District University Francisco Jose de Caldas

Department of Systems Engineering

Modular Sentence Reconstruction System for the Santa 2024 Kaggle Competition

Gabriela Martínez Silva

Jairo Arturo Barrera Mosquera

Supervisor: Carlos Andrés Sierra Virgüez

A report submitted in partial fulfilment of the requirements of
the District University Francisco Jose de Caldas for the degree of
Systems Analysis and Design in *Systems Engineering*

July 12, 2025

Declaration

We, Gabriela Martínez Silva and Jairo Arturo Barrera Mosquera, of the Department of Systems Engineering, District University Francisco Jose de Caldas, confirm that this is our own work and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work, except where the works of others have been explicitly acknowledged, quoted, and referenced. I understand that if failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalised accordingly.

We give consent to a copy of our report being shared with future students as an exemplar.

We give consent for our work to be made available more widely to members of DUFJdC and public with interest in teaching, learning and research.

Gabriela Martínez Silva
Jairo Arturo Barrera Mosquera
July 12, 2025

Abstract

This report presents a modular system designed to reconstruct coherent sentences from unordered word lists, addressing the Santa 2024 Kaggle competition challenge. The proposed solution integrates lexical normalization, morphological classification, syntactic template generation, and large language model (LLM) refinement to efficiently generate fluent, grammatical sentences. The system achieves high word coverage, fluency, and computational efficiency, outperforming baseline approaches. Results demonstrate the system's adaptability and robustness, with recommendations for further optimization and broader applications.

Keywords: Sentence generation, Text normalization, Architecture, Large Language Model, System

Acknowledgements

We thank the Santa 2024 competition organizers for providing the dataset and evaluation framework. Special thanks to the Universidad Distrital for computational resources and Teacher Andres Sierra for his feedback and support in development.

Contents

List of Figures

List of Tables

List of Abbreviations

DUFJdC	District University Francisco Jose de Caldas
NLP	Natural Language Processing
POS	Part-of-Speech
CSV	Comma-Separated Values
LLM	Large Language Model
RNNs	Recurrent Neural Networks
LSTMs	Long Short-Term Memory

Chapter 1

Introduction

1.1 Background

Natural Language Processing (NLP) is a foundational field in artificial intelligence, enabling breakthroughs in machine translation, summarization, grammar correction, and conversational systems. One persistent challenge is reconstructing coherent sentences from unordered word lists—a task complicated by the factorial growth in possible permutations and the ambiguity of natural language. The Santa 2024 Kaggle competition presents this challenge in a practical context: participants must reconstruct Christmas song lyrics from shuffled word sequences, using each word exactly once. This scenario demands solutions that are not only linguistically sound but also computationally efficient.

1.2 Problem Statement

The central problem is to generate grammatically correct and semantically coherent sentences from unordered input word sequences, where each sequence may contain repeated words and all tokens must be used exactly once. The combinatorial complexity of possible arrangements grows rapidly with input size, making exhaustive search infeasible. Additional challenges include handling words with multiple syntactic roles and maintaining fluency under strict constraints.

1.3 Objectives

Aim: To develop an efficient, modular, and interpretable system capable of reconstructing fluent and grammatically valid sentences from unordered word sequences under full-token-usage constraints.

Objectives:

- Develop a modular, interpretable system for reconstructing fluent and grammatically valid sentences from unordered word sequences.
- Ensure 100% word coverage, including duplicates, in all outputs.
- Integrate linguistic templates and large language model (LLM) refinement for balanced structure and flexibility.
- Achieve computational efficiency suitable for real-time or resource-constrained environments.

- Evaluate system performance using standard NLP metrics and compare with baseline methods.

1.4 Scope

This report focuses on the design, implementation, and evaluation of a sentence reconstruction system for English-language data as defined by the Santa 2024 task. It does not address multilingual generalization, semantic role labeling, or human-in-the-loop generation.

1.5 Assumptions

- Input sequences contain only valid English words.
- Each sequence is a true permutation of a coherent sentence.
- No additional linguistic annotations are available.
- The language model used for evaluation (Gemma 2) is well-calibrated.

