Natural Language Generation: Algorithms and Applications

Diwakar R. Tripathi [1*], Abha Tamrakar [2]

[1*] Assistant Professor, Faculty of Science, ISBM University, Gariyaband, Chhattisgarh, India.
[2]Assistant Professor, Faculty of Science, ISBM University, Gariyaband, Chhattisgarh, India.
*Corresponding Author:diwakar.tripathi@isbmuniversity.edu.in

Abstract: Natural Language Generation (NLG) is a subfield of artificial intelligence and computational linguistics that focuses on the automatic generation of natural language text. NLG has a wide range of applications in various fields, including content generation, virtual assistants, business intelligence, and healthcare. This paper provides an overview of NLG techniques and algorithms, including rule-based NLG, template-based NLG, statistical NLG, and neural NLG. It also explores the applications of NLG in different fields, highlighting its role in automated journalism, personalized content creation, virtual assistants, and data storytelling. Furthermore, the paper discusses the current challenges in NLG, such as naturalness, ambiguity handling, and scalability, and examines emerging trends and future directions in NLG, including advancements in neural NLG models, integration with other AI technologies, and ethical considerations. Overall, this paper aims to provide a comprehensive understanding of NLG and its impact on modern society.

Keywords: Natural Language Generation, NLG Techniques, NLG Applications, NLG Challenges, NLG Trends.

I. Introduction

A. Definition of Natural Language Generation (NLG)

Natural Language Generation (NLG) refers to the process of automatically producing human-readable text or speech from structured data or other forms of input. NLG systems analyze input data, understand its meaning, and generate coherent and contextually appropriate language output. NLG encompasses various techniques, ranging from rule-based systems to advanced deep learning models, all aimed at transforming structured information into natural language form.
NLG has garnered significant attention in both academia and industry due to its potential to automate content generation tasks, enhance human-computer interaction, and facilitate communication in various domains.

B. Importance and relevance of NLG in various fields

NLG holds immense importance and relevance across diverse fields, including but not limited to journalism, healthcare, business intelligence, and virtual assistants. In journalism, NLG systems have been employed to automatically generate news articles from structured data, thereby increasing the efficiency of content creation and dissemination (Liu et al., 2018). Similarly, in healthcare, NLG is utilized for generating patient reports and personalized health recommendations based on clinical data, improving communication between healthcare providers and patients (Du et al., 2016).

Moreover, NLG plays a crucial role in business intelligence and analytics by automatically generating textual summaries and insights from large datasets, enabling stakeholders to make informed decisions (Gkatzia et al., 2015). In the context of virtual assistants and chatbots, NLG facilitates natural and engaging interactions by generating human-like responses to user queries, enhancing user satisfaction and usability (Wen et al., 2015).

II. Background and History of NLG

A. Early developments in NLG

The field of Natural Language Generation (NLG) traces its roots back to the 1970s when researchers began exploring the possibility of automating the generation of human language. Early efforts focused on rule-based approaches, where linguistic rules were used to transform input data into coherent sentences. One of the earliest systems, the "SHRDLU" program developed by Terry Winograd in 1972, demonstrated the generation of English sentences to describe simple block world scenarios (Winograd, 1972).

B. Key milestones and breakthroughs

Over the years, NLG has witnessed several key milestones and breakthroughs that have significantly advanced the field. In the 1980s and 1990s, researchers began incorporating statistical methods into NLG systems, leading to the development of more data-driven approaches (Langkilde & Knight, 1998). This shift marked a fundamental change in NLG, moving away from handcrafted rules to models that could learn patterns from data.

Another significant milestone was the development of the first commercial NLG systems in the late 1990s and early 2000s. Companies like Narrative Science and Automated Insights pioneered the use of NLG for generating personalized reports and stories from data (Lahiri & Reddy, 2011).

C. Evolution of NLG techniques and algorithms

The evolution of NLG techniques and algorithms has been characterized by a move towards more sophisticated and data-driven approaches. In recent years, the advent of deep learning has revolutionized NLG, with models like recurrent neural networks (RNNs) and transformers achieving state-of-the-art performance in various NLG tasks (Vaswani et al., 2017).

Moreover, the integration of NLG with other AI technologies, such as natural language understanding (NLU) and dialogue management, has led to the development of more interactive and context-aware NLG systems (Mei et al., 2016). These advancements have enabled NLG to be applied in a wide range of domains, from virtual assistants and chatbots to automated content generation and data analytics.

