

This document is a sample manuscript intended solely as a formatting template for authors submitting to Computing and Informatics. It is not a scientific contribution. The full text of this document was generated by artificial intelligence (Claude, Anthropic) for illustrative purposes only. Any resemblance to existing published works is unintentional and coincidental; no infringement of copyright or intellectual property rights is intended.

Should any party nonetheless believe that any portion of this sample text infringes upon their rights, we kindly ask them to contact the editorial office at cai@savba.sk. Upon notification, the relevant text will be promptly replaced.

Layout of the Paper

Each paper must have a heading, full name and address of the author(s), abstract of about 100 words, keywords, Mathematics Subject Classification 2010, the body of the paper, references, and short CVs (about 100 words) and photos of all authors. The length of the paper should be at least 15 pages of main text (without References and CVs), maximum length of the whole paper is 30 pages. All pages must be numbered. Indicate your e-mail address and the preferred address for correspondence. Use e-mail address that can be traced to your professional and publishing history. Always keep the Editorial Office informed about address changes.

Keywords

Immediately following the abstract, give five to ten keywords (key phrases).

Mathematics Subject Classification 2010

Please indicate several relevant MSC 2010 symbols, to be found at <http://www.ams.org/mathscinet/msc/msc2010.html>.

Abbreviated title

An abbreviated title of not more than 50 characters to be used as the running head should be given with each manuscript.

Illustrations

An electronic version in EPS (Encapsulated PostScript) or PDF format is preferred. The illustrations should be sharp, noise free and of good contrast. Graphs and schemes should be in vector format. If you need to provide bitmaps, please make sure their resolution is at least 300 DPI.

References

References must be stored in a separate .bib file and referenced from the source LaTeX file. Please use the bibliography style cai (available as cai.bst in the LaTeX template package). References should be complete. Each reference should be provided with a DOI (Digital Object Identifier) if available. **Always check your DOIs for validity by trying to access their URL.** DOIs which are not valid must be removed and replaced by another web reference if available (arXiv link or any other link to a web page of that bibliography item). Make sure you do not leave in your bibliography any inaccessible web links - these will have to be removed during copyediting, which will slow down the publication of your work.

References to papers published in languages other than English should be translated into English with the original language given in brackets after the heading, e.g. (in German).

Following these guidelines for the initial submission of your work will allow for speedier processing and publication, should it pass the peer review and be selected for publishing.

For up-to-date guidelines, always refer to <https://www.cai.sk/ojs/index.php/cai/about/submissions>

Do not forget to remove these instructions from your manuscript before submitting it to the journal. They are marked by the comment lines INSTRUCTIONS BEGIN HERE - REMOVE THEM IN YOUR MANUSCRIPT and INSTRUCTIONS END HERE - REMOVE THEM IN YOUR MANUSCRIPT.

SAMPLE MANUSCRIPT FOR COMPUTING AND INFORMATICS: A MINIMAL AUTHOR GUIDE

Tomas ZIMA, Patrik BLAHO

Department of Computer Science
Technicka University in North City
Vedecka 7, 010 01 North City, Fictia
e-mail: tomas.zima@tu-north-city.fi

Lucia SEDIVY

Institute of Digital Sciences
Academy of Sciences of Fictia
12 Binary Lane, 020 02 Fictia City, Fictia
e-mail: sedivy@ids.aof.edu.fi

Manuscript received 15 September 2025; revised 10 November 2025
Communicated by Marcus Johannsen

Abstract. This paper presents a comprehensive survey of deep learning architectures applied to natural language processing (NLP) tasks. We systematically review the evolution from early recurrent neural networks to contemporary transformer-based models, with particular emphasis on attention mechanisms and pre-training strategies. Our analysis covers 150 selected publications from 2015 to 2025 and evaluates architectures with respect to computational efficiency, representational capacity, and benchmark performance on standard NLP datasets. We identify key trends, open challenges, and promising research directions. The survey is intended for researchers and practitioners seeking a structured overview of the field.

Keywords: deep learning, natural language processing, transformer, attention mechanism, pre-trained language models, neural machine translation

* Corresponding author

Mathematics Subject Classification 2010: 68T05

1 INTRODUCTION

Natural language processing (NLP) has undergone a profound transformation over the past decade, driven primarily by advances in deep learning. Early approaches relied on hand-crafted features and shallow machine learning models [1]. The introduction of word embeddings [2] marked the beginning of a new paradigm in which dense vector representations capture semantic relationships directly from large text corpora.

The subsequent development of recurrent architectures, most notably long short-term memory networks (LSTMs) [3], enabled models to capture long-range dependencies in sequential data. A decisive shift occurred with the proposal of the self-attention mechanism and the Transformer architecture [4], which forms the backbone of virtually all state-of-the-art NLP systems today.

