UDC (004.8+004.92)::81`32

Big semi-structured data & deep ANN in computational linguistics

Svitlana Goncharenko
Kyiv National University of Technologies and Design, Kyiv
https://orcid.org/0000-0002-7740-4658

Abstract. *Modern linguistics is experiencing a qualitative shift due to the active implementation of big data processing technologies and artificial neural networks. This paradigm shift is fueled by huge arrays of digital texts, particularly Big Semi-Structured Data (e.g., emails and social media posts), which combines flexibility with elements of organization. Deep artificial neural networks are a key tool for analyzing these resources, allowing to model complex dependencies and automate tasks such as machine translation, speech recognition and text generation. The integration of big data and artificial intelligence forms a new research paradigm that shifts the emphasis from static description to dynamic modeling and prediction of linguistic phenomena, increasing the interdisciplinarity and applied potential of linguistics.*
Keywords: *philology, computational linguistics, Big Data.*

Introduction.

Modern linguistics is experiencing a qualitative shift due to the active implementation of big data processing technologies and artificial neural networks. If earlier research was mainly based on limited data sets and traditional analytical approaches (which involved the use of symbolic AI (knowledge-based) [1] based on human expert experience [2] and the results of classical machine learning [3]), today the situation has changed dramatically. It is the accumulated huge arrays of digital texts, speech streams and multimodal materials that provide a unique opportunity to study language in its mobile, changing and living form.

Big structured, semi-structured and unstructured data [4] allow us to record new patterns (templates) [5] in the field of structure, semantics, dynamics, pragmatics and other aspects. It is semi-structured data that combines flexibility and organization [6]: they contain ordering elements (labels, tags, attributes), but are not subject to strict schemes. This allows them to be stored, analyzed, and processed on a large scale, making them extremely valuable for modern analytics and artificial intelligence applications [7]. Examples of such data include XML and JSON files, emails, social media posts, and event logs (log files). All of them combine structured elements (e.g., date, author, tags) and unstructured parts (message text, images, or other content).

It is Big Semi-Structured Data in philological research that reflects global processes in communication, cultural and social contexts, and the dynamics of language evolution.

Artificial neural networks [8], especially deep models [9], are becoming a central tool for analyzing such resources. Their potential in modeling complex relationships, learning on large-scale corpora, and detecting hidden dependencies makes them key for tasks such as machine translation, speech recognition, text generation, meaning interpretation, and the creation of intelligent communication systems.

Thus, the combination of big data and artificial intelligence forms a new research paradigm in linguistics. It is based on the transition from static description to dynamic modeling and prediction of language phenomena, which strengthens the interdisciplinarity of research and expands the applied potential of linguistic science.

The Main Part.

Modern linguistics is undergoing a period of intensive digitalization, with traditional methods of language analysis being actively supplemented by new technologies for processing large semi-structured data sets and deep neural networks. Semi-structured information (social media messages, online forums, blogs, multimodal texts with markup elements) combines features of structure and unstructuredness. The massive accumulation of such Big Data opens up fundamentally new opportunities for researchers–the identification of patterns and semantic connections that were previously impossible to detect using classical approaches.

Deep neural network architectures have become a key tool for analyzing such data. Their ability to train on colossal language corpora, extract hidden levels of abstraction, and work with real-world communicative practices makes them indispensable for solving problems of machine translation, speech recognition, automatic text analysis, linguistic construction generation, and the study of language system dynamics.

The application of these technologies is becoming especially important for cognitive, sociolinguistics, and computational linguistics. Semi-structured datasets capture live speech and digital interactions in natural environments, while deep neural networks enable the construction of adaptive models that account for cultural, social, and contextual factors. Thus, the integration of big data and AI technologies is shaping a new research paradigm, shifting the emphasis from descriptive analysis to predicting and modeling language processes.

Conclusions.

1. Large semi-structured data sets provide researchers with unique material for analyzing language in its dynamic and natural manifestation.

2. Deep artificial neural networks are becoming the main tool for working with such data, allowing to automate complex tasks - from morphosyntactic analysis to the detection of semantic structures.

3. The interaction of these technologies creates new opportunities for studying the nature of language and developing intelligent systems capable of natural dialogue with humans.

4. The use of approaches based on big data and AI expands the applied potential of linguistics in the field of machine translation, intelligent assistants, search engines and text content analysis.

5. The synthesis of semi-structured data and neural network models transforms modern linguistics, making it more interdisciplinary, innovative and practice-oriented.

Discussion.