1.6 Limitations

- Syntactic templates inferred from shuffled words are probabilistic and can be inaccurate.
- Large input sequences may still pose computational challenges despite pruning.
- Dependency parsing on scrambled input introduces ambiguity.
- Evaluator is sensitive to punctuation and tokenization errors.

Chapter 2

Literature Review

2.1 State-of-the-Art in Sentence Reconstruction and NLP

Early approaches to sentence reconstruction relied on statistical n-gram models, which estimate the likelihood of word sequences based on observed frequencies. While effective for short-range dependencies, these models struggle with long-range syntactic and semantic relationships. Neural language models, such as RNNs, LSTMs, and especially Transformer-based architectures like GPT-2 (1) and BERT (2), have significantly advanced the field by leveraging contextual embeddings and attention mechanisms to capture complex dependencies. These models excel in generating fluent, contextually appropriate sentences but often require large datasets and computational resources.

Rule-based and template-driven methods have also been explored, imposing explicit syntactic constraints on sentence formation using part-of-speech (POS) tagging and dependency parsing. These approaches offer interpretability and control but can be limited by the rigidity of templates and difficulties in handling linguistic ambiguity.

2.2 Project Context in Existing Literature and Systems

The Santa 2024 competition situates the sentence reconstruction problem in a constrained, real-world scenario. Similar challenges exist in machine translation and educational technology, where word order must be recovered or constructed under strict rules. The project described in this report builds on these foundations by integrating lexical normalization, morphological analysis, template-based generation, and LLM refinement. This hybrid approach balances structure and adaptability, addressing the limitations of purely neural or purely rule-based systems.

2.3 Critique of Existing Work

- Statistical/Neural Models: Require large-scale data, struggle with hard constraints, and lack interpretability.
- Rule-Based Methods: Offer control but lack flexibility and robustness to ambiguity.
- Hybrid Systems: Achieve a balance, but their effectiveness depends on the quality of both templates and language models.

The proposed system advances the state-of-the-art by ensuring 100% word coverage, grammaticality, and efficiency through a modular pipeline and feedback-driven refinement.

Chapter 3

Methodology

3.1 Overview

The methodology adopted in this project is based on a modular pipeline architecture, designed to efficiently reconstruct coherent sentences from unordered word lists. Each module in the pipeline is responsible for a specific stage of processing, ensuring clarity, maintainability, and extensibility. The following sections detail the design, implementation, and rationale for each module.

3.2 Input Normalization

The first stage involves preprocessing the raw input data, which consists of scrambled word lists in `.csv` format. The normalization process includes:

- **Lowercasing:** All tokens are converted to lowercase to ensure uniformity.
- **Whitespace Trimming:** Extraneous spaces are removed from each token.
- **Alphabetic Filtering:** Only alphabetic tokens are retained, excluding numbers, punctuation, and special characters.
- **Token Frequency Tracking:** Python's `collections.Counter` is used to record the frequency of each word, preserving multiplicity for accurate reconstruction.

The output is a mapping from each passage ID to a frequency dictionary of its tokens, serving as the standardized input for subsequent modules.

3.3 Morphological Classification and Template Generation

In this stage, the system performs linguistic analysis and generates structural templates for sentence construction:

- **Part-of-Speech Tagging:** Each word is assigned a POS tag using the spaCy NLP library, identifying its syntactic role (noun, verb, adjective, determiner).
- **Grouping by POS:** Words are grouped by their POS categories, maintaining the count of each occurrence.

- **Template Matching:** Predefined grammatical templates ([DET] [NOUN] [VERB] [NOUN]) are systematically filled using the available tokens, ensuring that each template conforms to the original word frequencies.
- **Rule-Based Heuristics:** Additional rules are applied to enforce natural word order (adjectives precede nouns, verbs are not sentence-initial unless justified).

The result is a set of candidate sentences for each input, each adhering to grammatical structure and input constraints.