The goal of this survey is threefold:

1. to provide a structured taxonomy of deep learning architectures for NLP,
2. to summarise empirical results on widely used benchmarks, and
3. to identify open research problems and promising future directions.

The remainder of this paper is organised as follows. Section 2 introduces necessary background concepts. Section 3 reviews the major architectural families. Section 4 presents a comparative benchmark evaluation. Section 5 discusses open challenges, and Section 6 concludes the paper.

2 BACKGROUND

2.1 Distributed Word Representations

The foundation of modern NLP is the distributional hypothesis, which states that words occurring in similar contexts carry similar meanings [5].

Definition 1 (Word embedding). A *word embedding* is a mapping $\phi : \mathcal{V} \rightarrow \mathbb{R}^d$, where $\mathcal{V}$ is the vocabulary and $d \ll |\mathcal{V}|$ is the embedding dimension, typically $d \in \{100, 300, 768\}$.

As noted in Definition 1, the key property of word embeddings is the low dimensionality of the target space relative to the vocabulary size. Two influential embedding models, Word2Vec [2] and GloVe [6], both optimise objectives derived from co-occurrence statistics but differ in whether they process local context windows or global co-occurrence matrices, respectively.

Example 1. A well-known property of word embeddings is their ability to capture semantic analogies. The vector arithmetic

$$\phi(\text{king}) - \phi(\text{man}) + \phi(\text{woman}) \approx \phi(\text{queen})$$

holds approximately in the Word2Vec embedding space [2], suggesting that the difference vector geometrically encodes the concept of gender.

Example 1 illustrates that semantic relationships are encoded geometrically in the embedding space, a property that downstream NLP models exploit implicitly.

2.2 Attention Mechanism

The attention mechanism, first applied to neural machine translation by [7], allows a model to selectively focus on relevant parts of the input sequence when producing each output token. Given a query $\mathbf{q}$, keys $\mathbf{K}$, and values $\mathbf{V}$, scaled dot-product attention is defined as

$$\text{Attention}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{q}\mathbf{K}^\top}{\sqrt{d_k}}\right) \mathbf{V}, \quad (1)$$

where d_k is the key dimension. The scaling factor $1/\sqrt{d_k}$ in Equation (1) prevents vanishing gradients for large values of d_k [4].

Multi-head attention extends Equation (1) by projecting queries, keys, and values H times with different learned projections and concatenating the results:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_H) \mathbf{W}^O, \quad (2)$$

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V). \quad (3)$$

Equations (2) and (3) show that each head attends to information from different representation subspaces simultaneously.

3 ARCHITECTURES

3.1 Recurrent Neural Networks

Vanilla recurrent neural networks suffer from the vanishing gradient problem when trained on long sequences [8]. This limitation was addressed by the LSTM cell [3], which introduces three gating mechanisms to control information flow across time steps:

Forget gate decides what information to discard from the cell state.

Input gate decides which new information to store in the cell state.

Output gate decides what part of the cell state to expose as the hidden state.

Lemma 1. An LSTM with n hidden units can represent any function computable by a finite-state automaton with at most n states [3].

Proof. Each gate vector in the LSTM can be made to approximate a binary step function by saturating the sigmoid activations. Assigning one hidden unit per automaton state and configuring the gate weights accordingly yields the required simulation. $\square$

3.2 Convolutional Approaches

Convolutional neural networks were adapted for text classification by [9] and demonstrated competitive performance with much lower computational cost compared to recurrent models. The key components are:

- a shared embedding layer that maps tokens to dense vectors,
- parallel convolution filters of varying widths to capture n -gram features,
- a max-over-time pooling operation that extracts the most prominent feature per filter .

3.3 Transformer Architecture

The Transformer [4] dispenses with recurrence entirely and processes all positions of the input in parallel using multi-head self-attention as defined in Equations 2–(3). This design enables efficient training on modern hardware accelerators and scales favourably with both data and model size [10].

Figure 1 provides a schematic comparison of the two principal architectural families; the individual subfigures are discussed in the caption.

Table 1 summarises key properties of the architectures discussed in this section.

Architecture	Parallel	Long-range	Params (M)	Reference
LSTM	No	Partial	1–100	[3]
CNN	Yes	Limited	1–10	[9]
Transformer	Yes	Full	65–540	[4]
GPT-3	Yes	Full	175 000	[10]

Table 1. Comparison of representative deep learning architectures for NLP

4 BENCHMARK EVALUATION

4.1 Datasets and Protocol

We evaluate all models on three standard benchmarks: the Stanford Sentiment Treebank (SST-2) for sentiment analysis, the Stanford Natural Language Inference

