate prod.	89.2	95.4	89.7	91.9	81.1	75.2
Backoff						
C3 $\rightarrow$ R3 $\rightarrow$ L3 $\rightarrow$ R2 $\rightarrow$ L2	88.1	94.0	88.9	89.4	83.3	76.2
C3 $\rightarrow$ L3 $\rightarrow$ R3 $\rightarrow$ L2 $\rightarrow$ R2	87.3	94.0	88.1	88.9	80.0	75.2

Table 2: Accuracy across interpolation and backoff configurations. Columns "V. Freq.", "Freq.", "Medium", "Rare", and "V. Rare" denote performance over groups of abbreviations binned by frequency. Left context refers to its inclusion in candidate production. Reliability-weighted models use scaling parameter k. In backoff configurations, C3, L3, R3, L2, and R2 denote center, left, and right context n-gram models (trigram or bigram, respectively).

for very rare abbreviations. This reflects the difference between the two strategies: interpolation aggregates multiple signals, whereas backoff relies on a single best-available context, which may be advantageous when most signals are weak.

Overall, the findings highlight three main points: candidate generation plays a decisive role in performance, reliability-aware interpolation provides the most robust overall strategy, and backoff remains competitive under extreme sparsity. Despite its simplicity, the framework demonstrates that surface-level distributional statistics capture sufficient contextual information to resolve many morphosyntactic distinctions implicitly. However, performance remains limited for very rare abbreviations and in cases where candidate expansions are orthographically similar, making contextual discrimination inherently difficult.

5.2. Computational Efficiency

The complete set of binary N-gram tables occupies 169.8 MB of disk space. All experiments were conducted on a standard laptop equipped with an AMD Ryzen 7 5700U CPU (1.80 GHz) and 16 GB RAM. No GPU acceleration was used.

During inference, the system processes approximately 2,375 abbreviation instances per second. These results confirm the computational efficiency of the proposed approach. The model remains compact in size and achieves high throughput, making it suitable for large-scale corpus processing and real-time preprocessing pipelines.

6. Conclusion

This work presented a lightweight, corpus-based N-gram approach to abbreviation expansion that relies exclusively on contextual statistics. The method requires no external linguistic resources, pretrained models, or explicit morphosyntactic analysis, making it suitable for domain-specific and resource-constrained environments.

The experiments demonstrate that reliability-aware interpolation provides the most robust overall solution, outperforming fixed-weight interpolation and generally surpassing strict backoff strategies. The results further show that excluding left context from candidate production significantly improves accuracy, highlighting the importance of carefully separating candidate retrieval from scoring. While backoff remains competitive in extremely sparse conditions, interpolation offers a more flexible and stable framework across frequency bands.

Overall, the findings confirm that distributional N-gram modeling remains a strong and practical approach in resource-constrained settings, particularly in specialized domains and morphologically rich languages. Future work may explore hybrid approaches that integrate lightweight statistical modeling with selective linguistic constraints or domain-adaptive neural components to further improve performance on rare and ambiguous cases.

7. Acknowledgements

This paper was conducted as part of Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language (MEZZANINE J74642), financed from the national budget by a contract between the Slovenian Research Agency and the Faculty of Computer and Information Science, University of Ljubljana.

8. Bibliographical References

Daphné Chopard and Irena Spasić. 2019. A deep learning approach to self-expansion of abbreviations based on morphology and context distance. In *Statistical Language and Speech Processing: 7th International Conference, SLSP 2019, Ljubljana, Slovenia, October 14–16, 2019, Proceedings*, page 71–82, Berlin, Heidelberg. Springer-Verlag.

Paul Cook and Suzanne Stevenson. 2009. An unsupervised model for text message normalization. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 71–78, Boulder, Colorado. Association for Computational Linguistics.

Kyle Gorman, Christo Kirov, Brian Roark, and Richard Sproat. 2021. Structured abbreviation expansion in context. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 995–1005, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amila Kugic, Stefan Schulz, and Markus Kreuzthaler. 2024. Disambiguation of acronyms in clinical narratives with large language models. *Journal of the American Medical Informatics Association*, 31(9):2040–2046.

Nima Shafiei Rezvani Nezhad, Meysam Mansouri, Rabih Abdulkarim Zakaria, and Ruhollah Abolhasani. 2025. Medical abbreviation disambiguation with large language models: Zero- and few-shot evaluation on the medal dataset. *bioRxiv*.

Serguei V. S. Pakhomov, Ted Pedersen, and Christopher G. Chute. 2005. Abbreviation and acronym disambiguation in clinical discourse. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, pages 589–93.

Alvin Rajkomar, Eric Loreaux, Yuchen Liu, Jonas Kemp, Benny Li, Ming-Jun Chen, Yi Zhang, Afroz Mohiuddin, Juraj Gottweis, et al. 2022. Deciphering clinical abbreviations with a privacy protecting machine learning system. *Nature Communications*, 13(1):7456.

Brian Roark and Richard Sproat. 2014. Hippocratic abbreviation expansion. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 364–369, Baltimore, Maryland. Association for Computational Linguistics.

Akira Terada, Takenobu Tokunaga, and Hozumi Tanaka. 2004. Automatic expansion of abbreviations by using context and character information. *Information Processing Management*, 40(1):31–45.