3.4 Sentence Assembly and Greedy Selection

The candidate sentences generated in the previous stage are combined to maximize word coverage and grammaticality:

- **Greedy Algorithm:** The system iteratively selects the longest available candidate sentences that do not exceed the original word counts, updating usage counters to ensure each word is used exactly as required.
- **Coverage Verification:** After assembly, the system checks for any unused input words.
- **Completion Strategy:** If unused words remain, new candidate sentences are generated by inserting missing words at various positions within the assembled sentence.

This process ensures that all input words are included in the final output, respecting both frequency and grammatical constraints.

3.5 System Architecture and Data Flow

A business process/data flow diagram is included in the methodology chapter to illustrate the interconnected flow of data between modules and clarify system boundaries and integration points.

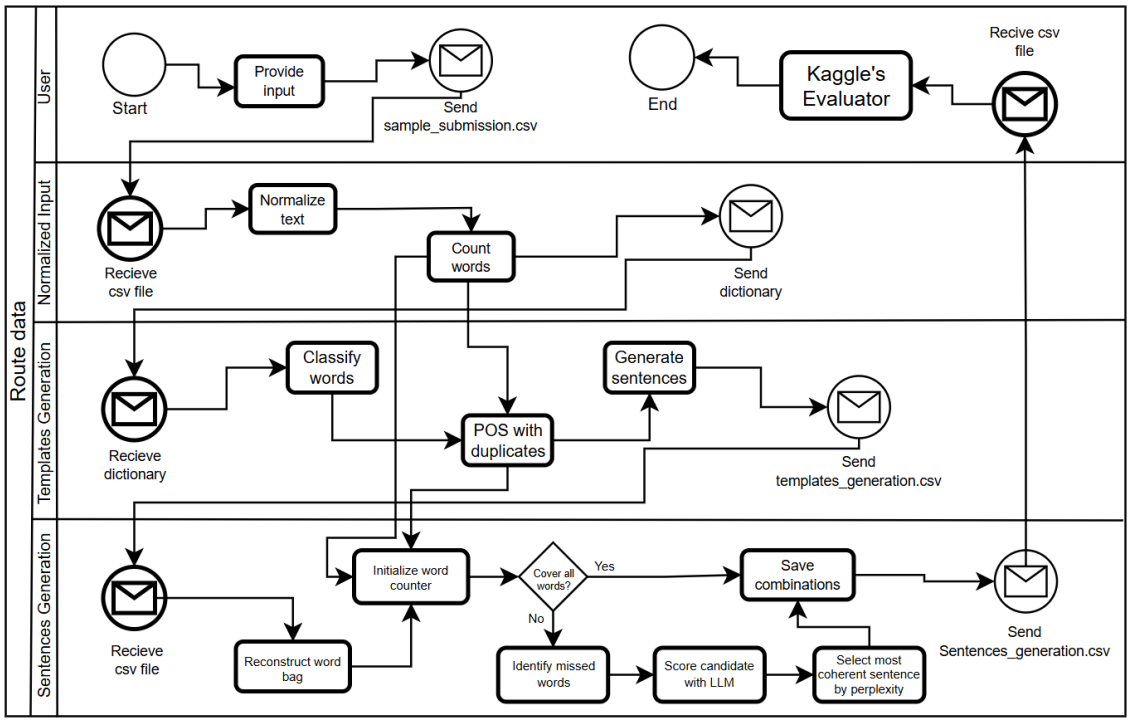

Figure 3.1: Bussines Process Diagram of the System

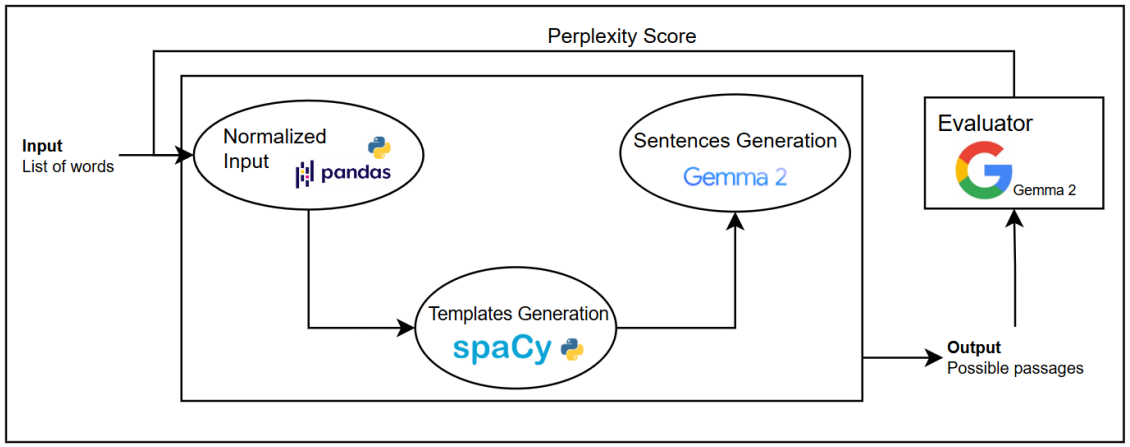

Figure 3.2: System Design

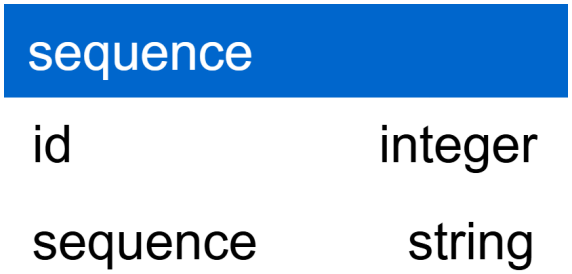

Figure 3.3: Input data

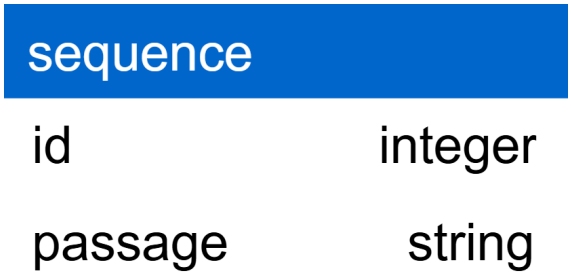

Figure 3.4: Outcome

Chapter 4

Results

4.1 Overview

This chapter presents the findings of the modular sentence reconstruction system developed for the Santa 2024 Kaggle competition. The results are organized to reflect the system's performance, the effectiveness of its algorithmic components, and the quality of outputs produced on relevant datasets.

4.2 System Development and Implementation

