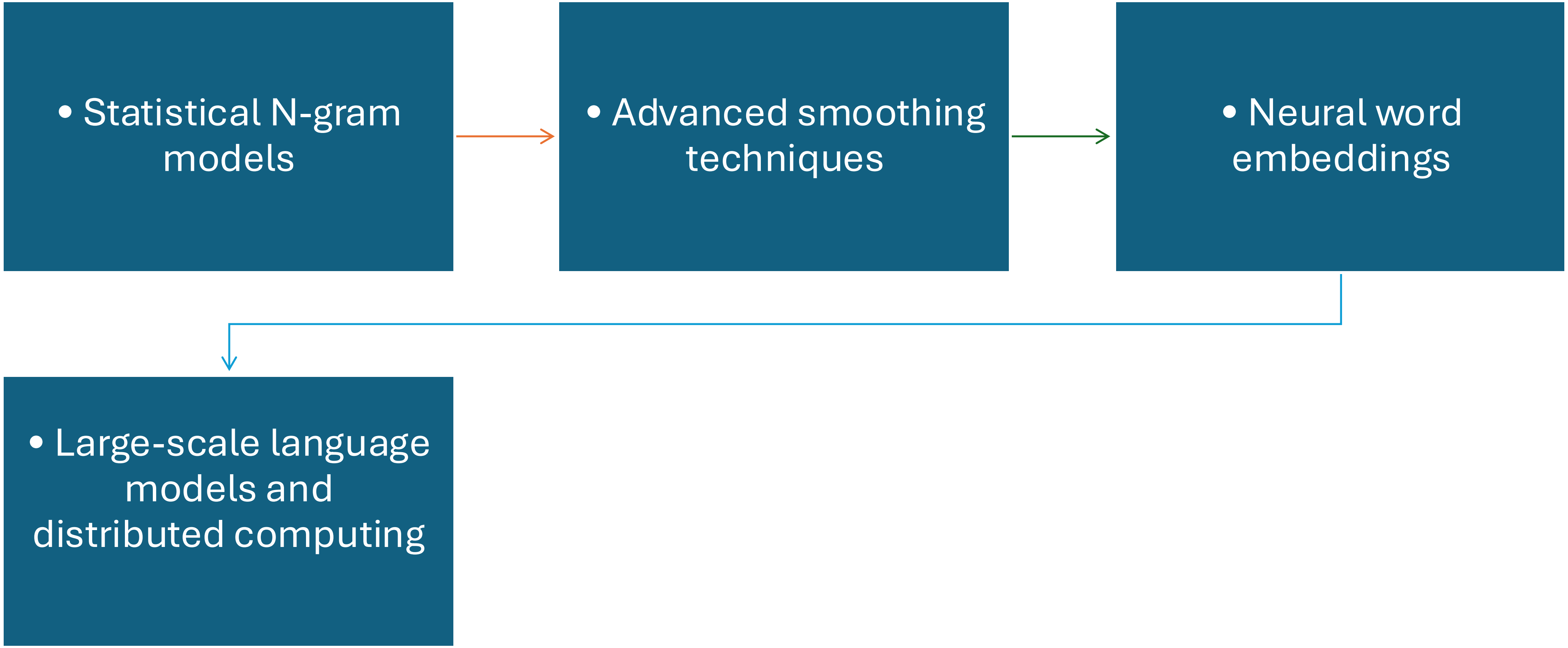
Tecnológico de Monterrey

Language modeling and N-grams: From Statistical Foundations to Neural Advancements

Week 3 Mini Survey

Luis Alberto Portilla López

The Evolution of Language Modeling



- Statistical N-gram models
- Advanced smoothing techniques
- Neural word embeddings
- Large-scale language models and distributed computing

Current Challenges

Balancing model complexity with computational efficiency

Handling of rare words and out-of-vocabulary terms

Incorporating long-range dependencies in language

Search Methodology & Criteria

CITATION CHAINING AND FORWARD CITATION

KEYWORD SEARCH

BOOLEAN SEARCH

SNOWBALLING

CRITERIA:

- Initial review of abstracts to assess relevance based on the title, publication venue, and year.
- Direct and indirect relevance to the paper being cross-referenced through the abstract.
- Consideration of the number of citations and field-weighted citation impact (fwci), a metric that measures the citation impact of a paper adjusted for disciplinary differences.

Preliminary Terms

🔑 **Key terms identified during the week:**

- Word Embedding
- Language Model
- Smoothing
- N-gram

Document Comparison

“An Empirical Study of Smoothing Techniques for Language Modeling”

“Linguistic Regularities in Continuous Space Word Representations”

“Building Wikipedia N-grams with Apache Spark”

“The Role of n-gram Smoothing in the Age of Neural Networks”

References

1. Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. Computer Speech & Language, 13(4), 359-394. https://doi.org/10.1006/csla.1999.0128

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. https://doi.org/10.1109/ICOIN53446.2022.9687193

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. https://doi.org/10.1016/j.eng.2021.03.023