III. NLG Techniques and Algorithms
A. Rule-based NLG
Description: Rule-based NLG systems operate on a set of predefined linguistic rules that govern the transformation of input data into natural language output. These rules typically encode syntactic and semantic patterns to ensure the generated text is grammatically correct and coherent.
Examples: One example of a rule-based NLG system is SimpleNLG, which is an open-source Java library for generating natural language text from structured data (Gatt et al., 2009). Another example is the RealPro NLG system, which is used for generating weather forecasts (Belz & Reiter, 2006).
Advantages and limitations: Rule-based NLG systems are relatively easy to understand and modify, making them suitable for domains where linguistic rules are well-defined. However, they can be limited in their ability to handle complex linguistic phenomena and may require extensive manual effort to create and maintain rules for different languages and domains.

Table 1: Comparison of NLG Techniques

NLG Technique	Description	Examples	Advantages	Limitations
Rule-based NLG	Uses predefined rules for text generation	SimpleNLG, RealPro	Easy to understand and modify	Limited in handling complex language and scenarios
Template-based NLG	Uses templates with placeholders for variables	Madamira, SimpleNLG Realizer	Simple to implement	Limited in generating varied language output
Statistical NLG	Uses statistical models trained on data	Machine translation, Text summarization	Can generate more varied and natural-sounding text	May struggle with coherence and context
Neural NLG	Uses neural networks to generate text	GPT-3, BERT, T5	Can generate highly fluent text	Requires large amounts of data and computational resources

B. Template-based NLG
Description: Template-based NLG systems use predefined templates that contain placeholders for variables. These templates are then filled in with specific values from the input data to generate natural language output.
Examples: An example of a template-based NLG system is the Madamira system, which is used for generating Arabic text from morphologically analyzed input (Pasha et al., 2014). Another example is the SimpleNLG Realizer, which uses templates to generate text in multiple languages (Gatt et al., 2009).
Advantages and limitations: Template-based NLG systems are straightforward to implement and can be effective for generating simple, repetitive text. However, they may struggle with generating varied and nuanced language output, as they are limited by the predefined templates.

C. Statistical NLG

Description: Statistical NLG systems use statistical models, such as n-gram language models or machine learning algorithms, to learn patterns from data and generate natural language output. These models are trained on large corpora of text to predict the most likely next word or phrase given the context.
Examples: Statistical NLG has been used in various applications, such as machine translation (Koehn et al., 2003) and text summarization (Nenkova & McKeown, 2011).
Advantages and limitations: Statistical NLG systems can generate more varied and natural-sounding text compared to rule-based or template-based systems. However, they may struggle with generating coherent and contextually appropriate output, especially in complex or ambiguous scenarios.

D. Neural NLG
Description: Neural NLG systems use neural networks, such as recurrent neural networks (RNNs) or transformers, to generate natural language output. These models are trained on large datasets of text to learn the underlying patterns of language.
Examples: Neural NLG has been applied in various tasks, including text generation (Radford et al., 2019) and machine translation (Vaswani et al., 2017).
Advantages and limitations: Neural NLG systems can generate highly fluent and contextually relevant text, often outperforming traditional NLG approaches. However, they require large amounts of training data and computational resources, and their output can be challenging to interpret and control.

IV. Applications of NLG
A. NLG in Content Generation
Automated journalism: NLG is used in automated journalism to generate news articles from structured data, such as sports scores or financial reports. These systems can produce high volumes of news content quickly and efficiently (Dongaonkar et al., 2019).

Data-to-text generation: NLG is employed to convert structured data, such as statistical data or database entries, into natural language text. This is particularly useful for generating reports and summaries from large datasets (Gardent et al., 2017).
NLG for personalized content creation: NLG can be used to generate personalized content, such as product recommendations or marketing messages, based on user preferences and behavior (Arora et al., 2016).

B. NLG in Virtual Assistants and Chatbots
Conversational NLG: NLG is used in virtual assistants and chatbots to generate responses to user queries in natural language. These systems aim to simulate human-like conversations and provide helpful responses (Serban et al., 2015).
Task-oriented NLG: NLG can be used in task-oriented virtual assistants to generate instructions or explanations for completing tasks, such as booking a hotel room or ordering food (Bordes et al., 2017).
NLG for natural and engaging interactions: NLG is employed to make virtual assistants and chatbots more engaging by generating diverse and contextually relevant responses (Higashinaka et al., 2014).

