N-gram Models in Language Processing: Their Strengths, Limitations, and Future Directions

José Martín Véliz Zambrano.
Tecnológico de Monterrey, Querétaro, México.

Abstract

Language processing systems have heavily utilized n-gram models due to their simplicity and effectiveness. While advancements in neural networks suggest a shift towards complex modelling, there remains an undervalued potential in n-gram models. This review examines recent research on the utility of n-grams in language modelling by addressing their strengths and limitations and reflecting on future implications and directions.

1 Introduction

Research in language processing has largely centred on n-gram models, which consider the probability of a word based on a fixed number of preceding words. While the power of neural language models has introduced fresher, more complex perspectives, n-gram models continue to be relevant ([2]Shareghi et al.). This review looks at recent research on the utility of sophisticated linguistic knowledge, n-gram models, distributed processing, language modelling in speech recognition, and application in smart home automation.

2 Survey

The need to incorporate sophisticated linguistic knowledge into language modelling has been a point of continuous exploration. In the study by [1]Brill et al., it becomes evident that humans have the potential to significantly enhance speech recognition models' outputs using proficient closed-class word choices and world knowledge, among others. This emphasizes the necessity of integrating intricate linguistic information into language models.

To assert the importance of n-gram models, [2]Shareghi et al. tackled the comparative performance of n-gram models versus LSTM models for 42 languages. Their findings duly noted that despite the successes credited to neural models, n-gram models' potential has been left untapped. The research compounded the argument for the consideration of n-gram models, particularly in resource-constrained or morphologically rich languages.

Nevertheless, while n-gram models remain valuable, their traditional development has limitations in scalability and efficiency. The study by the [3]authors who introduced a distributed data processing approach using Apache Spark substantiates this argument. By building an N-gram corpus from an English Wikipedia archive, they brought about a scalable, robust, and inexpensive alternative to the convention of using single-machine databases, consequently indicating the suitability of distributed techniques for handling large-scale datasets.

Subsequently, in the realm of Automatic Speech Recognition (ASR), the investigation into the use of pre-trained language models (PLMs) carried out by a team of [4]researchers presented GPT-2 as a beneficial resource when converted into an n-gram model through text sampling and probability conversion. Therefore, in low data scenarios, pre-trained language models (like GPT-2) displayed considerable potential, though their performance dwindled as data volumes increased, suggesting a still strong place for traditional trigram models.

Finally, the unique application of n-gram language modelling in the smart home domain, as depicted in a [5]study on a system called Helion, underscores its broad relevance and dynamic applicability. The system enabled the generation of realistic home automation scenarios for testing purposes, which was a marked improvement from the conventional, random event permutations. The extracted success points to the possibility of using advanced n-gram modelling in innovative fields to solve emerging challenges.

3 Discussion and Future Directions

This integrated review underscores the considerable promise n-gram models still hold despite advances in language modelling technologies. The models offer consistent and reliable outcomes, especially in resource-limited settings and where complex morphological language processing is required. Nonetheless, improvements are necessary to handle large-scale datasets or integrate sophisticated linguistic knowledge beneficial for systems such as ASR.

The future seems ripe for a thorough exploration of advanced or hybrid n-gram models. There is potential in leveraging these models in new domains and research areas, such as smart home automation. Additionally, further studies to enhance and accelerate the application of n-gram models in ASR using PLMs or other technologies should be encouraged. Lastly, the development and evolution of processing technologies, such as Apache Spark, for n-gram models, opens up prospects for increased speed, scalability, and efficiency, allowing for robust handling of larger datasets.

4 Conclusion

The pervasive utility and relevance of n-gram models in language processing are well established. Despite advancements in neural networks, the future still holds intriguing possibilities for n-gram models, particularly in areas where resources are limited or linguistic characteristics are complex. The call is toward improved integration of linguistic knowledge and world understanding into language models, thorough investigation of distributed processing, adoption of cutting-edge language models such as GPT-2 in speech recognition, and exploring applications in emerging fields such as smart home automation. The overarching goal is to refine and expand the breadth and depth of n-gram models' application, pushing the boundaries of language modelling and processing.

5 References

[1] Brill, E., Florian, R., Henderson, J.C., Mangu, L. (Date Unknown). Beyond N-Grams: Can Linguistic Sophistication Improve Language Modeling? Department of Computer Science, Johns Hopkins University, Baltimore, Md. 21218 USA.

[2] Ehsan Shareghi, Daniela Gerz, Ivan Vulić, Anna Korhonen. "Show Some Love to Your *n*-grams: A Bit of Progress and Stronger *n*-gram Language Modeling Baselines." Language Technology Lab, DTAL, University of Cambridge.

[3] Esmaeilzadeh, A., Fonseca Cacho, J.R., Taghva, K., Nojoo Kambar, M.E.Z. & Hajiali, M. (2022). Building Wikipedia N-grams with Apache Spark. In K. Arai (Ed.): SAI 2022, LNNS 507, pp. 672–684. Springer Nature Switzerland AG.

[4] Krishnan, A., Alabi, J. O., Klakow, D. (2023) "On the N-gram Approximation of Pre-trained Language Models", Saarland University, Germany.

[5] Mandal, P., Manandhar, S., Kafle, K., Moran, K., Poshyvanyk, D., & Nadkarni, A. (2023). Helion: Enabling Natural Testing of Smart Homes. In Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE ’23), December 3–9, 2023, San Francisco, CA, USA.