**Міністерство освіти і науки, молоді та спорту**

**Національний технічний університет України**

**“Київський політехнічний інститут”**

Факультет біомедичної інженерії

Кафедра біомедичної інженерії

**РЕФЕРАТ**

з дисципліни “Англійська мова професійного спрямування”(ІІІ курс)

на тему: **“Language Models and Natural Language Processing”**

Виконав: Азіза Е.А.

група БС-12

Викладач: Компанець Н.М.

Київ 2023

Contents

[**Introduction 3**](#_heading=h.3znysh7)

[**Fundamentals of Natural Language Processing (NLP) 4**](#_heading=h.tyjcwt)

[**History and development of language models 4**](#_heading=h.9v953lvyec9m)

[**Modern language models 6**](#_heading=h.rtrtb6bojq1t)

[**Applying Language Models 7**](#_heading=h.5e8fck6nkaok)

[Automatic machine translation 7](#_heading=h.50lxfymi4ov2)

[Virtual assistants 7](#_heading=h.3kxoh6ixi6vg)

[Chatbots 7](#_heading=h.vzplxy2x0hzj)

[Sentiment Analysis 7](#_heading=h.uecr8voiilg6)

[Speech recognition 7](#_heading=h.qwnf00r0rvb6)

[Education 7](#_heading=h.we7tuy3r9yk6)

[**Summary in English 8**](#_heading=h.3rdcrjn)

[**Summary translation 9**](#_heading=h.lnxbz9)

[**References 10**](#_heading=h.35nkun2)

[**Glossary 11**](#_heading=h.44sinio)

# 

# 

# 

# Introduction

Artificial intelligence (AI) has made significant breakthroughs in the past few decades, and one of its most important application areas is natural language processing (NLP). Natural language processing is a branch of computer science and artificial intelligence that deals with the interaction between computers and human (natural) language. Thanks to the development of NLP, computers are now able to understand, interpret, and respond to human language in ways that once seemed inconceivable.

This article explores the history and evolution of language models, from early statistical methods to modern neural network architectures such as Transformer. We will also discuss the core tasks of NLP, covering a wide range from machine translation to text sentiment analysis. In addition, modern applications of language models have been explored in various fields, including machine translation, chatbots, sentiment analysis, etc.

The importance of developing effective language models and NLP techniques cannot be underestimated. In a globalized world where information and communications are crucial, the ability of machines to understand and generate natural language has become critical. From banking to healthcare, from education to tourism – language models are making their way into every industry. Their impact extends far beyond technical disciplines, into social, cultural and economic spheres.

Therefore, a deep understanding of language models and natural language processing opens up new horizons for innovation and development in many fields of human activities. However, despite impressive progress, natural language processing still faces many issues and challenges, including issues of ethics, bias, and environmental impact. Finally, we examine current trends and possible futures in this dynamic field. This summary provides a comprehensive understanding of natural language processing and language models and their importance and impact on the modern world.

# 

# Fundamentals of Natural Language Processing (NLP)

In Natural Language Processing, the machine learning training algorithms study millions of examples of text — words, sentences, and paragraphs — written by humans. By studying the samples, the training algorithms gain an understanding of the “context” of human speech, writing, and other modes of communication. This training helps NLP software to differentiate between meanings of various texts. The five phases of NLP involve lexical (structure) analysis, parsing, semantic analysis, discourse integration, and pragmatic analysis. Some well-known application areas of NLP are Optical Character Recognition (OCR), Speech Recognition, Machine Translation, and Chatbots.

That’s the bridge between human language and machine understanding, enabling computers to read, understand, interpret and even generate human language. The discipline includes many tasks, each of which is unique and presents certain difficulties.

The first step in NLP is text segmentation or tokenization, which involves breaking down text into smaller components such as words, phrases, or sentences. Next is morphological analysis, which determines the grammatical attributes of the word, such as part of speech, tense, number, gender, etc. The next step is syntactic analysis, which involves making connections between words and phrases and building a syntactic tree of the sentence. This helps understand the structure and hierarchy of elements in a sentence. Syntactic analysis is followed by semantic analysis, where the machine attempts to understand the hidden meaning and meaning of words and phrases in context. The last stage is pragmatic analysis, which involves understanding the speaker's intention and using context to find out the true meaning of the message. This level of understanding requires complex analysis and is often the hardest for machines to achieve.

Together with that, all these steps provide the basic natural language understanding needed to perform more complex tasks such as machine translation, automatic summarization, text generation, etc. This makes NLP a fundamental building block for human-computer interaction in the field of artificial intelligence.

# History and development of language models

The development of language models began with early attempts to create machines capable of understanding and generating natural language. The story goes back decades, each with its own breakthroughs and discoveries.