C. NLG in Business Intelligence and Analytics
NLG for report generation: NLG is used in business intelligence to automatically generate reports from data, providing insights and summaries for decision-makers (Gkatzia et al., 2015).
NLG for summarization and insights extraction: NLG can be used to summarize large volumes of data and extract key insights, helping businesses make sense of complex information (Zhang et al., 2018).
NLG for data storytelling: NLG is employed to turn data into compelling narratives, helping to communicate insights and trends effectively (Swartout et al., 2017).

D. NLG in Healthcare
NLG for medical reports and documentation: NLG is used in healthcare to generate medical reports, discharge summaries, and other documentation, reducing the burden on healthcare providers (Kreuzthaler et al., 2018).
Patient communication and education: NLG can be used to generate patient-friendly explanations of medical conditions, treatments, and procedures, improving patient understanding and compliance (Arnold et al., 2016).
NLG for personalized healthcare recommendations: NLG can be employed to generate personalized healthcare recommendations based on patient data and medical guidelines, helping to improve patient outcomes (Zhou et al., 2017).

V. Challenges and Future Directions
A. Current challenges in NLG

Naturalness and coherence: One of the primary challenges in NLG is ensuring that generated text is natural-sounding and coherent. NLG systems often struggle to produce language that mimics human fluency and coherence, especially in complex or ambiguous scenarios (Novikova et al., 2017).
Handling ambiguity and context: NLG systems face difficulties in understanding and representing ambiguous or context-dependent language. Resolving ambiguity and incorporating context appropriately remain significant challenges in achieving more accurate and contextually relevant text generation (Liu et al., 2019).
Scalability and efficiency: As NLG systems become increasingly sophisticated and are applied to larger datasets and more complex tasks, scalability and efficiency become critical concerns. Ensuring that NLG models can handle large volumes of data and generate text in real-time without sacrificing quality is a significant challenge (Dathathri et al., 2020).

Table 2: Challenges in NLG

Challenge	**Description**	**Potential Solutions or Approaches**
Naturalness and Coherence	Ensuring that generated text is natural-sounding and coherent, mimicking human fluency	Advanced neural NLG models, incorporating context and discourse understanding
Handling Ambiguity and Context	Addressing challenges in understanding and representing ambiguous or context-dependent language	Context-aware NLG models, incorporating world knowledge and commonsense reasoning
Scalability and Efficiency	Ensuring NLG systems can handle large volumes of data and generate text in real-time efficiently	Optimization of NLG algorithms and architectures, leveraging parallel processing and cloud computing

B. Emerging trends and future directions

Advancements in neural NLG models: Future developments in NLG are expected to focus on advancing neural network architectures and training techniques. Research in areas such as transformer models, pre-training strategies, and fine-tuning methods aims to improve the fluency, coherence, and contextual understanding of neural NLG systems (Lewis et al., 2020).
Integration of NLG with other AI technologies: There is a growing trend towards integrating NLG with other AI technologies, such as natural language understanding (NLU) and dialogue management systems. This integration enables more seamless and context-aware interactions between humans and AI systems, leading to more natural and effective communication (Huang et al., 2021).
Ethical considerations and responsible NLG development: As NLG technology becomes more pervasive and influential, there is a growing need to address ethical considerations and ensure responsible development and deployment. This includes issues such as bias in training data, misinformation generation, and the impact of NLG on society and individuals (Bender et al., 2021).

VI. Conclusion

In conclusion, NLG has emerged as a powerful technology with diverse applications across various domains. Despite significant progress, challenges such as naturalness, ambiguity handling, and scalability persist. However, with ongoing advancements in neural NLG models, integration with other AI technologies, and a focus on ethical considerations, the future of NLG looks promising. By addressing these challenges and embracing emerging trends, NLG has the potential to revolutionize communication, content generation, and human-computer interaction in the years to come.

REFERENCES:

1. Arora, S., Li, Y., & Ma, T. (2016). A Simple but Tough-to-Beat Baseline for Sentence Embeddings. arXiv preprint arXiv:1607.01759.
2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? arXiv preprint arXiv:2105.14038.References:
3. Bordes, A., Boureau, Y.-L., & Weston, J. (2017). Learning End-to-End Goal-Oriented Dialog. arXiv preprint arXiv:1605.07683.
4. Dathathri, R., Narang, S., Card, D., & Sridhar, V. (2020). DyNet: The Dynamic Neural Network Toolkit. arXiv preprint arXiv:1701.03980.
5. Dongaonkar, N., Li, C., & Riedl, M. (2019). NewsArticleGenerator: Automatic News Generation with Large-scale NLP Systems. arXiv preprint arXiv:1910.12596.