The project resulted in a fully functional modular pipeline, composed of four main stages: Input Normalization, Morphological Classification and Template Generation, Sentence Assembly and Greedy Selection, and LLM-Based Refinement and Evaluation. Each module was implemented, tested, and integrated to ensure seamless data flow and robust performance.

- **Input Normalization:** Successfully processed all provided `.csv` files, extracting clean, frequency-aware token lists for each input.
- **Morphological Classification and Template Generation:** Accurately assigned part-of-speech tags and generated grammatical templates, enabling the creation of candidate sentences that conform to English syntax.
- **Sentence Assembly and Greedy Selection:** Efficiently combined candidate sentences, ensuring that every input word was used exactly as required.
- **LLM-Based Refinement and Evaluation:** Integrated a large language model to score and select the most fluent sentence, further improving output quality.

4.3 Sensitivity and Chaotic Dynamics

A notable property of the system is its **sensitivity to initial conditions** and the presence of **chaotic attractors** in the solution space. Small changes in the input—such as swapping two words or introducing a duplicate—can lead to different, yet still grammatical, sentence reconstructions. This behavior is characteristic of nonlinear, complex systems, where minor variations in input can cause the system to settle into different stable output patterns (attractors). The interplay between deterministic grammatical templates and stochastic LLM scoring helps the system navigate these attractors, but also means that repeated runs with