The author reasonably argues:

1) that it is hybrid deep learning models that combine the advantages of different neural network approaches (for example, transformers, recurrent and convolutional structures) that can significantly increase the efficiency of data processing [10]. Their use opens up opportunities for solving such tasks as automatic translation, intelligent speech analysis, detection of semantic connections and prediction of language changes. The integration of big data and hybrid neural technologies contributes to the construction of adaptive systems that take into account not only the structural aspects of language, but also the social, cognitive and cultural features of communication. This forms a new paradigm in linguistics, where the emphasis is on interdisciplinarity, innovation and close interaction with artificial intelligence.

2). In modern applied scientific research (in particular in the field of philology), the synergy effect is becoming increasingly important for complex, dynamic and interdisciplinary scientific research [11], especially those involving the use of big data and hybrid deep neural networks. This effect is manifested in the fact that the combination of powerful analytical tools and large-scale language corpora allows you to obtain a result that exceeds the sum of the individual contributions of each approach. The synergy of such approaches opens up new opportunities for studying the structure, semantics, pragmatics and sociocultural aspects of language. It allows you to combine classical linguistic methods with analytical and predictive models, creating an innovative, adaptive and interdisciplinary paradigm of modern linguistics.

References

1. Tuhaienko V., Krasniuk S. Effective application of knowledge management in current crisis conditions. *International scientific journal "Grail of Science".* 2022. № 16. pp. 348-358.

2. Naumenko, M. (2024). Models of business knowledge in artificial intelligence systems for an effective competitive enterprise. *International scientific journal "Internauka". Series: "Economic Sciences"*. № 6. DOI: https://doi.org/10.25313/2520-2294-2024-6-10010 [In Ukrainian].

3. Naumenko, M. (2024). Efektyvne zastosuvannia klasychnykh alhorytmiv mashynnoho navchannia pry pryiniatti adaptyvnykh upravlinskykh rishen [Effective application of classic machine learning algorithms when making adaptive management decisions]. *Scientific perspectives (special edition)*, 5 (47). DOI: https://doi. org/10.520 58/2708-7530-2024-5(47)-855-875 [in Ukrainian].

4. Krasnyuk, M., Nevmerzhytska, S., & Tsalko, T. (2024). Processing, analysis & analytics of big data for the innovative management. *Grail of Science*, (38), 75–83. https://doi.org/10.36074/grail-of-science.12.04.2024.011.

5. Krasnyuk M, Elishis D (2024). Perspectives and problems of big data analysis & analytics for effective marketing of tourism industry. *Наука і Техніка Сьогодні*, 4(32). https://doi.org/10.52058/2786-6025- 2024-4(32)-833-857.

6. Krasnyuk M., Krasnuik I. (2024) Big data analysis and analytics for marketing and retail. *Proceedings of the International Scientific Conference "Artificial Intelligence in Science and Education" (AISE).* – Kyiv, March 2024. pp. 459-463.

7. Naumenko, M. (2024). Analiz ta analityka velykykh danykh v marketynhu ta torhivli konkurentnoho pidpryiemstva [Analysis and analytics of big data in marketing and trade of a competitive enterprise]. *Grail of Science*, (40), 117–128. DOI: https://doi.org/10.36074/grail-of-science.07.06.2024.013 [in Ukrainian].

8. Maksym Naumenko (2024). Regression analysis using shallow artificial neural networks in the management of an efficient and competitive enterprise. *Věda a perspektivy*, 7(38) (2024), pp. 17-32. https://doi.org/10.52058/2695-1592-2024-7(38)-17-32.

9. Naumenko, M. (2024). Optymalne vykorystannia alhorytmiv hlybokoho mashynnoho navchannia v efektyvnomu upravlinni pidpryiemstvom [Optimal use of deep machine learning algorithms in effective enterprise management]. *Successes and achievements in science*, 4 (4). DOI: https://doi.org/10.52058/3041-1254-2024-4(4)-776-794 [in Ukrainian].

10. Krasnyuk, M. (2014). Hybridization of intelligent methods of business data analysis (anomaly detection mode) as a standard tool of corporate audit. *The state and prospects of the development Education and science of today:* materials of the III International science and practice conf. [m. Ternopil, October 10-11. 2014]. TNEU, 2014. pp. 211-212 [in Ukrainian].

11. Derbentsev, V. D., Serdiuk, O. A., Soloviov, V. M., & Sharapov, O. D. (2010). *Synergistic and econophysical methods of studying dynamic and structural characteristics of economic systems*. Cherkasy: Brama-Ukraine. - 2010 [in Ukrainian].