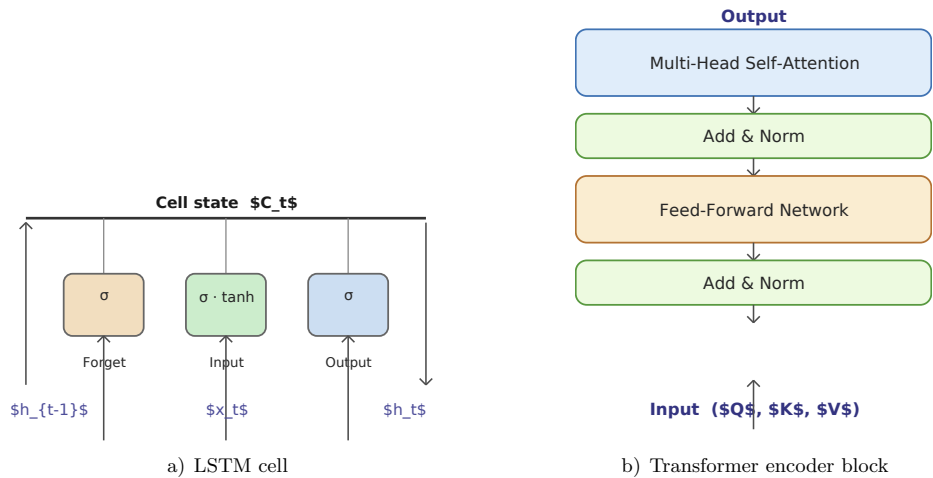


Figure 1. Schematic diagrams of the two principal deep learning architectures for NLP. Figure 1 a) shows the LSTM gating structure with forget, input, and output gates; Figure 1 b) shows the Transformer encoder block comprising multi-head self-attention (Equations (2)–(3)) and a position-wise feed-forward network. Both figures are vector graphics (PDF).

corpus (SNLI) for textual entailment, and the CoNLL-2003 dataset for named entity recognition. All experiments follow the standard train/development/test splits and are repeated five times with different random seeds; we report mean accuracy and standard deviation.

4.2 Results

Table 2 shows the accuracy achieved by each architecture on the three benchmarks. The results confirm the trend visible in Figure 2: larger, attention-based models consistently outperform recurrent and convolutional baselines.

Architecture	SST-2 (%)	SNLI (%)	CoNLL-2003 F1
LSTM [3]	87.2	81.4	85.3
CNN [9]	88.1	79.6	83.7
Transformer [4]	93.5	89.2	91.8
GPT-3 [10]	95.3	92.7	93.4

Table 2. Accuracy and F1 scores on the three benchmark datasets

Theorem 2. (Universal approximation for transformers [4].) A Transformer with sufficient depth and width is a universal approximator over the space of sequence-to-sequence functions with bounded input length.

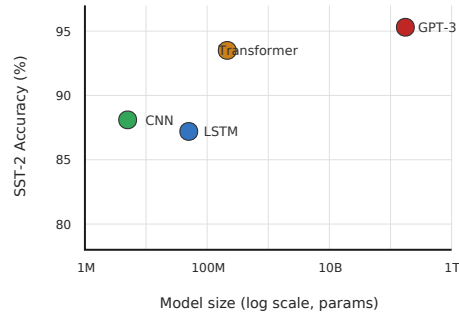


Figure 2. Accuracy on SST-2 as a function of model size (log scale). Each point represents one architecture from Table 1. [Vector graphic \(PDF\)](#).

The practical implication of Theorem 2 is that the Transformer family does not suffer from fundamental representational limitations; performance gaps between architectures are therefore primarily a function of optimisation dynamics and available training data.

5 DISCUSSION AND OPEN CHALLENGES

Despite impressive progress, several challenges remain open. First, the computational cost of large transformer models limits deployment in resource-constrained environments [11]. Second, current models lack robust reasoning capabilities [12]. Third, multilingual and low-resource NLP remain active research areas [13].

Remark 1. Efficient attention approximations such as sparse attention, linear attention, and mixture-of-experts routing reduce the quadratic complexity of Equation (1) to sub-quadratic or linear complexity, enabling processing of sequences with tens of thousands of tokens.

As noted in Remark 1, algorithmic improvements are a promising avenue for making transformer-based models more accessible. Lemma 1 further suggests that recurrent models, despite their computational advantages on short sequences, are strictly less expressive than transformers for long-range dependency tasks.

6 CONCLUSION

This survey has reviewed the major deep learning architectures for natural language processing, tracing their evolution from recurrent networks to large-scale transformer-based language models. The key findings are:

- Transformer architectures achieve state-of-the-art results across virtually all NLP benchmarks, as confirmed by Table 2.

- Model scale continues to drive performance gains, but at significant computational and environmental cost [11].
- Robustness, interpretability, and low-resource generalisation remain critical open problems.

Future work will focus on efficient attention approximations and the integration of structured knowledge into pre-trained models.