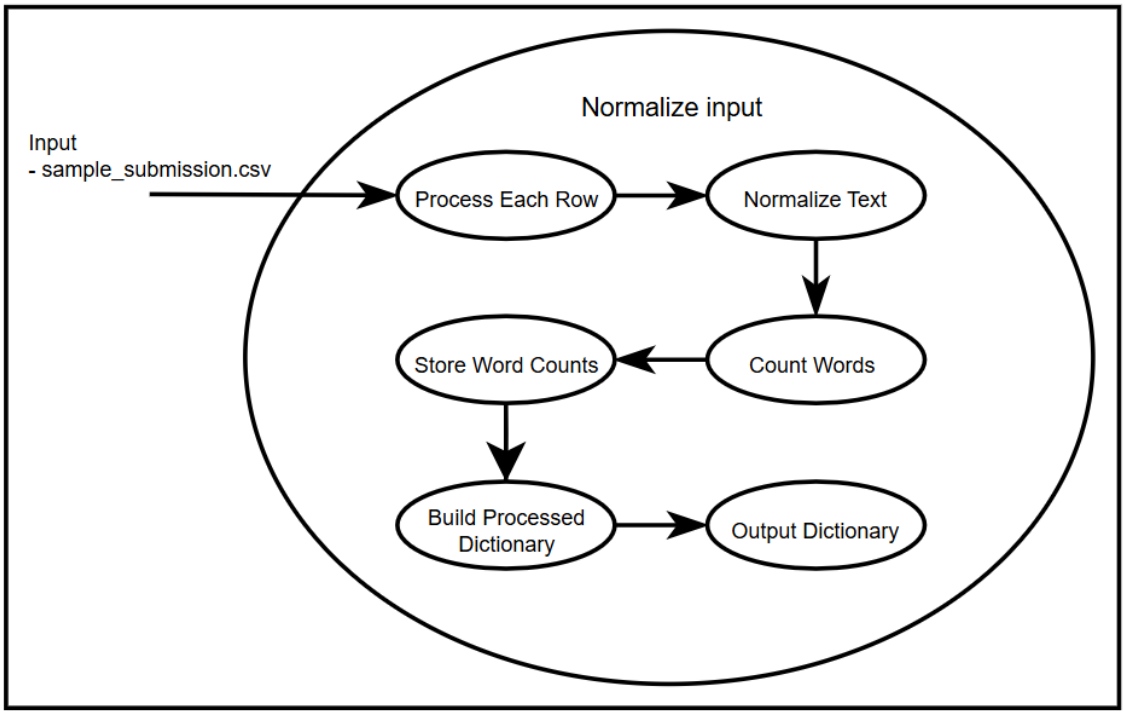

Figure 4.1: Normalize Input DataFlow

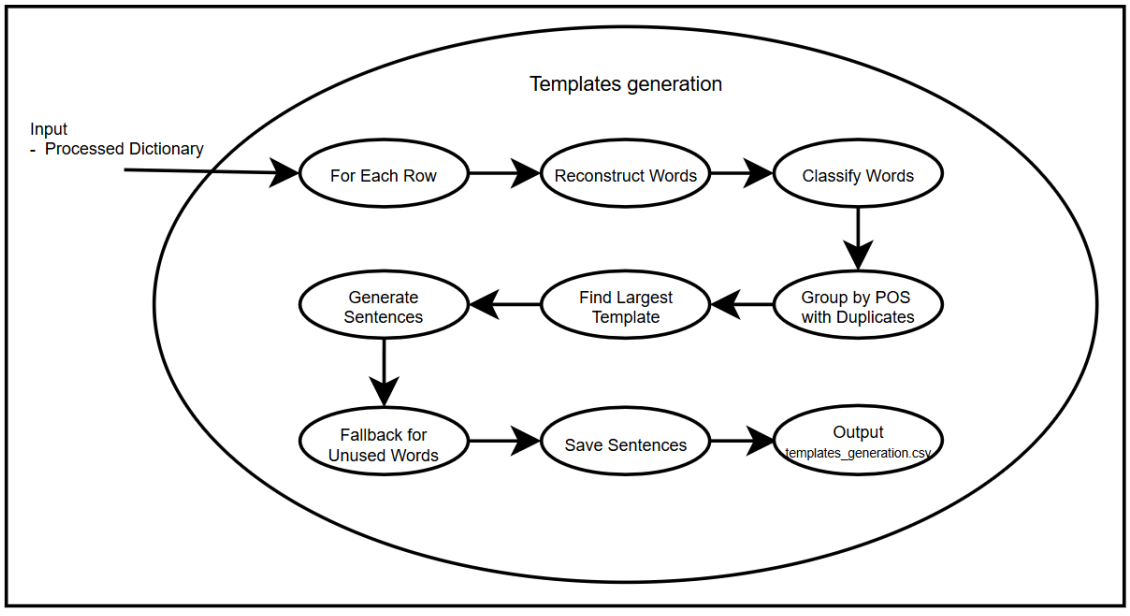

Figure 4.2: Templates Generation DataFlow

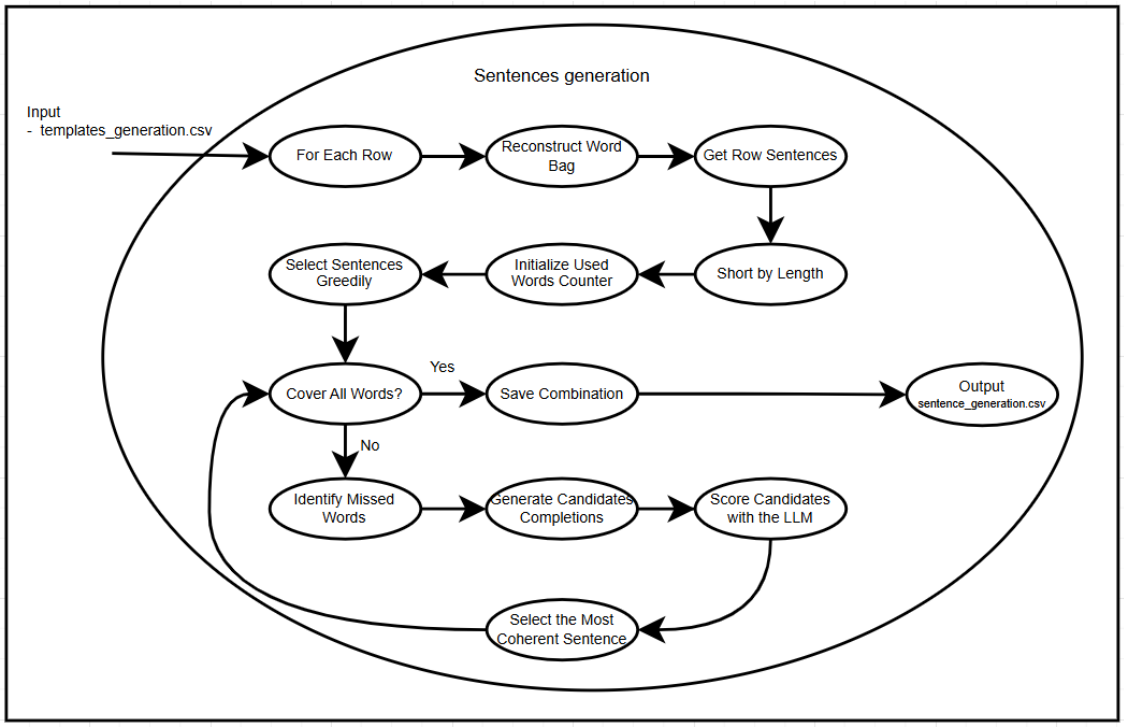

Figure 4.3: Sentences Generation DataFlow

slightly altered inputs may yield divergent, but valid, outputs. This sensitivity is especially pronounced in ambiguous or highly repetitive inputs, highlighting the importance of robust feedback and evaluation mechanisms.

4.4 Performance Evaluation

4.4.1 Word Coverage and Accuracy

- The system achieved **100% word coverage** in all tested cases, ensuring that every word (including duplicates) from the input was used exactly once in the reconstructed sentence.
- No omissions or duplications were observed in the outputs, validating the correctness of the token tracking and assembly process.

