DOI https://doi.org/10.30525/978-9934-26-475-7-2

NEURAL NETWORK-BASED OPTIMIZATION FOR DOMAIN-LEVEL COMPUTATIONAL LINGUISTICS

НЕЙРОННО-МЕРЕЖЕВА ОПТИМІЗАЦІЯ ДЛЯ ДОМЕННО-ОРІЄНТОВАНОЇ МАТЕМАТИЧНОЇ ЛІНГВІСТИКИ

Dychka I. A.
*Doctor of Technical Sciences,
Dean of Faculty
of Applied Mathematics,
Professor at the Computer Systems
Software Department
National Technical University
of Ukraine
"Igor Sikorsky Kyiv
Polytechnic Institute"
Kyiv, Ukraine*

Дичка І. А.
*доктор технічних наук,
декан факультету прикладної
математики,
професор кафедри програмного
забезпечення комп'ютерних систем
Національний технічний університет
України
"Київський політехнічний інститут
імені Ігоря Сікорського"
м. Київ, Україна*

Potapova K. R.
*Candidate of Technical Sciences,
Associate Professor at the Department
of System Programming
and Specialized Computer Systems
National Technical University
of Ukraine
"Igor Sikorsky
Kyiv Polytechnic Institute"
Kyiv, Ukraine*

Потапова К. Р.
*кандидат технічних наук,
доцент кафедри системного
програмування і спеціалізованих
комп'ютерних систем
Національний технічний університет
України
"Київський політехнічний інститут
імені Ігоря Сікорського"
м. Київ, Україна*

Meliukh V. V.
*Postgraduate Student at the Computer
Systems Software Department
National Technical University
of Ukraine
"Igor Sikorsky
Kyiv Polytechnic Institute"
Kyiv, Ukraine*

Мелюх В. В.
*аспірант кафедри програмного
забезпечення комп'ютерних систем
Національний технічний університет
України
"Київський політехнічний інститут
імені Ігоря Сікорського"
м. Київ, Україна*

The growth of artificial intelligence (AI) usage in different areas of human relations led to the integration of AI as a cornerstone of the new technological era [1, p.8]. Natural language processing (NLP), a branch of the AI field concerned with machine processing of text-based information, is responsible for furthering computational linguistics techniques to reach an automated analysis and comprehension of human language across diverse domains. Law, science, finance and healthcare – all benefit from the ability of NLP models to tailor their needs by handling unique vocabularies, syntactic structures, and contextual nuances that are different from general language manipulation. However, the uncharacterized uniqueness of each language and applied tasks affects the ability of models to respond to the problem posed. Therefore, customized model interpretations require novel approaches and strategies within neural network-assisted computational linguistics. It entails the realization of techniques regarding data selection and preprocessing, feature engineering, transfer learning, meta-learning or the selection of appropriate models to enhance performance in various fields. Addressing these factors can lead to the development of more effective and accurate systems for specialized applications with incorporated domain-specific knowledge and unique characteristics pertaining to neural network models.

Recent findings in natural language processing (NLP) convey promising results, indicating the increase in models' capability to capture intricate linguistic nature and achieve state-of-the-art output on numerous testing benchmarks. Particularly, general-purpose models with a strong data foundation such as BERT [2, p. 1], ELMo and GPT have proven to be quite potent at a variety language tasks, including but not limited to question answering, translation, and text generation [3, p. 261].

Transfer learning proves its worth in tasks related to the adaptation of general-purpose NLP models by fine-tuning them on customized datasets in accordance with their syntax and vocabulary. Domain-specific word embeddings, such as those derived from science-related or biomedical data corpora, provide better performance with more accurate feature representations by capturing the nuanced meanings of words. Consequently, BioBERT (a pre-trained biomedical data version of BERT) outperforms the general model in biomedical tasks [4, p. 1234]. However, while recent studies spiked an interest in exploring transfer learning and area-bounded adaptations in the face of BioBERT or FinBERT, domain-specific contexts still need supervision to break a gap for fully optimized neural network-based approaches. Present solutions are more fixated on one-dimensional pre-training data clusters without diving deeper into a more comprehensive approach that combines model-centric and data-centric techniques, creating hybrid paradigms [5, p. 15].