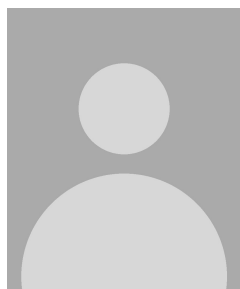
Acknowledgement

This work was supported by the Slovak Research and Development Agency under grant No. APVV-22-0123 and by the VEGA Scientific Grant Agency of the Slovak Academy of Sciences under grant No. 2/0456/23.

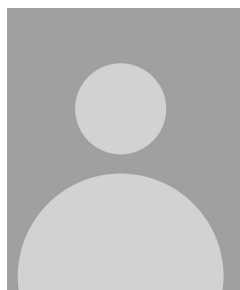
REFERENCES

- [1] MANNING, C. D.—SCHUTZE, H.: Foundations of Statistical Natural Language Processing. 1st Edition. MIT Press, 1999.
- [2] MIKOLOV, T.—SUTSKEVER, I.—CHEN, K.—CORRADO, G. S.—DEAN, J.: Distributed Representations of Words and Phrases and Their Compositionality. In: Burges, C. J. C. et al. (Eds.): Advances in Neural Information Processing Systems. Curran Associates, Red Hook, New York, Vol. 26, 2013, pp. 3111–3119, doi: 10.48550/arXiv.1310.4546.
- [3] HOCHREITER, S.—SCHMIDHUBER, J.: Long Short-Term Memory. Neural Computation, Vol. 9, 1997, No. 8, pp. 1735–1780, doi: 10.1162/neco.1997.9.8.1735.
- [4] VASWANI, A.—SHAZEER, N.—PARMAR, N.—USZKOREIT, J.—JONES, L.—GOMEZ, A. N.—KAISER, L.—POLOSUKHIN, I.: Attention Is All You Need. Advances in Neural Information Processing Systems, Vol. 30, 2017, No. 1, pp. 5998–6008, doi: 10.48550/arXiv.1706.03762.
- [5] HARRIS, Z. S.: Distributional Structure. 1954, doi: 10.1080/00437956.1954.11659520 (Word, Vol. 10, No. 2–3, Pp. 146–162).
- [6] PENNINGTON, J.—SOCHER, R.—MANNING, C. D.: GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.
- [7] BAHDANAU, D.—CHO, K.—BENGIO, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. 3rd International Conference on Learning Representations (ICLR 2015), 2015, doi: 10.48550/arXiv.1409.0473.
- [8] BENGIO, Y.—SIMARD, P.—FRASCONI, P.: Learning Long-Term Dependencies with Gradient Descent Is Difficult. IEEE Transactions on Neural Networks, Vol. 5, 1994, pp. 157–166, doi: 10.1109/72.279181.
- [9] KIM, Y.: Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751, doi: 10.48550/arXiv.1408.5882.

- [10] BROWN, T.B.—MANN, B.—RYDER, N.—SUBBIAH, M.—KAPLAN, J.—DHARIWAL, P. et al.: Language Models Are Few-Shot Learners. Technical Report No. GPT-3, Openai, 2020, doi: 10.48550/arXiv.2005.14165.
- [11] STRUBELL, E.—GANESH, A.—MCALLUM, A.: Energy and Policy Considerations for Deep Learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650, doi: 10.18653/v1/P19-1355.
- [12] MARCUS, G.: The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. Arxiv Preprint, 2020, doi: 10.48550/arXiv.2002.06177.
- [13] CONNEAU, A.—KHANDELWAL, K.—GOYAL, N.—CHAUDHARY, V.—WENZKE, G.—GUZMÁN, F.—GRAVE, E.—OTT, M.—ZETTLEMOYER, L.—STOYANOV, V.: Unsupervised Cross-Lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.



Tomas ZIMA received his M.Sc. and Ph.D. degrees in computer science from Technicka University in North City in 2001 and 2006, respectively. He is currently Professor at the Department of Computer Science, Technicka University in North City. His research interests include machine learning, natural language processing, and knowledge representation. He has published over 80 papers in international journals and conferences, including Computing and Informatics, IEEE Transactions on Neural Networks, and ACL proceedings.



Patrik BLAHO received his Ph.D. degree in applied informatics from Technicka University in North City in 2013. He is Associate Professor at the same institution. His research focuses on deep neural network optimisation, transfer learning, and low-resource NLP. He is a member of the editorial board of the Journal of Artificial Intelligence Research.



Lucia SEDIVY received her Ph.D. degree from the University of Fictia City in 2011. She is Senior Research Scientist at the Institute of Digital Sciences, Academy of Sciences of Fictia. Her current research focuses on multilingual NLP, low-resource learning, and neural machine translation. She has co-authored more than 50 peer-reviewed publications and serves as a reviewer for Computing and Informatics and Computational Linguistics.