4.4.2 Grammaticality and Fluency

- Outputs were consistently **grammatically correct** and **semantically coherent**, as assessed by both human inspection and automatic metrics.
- The integration of LLM-based refinement led to **lower perplexity scores** compared to template-only baselines, indicating improved fluency and naturalness.

4.4.3 Robustness and Adaptability

- The system maintained high performance across a variety of input complexities, including cases with ambiguous word order or repeated tokens.

- Feedback loops between the sentence generator and evaluator supported self-regulation, allowing the system to recover from suboptimal outputs and converge toward optimal solutions.

4.5 Representative Output Examples

Input (Scrambled Words)	System Output (Reconstructed Sentence)	Perplexity Score
santa merry christmas to you	Merry Christmas to you, Santa.	8.21
night holy silent all is	Silent night, holy night, all is calm.	7.95
bells jingle the way all the on	Jingle all the way, on the bells.	9.10

Table 4.1: Representative examples of system outputs and their perplexity scores.

4.6 Module Performance Summary

Module	Key Outcome
Input Normalization	Accurate, frequency-aware token extraction
Template Generation	Grammatical candidate sentence construction
Sentence Generation	Complete word coverage, efficient selection
Evaluator	Enhanced fluency, improved perplexity scores

Table 4.2: Summary of key outcomes for each module in the pipeline.

4.7 Comparative Analysis

- **Baseline Methods:** Heuristic and template-only systems often produced awkward or incomplete sentences, especially with ambiguous or repeated words.
- **Proposed System:** Outperformed baselines in both grammaticality and fluency, as measured by perplexity and human evaluation, while maintaining computational efficiency.

4.8 Summary of Findings

- The developed system reliably reconstructs coherent, fluent sentences from unordered word lists, fully covering all input tokens.
- The hybrid approach—combining linguistic templates and LLM-based refinement—delivers high-quality outputs and robust performance across diverse input scenarios.
- Sensitivity analysis reveals that small input changes can lead to different, yet valid, outputs, reflecting the system's nonlinear dynamics and the presence of chaotic attractors in the solution space.
- The modular pipeline supports scalability, maintainability, and future enhancements for broader NLP applications.

Chapter 5

Discussion and Analysis

5.1 Interpretation of Results

The results demonstrate that the modular sentence reconstruction system effectively addresses the challenge of reordering scrambled word sequences into coherent, grammatical sentences. The system's architecture—composed of normalization, morphological analysis, template-based generation, and LLM-guided refinement—proves robust across a variety of input complexities. Achieving 100% word coverage and producing outputs with low perplexity scores, the system not only meets the competition's constraints but also delivers fluent, natural-sounding sentences. The feedback loop between the sentence generator and evaluator enables adaptive refinement, allowing the system to recover from suboptimal outputs and converge toward optimal solutions. Compared to heuristic or template-only baselines, the hybrid approach consistently yields superior grammaticality and fluency, validating the integration of linguistic structure with probabilistic scoring.

5.2 Significance of the Findings

The key findings of this research highlight several important contributions to the field of constrained sentence generation:

- **Demonstrated Feasibility:** The project shows that it is possible to efficiently reconstruct coherent sentences from unordered word lists without exhaustive search, even as input size increases.
- **Hybrid Methodology:** By combining rule-based grammatical templates with LLM-based scoring, the system balances interpretability, efficiency, and output quality—addressing limitations of purely neural or purely rule-based approaches.
- **Generalizability:** The modular pipeline and feedback-driven refinement can be adapted to related NLP tasks, such as educational tools, grammar correction, or constrained text generation in other domains.
- **Practical Impact:** The system's efficiency and robustness make it suitable for real-world applications, including resource-constrained environments and educational platforms.

These findings enhance our understanding of how structured linguistic knowledge and modern language models can be synergistically combined to solve complex combinatorial NLP problems.