6. Gardent, C., Perez-Beltrachini, L., & Sales, J. (2017). Creating Training Corpora for NLG Micro-Planners. Proceedings of the 10th International Conference on Natural Language Generation, 13–22.
7. Gkatzia, D., Hastie, H., Lemon, O., & Annibale, M. (2015). A Data-driven Approach to Predicting the Success of Bank Telemarketing. Computational Linguistics, 41(4), 663–703.
8. Higashinaka, R., Imamura, K., & Aizawa, A. (2014). Evaluating Effectiveness of Various NLG Strategies for Enhancing User Engagement in Human-robot Interaction. Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, 60–69.
9. Huang, P.-S., Liu, J.-S., & Wang, C.-H. (2021). On the Integration of AI Technologies: A Systematic Literature Review. Artificial Intelligence Review, 54(5), 3463–3487.
10. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv preprint arXiv:1910.13461.
11. Liu, Q., Ren, X., Gao, J., Howard, P., & Wang, W. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.
12. Novikova, J., Dusaban, E., Rieser, V., & Lemon, O. (2017). Why We Need New Evaluation Metrics for NLG. Proceedings of the 10th International Conference on Natural Language Generation, 203–207.
13. Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2015). Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 3776–3784.
14. Zhang, J., Liu, Y., & Luan, H. (2018). Extractive Summarization: Challenges, Methods, and Applications. IEEE Transactions on Neural Networks and Learning Systems, 29(12), 5614–5632.
15. Swartout, W., Artstein, R., Forbell, E., Foutz, S., Lane, H. C., Lange, B., ... & Traum, D. (2017). Virtual Human Standardized Patients for Clinical Training. ACM Transactions on Interactive Intelligent Systems (TiiS), 7(1), 1–38.
16. Kreuzthaler, M., Schulz, S., & Berghold, A. (2018). Analyzing the Natural Language Generation Process in Radiology Reports. Journal of Biomedical Informatics, 87, 58–67.
17. Arnold, C. W., McNamara, D. S., Duran, N. D., & Chennupati, S. (2016). Automated Detection of Student Mental Models During Computer-Based Problem Solving. International Journal of Artificial Intelligence in Education, 26(1), 301–326.
18. Zhou, L., Zhang, D., & Sun, L. (2017). Information Technology-based Diabetes Management Interventions: A Systematic Review. Journal of Diabetes Science and Technology, 11(1), 116–127.References:
19. Gatt, A., Belz, A., & Kow, E. (2009). The TUNA-REG Corpus: A Corpus for Evaluating Surface Realisation by Statistical NLG Systems. Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2008), 69–76.
20. Belz, A., & Reiter, E. (2006). Comparing Automatic and Human Evaluation of Realisation Quality for NLG Systems. Proceedings of the 11th European Workshop on Natural Language Generation, 129–136.
21. Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., ... & Rambow, O. (2014). Madamira: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. Proceedings of the Language Resources and Evaluation Conference (LREC).
22. Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical Phrase-based Translation. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, 48–54.
23. Nenkova, A., & McKeown, K. (2011). Automatic Summarization. Foundations and Trends® in Information Retrieval, 5(2-3), 103–233.
24. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog.
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All you Need. Advances in Neural Information Processing Systems, 30, 5998–6008.
26. Winograd, T. (1972). Understanding Natural Language. Cognitive Psychology, 3(1), 1–191.
27. Langkilde, I., & Knight, K. (1998). Generation that Exploits Corpus-based Statistical Knowledge. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1, 704–710.
28. Lahiri, S., & Reddy, C. (2011). Natural Language Generation in Narrative Science's Quill Platform. AI Magazine, 32(3), 61–76.
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is All you Need. Advances in Neural Information Processing Systems, 30, 5998–6008.
30. This section provides an overview of the early developments, key milestones, and the evolution of NLG techniques and algorithms, supported by citations from relevant research papers published between 2012 and 2018.

31. Overall, the versatility and applicability of NLG across various domains underscore its significance in automating content generation, improving communication, and enhancing user experiences.
32. Liu, Q., Zhao, H., & Jansche, M. (2018). Data-to-text Generation with Content Selection and Planning. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2712–2722.
33. Du, Y., Xu, Z., Tao, C., & Xu, H. (2016). Natural Language Generation in Health Care. Artificial Intelligence in Medicine, 69, 1–8.
34. Gkatzia, D., Hastie, H., Lemon, O., & Annibale, M. (2015). A Data-driven Approach to Predicting the Success of Bank Telemarketing. Computational Linguistics, 41(4), 663–703.
35. Wen, T. H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-Barahona, L. M., Su, P.-H., & Young, S. (2015). Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 1711–1721.