
ITESM Campus Monterrey

# Proof of Concept: Enhancing Language Understanding with Modern NLP

Luis Alberto Portilla López
August 16th, 2024

Research Stay - Going beyond Artificial Intelligence: Artificial Emotions

TC3073 | Group 573

# Introduction

Natural Language Processing (NLP) plays a crucial role in how AI systems understand and interact with human language. Traditional methods like N-grams have been foundational, but they often fall short when dealing with complex language patterns, context, and rare words. With the rise of neural networks, there's an opportunity to enhance these traditional methods to build more robust language models that can better grasp the intricacies of human communication.

This Proof of Concept (PoC) aims to combine these older methods with modern techniques to improve the accuracy and efficiency of language understanding in AI systems. By doing so, we seek to create models that can more effectively process both common and rare language patterns, maintain context over long texts, and do so efficiently enough to be used in real-world applications.

## Business Problem

As AI becomes integral to business operations, especially in areas like customer service, content generation, and data analysis, the limitations of current NLP models become more apparent:

1. **Contextual Understanding:** Current models often struggle to maintain the context across sentences, leading to misinterpretations in tasks like summarization and conversation generation.

2. **Rare Words and Domain-Specific Language:** Traditional models have difficulty accurately processing rare or domain-specific terms, which can be critical in specialized fields like medicine or law.

3. **Handling Large Texts:** When working with longer documents, AI models often lose track of the broader context, resulting in less coherent outputs.

4. **Resource Demands:** Modern NLP models, especially deep learning-based ones, require significant computational resources, which can be costly and time-consuming.

These challenges lead to AI applications that may fail to meet user expectations, particularly in specialized or high-stakes environments where accuracy is paramount.

# Proposed Solution

To address these challenges, this PoC proposes a hybrid approach that integrates traditional N-gram techniques with advanced neural network models:

1. **Combining N-grams and Neural Networks:**

    • N-grams will be used for their strength in capturing frequent patterns in language, which is essential for tasks like language modeling and speech recognition.
    • Neural networks, particularly those based on recurrent and transformer architectures, will be employed to handle the nuances of context and rare word interpretation.

2. **Subword Tokenization:**

    • Implementing subword tokenization techniques (such as Byte-Pair Encoding or WordPiece) will help in breaking down rare or complex words into smaller, more manageable units. This allows the model to better understand and generate text involving uncommon terminology.

3. **Contextual Embeddings:**
    • Leverage models like BERT or GPT that can maintain context across longer sequences of text. This will be critical in improving the coherence and relevance of outputs, especially in tasks that require a deep understanding of context, such as summarization or dialogue generation.

4. **Distributed Computing for Efficiency:**

    • Utilize distributed computing frameworks, such as Apache Spark, to manage the computational load. This will enable the model to process large datasets more efficiently, making the solution scalable for real-world applications.

# Expected Outcomes

By integrating these techniques, the proposed solution aims to achieve the following outcomes:

- **Improved Accuracy in Language Understanding:** The hybrid approach is expected to provide a more nuanced understanding of both common and rare language patterns, leading to more accurate text processing and generation.
- **Better Context Management:** By using contextual embeddings, the system will be better at maintaining and utilizing context across longer texts, improving the quality of tasks like summarization and conversation.
- **Efficient Processing of Large Datasets:** The use of distributed computing will ensure that the model can handle large volumes of data efficiently, reducing processing time and computational costs.

# Conclusion

This PoC outlines a practical and advanced approach to enhancing NLP models by combining the strengths of traditional methods with the capabilities of modern neural networks. By addressing key challenges such as context management, rare word handling, and computational efficiency, this approach aims to create NLP models that are more accurate, robust, and scalable for real-world applications.

# References


1. Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. Computer Speech & Language, 13(4), 359-394.

2. Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 746-751).

3. Fonseca, J., Freitas, A., & Carvalho, J. (2022). Building Wikipedia N-grams with Apache Spark. In 2022 International Conference on Information Networking (ICOIN) (pp. 589-594). IEEE.

4. Wang, W., Tao, J., & Gao, Y. (2021). From N-gram-based to Neural Language Models: Developments in Half a Century. Engineering, 7(9), 1235-1251.