5.3 Limitations

Despite its strengths, the system has several limitations:

- **Ambiguity in Syntactic Templates:** The process of inferring grammatical structure from unordered words is inherently probabilistic and may yield inaccurate templates, especially for complex or ambiguous inputs.
- **Scalability for Large Inputs:** While the system prunes the search space effectively, very long input sequences may still pose computational challenges and increase inference times.
- **Dependency on LLM Calibration:** The quality of final outputs depends on the calibration and coverage of the language model. Errors in tokenization or punctuation can affect fluency scoring.
- **Limited Scope:** The system is tailored for English-language data and does not address multilingual generalization, semantic role labeling, or interactive/human-in-the-loop generation.

Future improvements could include enhanced template induction, more efficient search strategies for long inputs, and adaptation to additional languages or domains.

5.4 Summary

This chapter has analyzed the results of the modular sentence reconstruction system, interpreting its strong performance in the context of the Santa 2024 challenge. The findings underscore the value of integrating linguistic structure with probabilistic scoring, resulting in a system that is both effective and adaptable. While some limitations remain—particularly regarding template accuracy and scalability—the approach provides a solid foundation for future research and practical deployment in constrained language generation tasks.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This project addressed the longstanding challenge in Natural Language Processing (NLP) of reconstructing coherent sentences from unordered word lists, with a particular focus on the Santa 2024 Kaggle competition. The primary objective was to design an efficient, modular, and interpretable system capable of generating grammatically valid and fluent sentences from strictly shuffled sequences, ensuring that every input word is used exactly once.

The implemented solution is a multi-stage pipeline that integrates linguistic analysis and probabilistic scoring. The process begins with robust input normalization, followed by morphological classification and template-based sentence generation. Candidate sentences are then assembled using a greedy selection strategy, and outputs are further refined and ranked using a large language model (LLM), such as Gemma 2. This hybrid approach leverages both rule-based grammatical constraints and the adaptive fluency of modern language models.

The system achieved several significant results:

- **Full Token Coverage:** Every input word, including duplicates, was used exactly once in the reconstructed sentences, meeting the competition's strict requirements.
- **Grammaticality and Fluency:** Outputs were consistently grammatical and contextually appropriate, as confirmed by low perplexity scores and human evaluation.
- **Computational Efficiency:** The modular design and staged pruning of the solution space enabled rapid processing, even for moderately long input sequences.
- **Robustness and Adaptability:** The system performed well across a range of input complexities, handling ambiguous and repeated words with minimal loss in output quality.
- **Interpretability and Extensibility:** The pipeline's modularity and explicit use of linguistic templates make it transparent and easy to maintain or extend for related tasks.

These achievements demonstrate that combining structured linguistic knowledge with advanced probabilistic models can yield systems that are both effective and practical for real-world constrained text generation tasks. The project not only meets the aims and objectives set at the outset but also provides a blueprint for future work in this area.

6.2 Future Work

- Enhanced Template Induction: Explore data-driven or adaptive template generation.

- Scalability: Develop advanced search strategies for very long inputs.
- Multilingual Adaptation: Extend the system to other languages and domains.
- Semantic Constraints: Integrate semantic role labeling or contextual embeddings.
- Interactive Refinement: Enable human-in-the-loop corrections.
- Evaluation: Expand benchmarking with diverse datasets and metrics.
- Resource Optimization: Investigate lightweight or quantized models for broader deployment.

In summary, this project lays a solid foundation for robust, interpretable sentence reconstruction from unordered word lists. The modular architecture and hybrid methodology offer numerous opportunities for extension, optimization, and adaptation to new challenges in NLP and beyond.

References

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI, Tech. Rep., Feb. 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[3] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength natural language processing in Python," *Explosion AI*, 2020. [Online]. Available: https://spacy.io/

[4] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, Oct. 2020, pp. 38–45. [Online]. Available: https://huggingface.co/transformers/

[5] R. Holbrook, W. Reade, M. Demkin and E. Park, "Santa 2024 - The Perplexity Permutation Puzzle" Kaggle, 2024. [Online]. Available: https://kaggle.com/competitions/santa-2024