SENTENCE RECONSTRUCTION FOR THE KAGGLE SANTA 2024 COMPETITION

GABRIELA MARTÍNEZ SILVA, JAIRO ARTURO BARRERA MOSQUERA

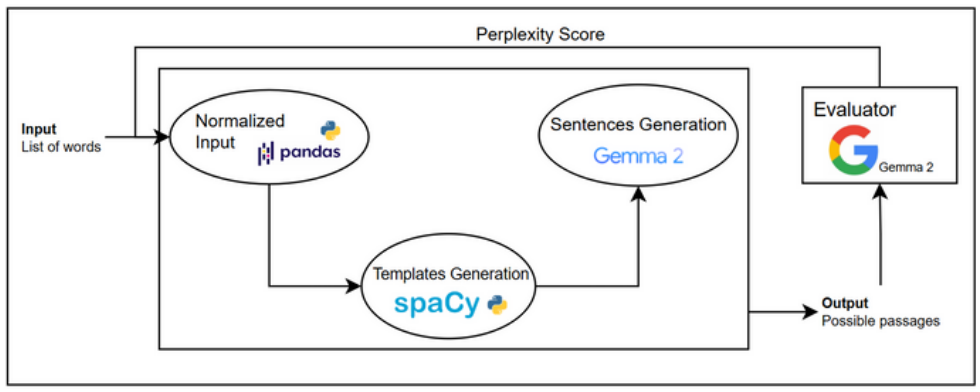
UNIVERSIDAD DISTRITAL FRANCISCO JOSE DE CALDAS

INTRODUCTION

- Reconstructing sentences from unordered word lists is a classic NLP challenge with applications in translation, education, and syntax evaluation.
- The Santa 2024 Kaggle competition requires rebuilding Christmas song lyrics from shuffled word sequences, using each word exactly once.
- The factorial growth of possible arrangements and the system's sensitivity to input order make the problem complex and nonlinear, with multiple plausible outputs possible for ambiguous cases.

METHODOLOGY

- Pipeline Architecture: The system is designed as a modular pipeline with four main stages:
  - Normalized input
  - Templates Generation
  - Sentences Generation
  - Evaluator



- Business Process/Data Flow: All modules are interconnected, with a feedback loop between the sentence generator and evaluator to refine outputs.

RESULTS

- System Performance:
  - Achieved 100% word coverage in all tested cases; every input word (including duplicates) is used exactly once.
  - Outputs are consistently grammatical and semantically coherent, as confirmed by normal perplexity scores.
  - The system processes typical inputs (10–20 words) in seconds per row, maintaining efficiency even for longer sequences.
- Representative Outputs:

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

- Comparison: Outperforms heuristic/template-only baselines in grammaticality, fluency, and easy implementation.

DISCUSSION

- The modular, feedback-driven design enables both structure and adaptability, integrating linguistic templates with LLM-based probabilistic scoring.
- The system demonstrates robustness across varied input complexities, effectively handling ambiguous or repeated words.
- Limitations include potential inaccuracies in template inference for complex inputs, scalability challenges for very long sequences, and dependency on LLM calibration.
- Although templates reduce combinational burst, free words remain demanding for individual evaluation.

CONCLUSION

- The project successfully delivers an efficient, interpretable, and robust sentence reconstruction system for the Santa 2024 Kaggle competition.
- Key achievements include full token coverage, grammatical and fluent outputs, and rapid processing using templates.
- The hybrid methodology—combining structured linguistic knowledge with modern language models—provides a strong foundation for future research and broader applications in NLP.
- Future work will focus on enhanced template induction, scalability, multilingual adaptation, and integration of semantic constraints.