The reason why traditional supervised learning approaches are leaving the mainstream scientific discussion lies in the scarcity of available labeled data in many domain-specific fields. On the other hand, methods like active learning and pseudo-labeling mitigate the rising problem by building synthetic or selectively labeled information to enhance language pattern learning in specialized areas [6, p. 16]. Active learning strategy employs models that identify the most sensitive and informative examples from the set that need to be labeled by experts, challenging data-limited environments and improving learning efficiency. The fusion of data-centric strategies around specific domains and advanced neural network-assisted optimization techniques like meta-learning presents a fresh systematic manner of tackling model generalization variability among various tasks within a single knowledge sphere, potentially changing how domains interact with language processing tasks.

A hybrid approach with integrated model-centric and data-centric strategies allows for balanced architecture improvements with increased data quality of the specialized field. Improved handling of the unique linguistic characteristics of the domain may be mediated by additional layers with custom attention mechanisms or task-specific embeddings. Polyfunctional systems, alternatively, can employ multi-task learning to optimize the model's performance or even design task-specific loss functions to account for critical errors specific to the restricted problem. For instance, combining fine-tuned neural network architectures with custom embeddings accounts for an optimized model structure and input data for a given domain.

The development of AI-powered systems brings remarkable changes for the future of NLP adjacent domains as more advanced and refined methods and techniques enter the space. For the time being, the optimization of established methodologies leads to more accurate and reliable output with reduced rates of errors in critically important fields [1, p. 10–14]. The hybrid approach shows high scalability and adaptability of methods in multiple sectors, ensuring the possibility of customization for any specialized field. In the healthcare or finance industry, improved named entity recognition can be attributed to a higher rate of positive decision-making processes and more streamlined classification [8, p. 87]. It makes a desirable and versatile solution for industries that deal with complex domain-specific language.

Resource-constricted regions of different disciplines elicit the need for high-performing models without heavy spending on overcoming the challenges of limited data brought to the scene such effective strategies as active learning, meta-learning and pseudo-labeling. Custom loss functions and domain-tailored training reduce error rates and improve task-specific accuracy, minimizing the need for extensive manual intervention. The findings pave the way for more advancements in the use of neural networks to specific NLP tasks. New research into more intricate multi-domain or cross-domain NLP models is sparked by the success of transfer learning and hybrid techniques in domain-specific situations, which push the limits of automated language processing.

Bibliography:

1. Bughin J., Hazan E., Sree Ramaswamy P., DC W., Chu M. Artificial intelligence: the next digital frontier. 2017.

2. Devlin J. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint. 2018. Available at: arXiv:1810.04805.

3. Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. 3rd edition. Online manuscript released August 20, 2024. Available at: https://web.stanford.edu/~jurafsky/slp3

4. Zhuang F., Qi Z., Duan K., Xi D., Zhu Y., Zhu H., Xiong H., He Q. A comprehensive survey on transfer learning. *Proceedings of the IEEE*. 2020. Vol. 109, № 1. P. 43–76. DOI: arXiv:1911.02685

5. Lee J., Yoon W., Kim S., Kim D., Kim S., So C.H., Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020. Vol. 36, № 4. P. 1234–1240. DOI: 10.1093/bioinformatics/btz682

6. Wilson G., Cook D. J. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2020. Vol. 11, № 5. P. 1–46. DOI: 10.1145/3400066

7. Hu Z., Dychka I., Potapova K., Meliukh V. Augmenting sentiment analysis prediction in binary text classification through advanced natural language processing models and classifiers. *International Journal of Information Technology and Computer Science*. 2024. Vol. 16. P. 16–31. DOI: 10.5815/ijitcs.2024.02.02

8. Дичка І., Потапова К., Вовк Л., Мелюх В., Веденєєва О. Adaptive domain-specific named entity recognition method with limited data. *Measuring and Computing Devices in Technological Processes*. 2024. № 1. P. 82–92. DOI: 10.31891/2219-9365-2024-77-11