The first steps in the field of natural language processing were taken in the 1950s and 1960s. Projects such as ELIZA and SHRDLU demonstrate a basic understanding of natural language, although they are limited to very narrow application areas. These early systems were based on simple rules and patterns, without the ability to learn or adapt. The 1970s and 1980s saw a shift from rules to statistical methods. It is recognized that machines need to analyze large amounts of text and learn from examples to understand natural language. This has led to the development of statistical models such as Hidden Markov Models (HMMs) and probabilistic context-free grammars. As the 1990s and early 2000s dawned, machine learning and neural networks emerged and revolutionized the approach to natural language processing. It ushered in the era of more complex algorithms, such as recurrent neural networks (RNN), and especially their variants, such as long short-term memory (LSTM), which allowed models to remember and use information over long periods of time. The introduction of the Transformer model in 2017 brought a huge leap forward, ushering in the era of Transformer-based language models such as BERT (Bidirectional Encoder Representation of Transformers) and GPT (Generative Pretrained Transformer). These models can significantly improve many NLP tasks by significantly improving understanding of context and semantics.

Today, language models are constantly evolving, becoming more powerful and versatile. By pre-training and developing transfer learning methods using huge data sets, it is possible to create models suitable for a variety of tasks.

The development of language models is a clear indication that the fields of artificial intelligence and natural language processing are constantly changing and pushing the boundaries of what is possible.

# 

# Modern language models

The current phase of language model development is characterized by the use of deep learning architectures and, in particular, based on transformers, which have significantly modified approaches to natural language processing.course.They have significantly improved context understanding and natural speech production, which has found many applications in a variety of tasks.

One of the first and most influential Transformer-based models was BERT (Bidirectional Encoding Regressions from Transformers), developed by Google in 2018. BERT used the attention mechanism to better understand the context of words in sentences, leading to significant performance improvements in tasks such as text classification, question answering systems, and machine translation.

The next step in the development of language models is the GPT (Generative Pre-training Transformer) from OpenAI. GPT-2 and GPT-3 templates have demonstrated highly natural text generation capabilities, opening up new horizons in areas such as automatic article generation, creative content creation, and even programming.

Recent developments in this field focus on technologies such as differentiable neural architectures, which allow models to independently adjust their structure to optimally solve specific problems . There is also a trend to create multilingual templates that can work with text in different languages ​​without prior configuration.

Modern language models open new horizons for many applications, such as advanced chatbots, advanced information retrieval, personalized news aggregation, and even innovative applications like creating music and writing.They are increasingly integrated into our everyday lives, providing more natural and richer ways of interacting with technology. These advances show that modern language models are not just data processing tools but also catalysts for new forms of communication and creativity in the age of artificial intelligence.

# 

# Applying Language Models

Language models are widely used today in many different fields, from information technology to the creative industries. Their ability to understand, interpret, and produce natural language opens the door to many creative applications.

### *Automatic machine translation*

Models have significantly improved the quality of machine translation, making it smoother and more natural. Systems like Google Translate are constantly being improved with the latest advances in language modeling.

### *Virtual assistants*

Siri, Alexa, and other virtual assistants use language modeling to understand user requests and generate natural responses. This allows users to communicate with devices as naturally as they communicate with other people.

### *Chatbots*

Many businesses use chatbots to automate customer service. Language models help chatbots understand queries and provide useful responses, thereby significantly improving service quality and reducing the burden on humans.

### *Sentiment Analysis*

Businesses and researchers use language modeling to analyze the emotional tone of text, allowing them to gauge public opinion, customer feedback or even predict changes in financial markets.

### *Speech recognition*

Language models are used to convert spoken language into text and are essential for developing accessible technology and improving human-machine interfaces.

### *Education*

Models build educational resources by providing personalized learning materials and adaptive tests to support effective student-centered learning.

These examples are just the tip of the iceberg in the wide range of applications of language models. As language models continue to be developed and refined, they will find even more innovative and impressive uses in everyday life and professional activities.

# 

# Summary in English

Natural language processing (NLP) is an important branch of artificial intelligence that aims to create systems that can understand, interpret and produce natural language in the same way as humans. An introduction to the field recognizes the ongoing development and importance of language models as an essential component of modern NLP systems.

The basics of natural language processing include elements such as morphological analysis, syntactic analysis, semantic analysis, and understanding context and intent. The history and development of language modeling shows an evolution from simple statistical approaches to more complex neural networks and transducers, especially models such as BERT and GPT, which have greatly improved our ability to understand and produce natural language.

Modern language models using deep learning architectures and transducers open new horizons for natural language processing tasks. These models are applied in fields ranging from machine translation and virtual assistants to sentiment analysis, automatic text generation and creative industries. These models have become an integral part of everyday life, improving the human-machine interface and opening up new possibilities for communication and creativity.

In conclusion, the field of natural language processing and language modeling has made significant progress and is expected to continue to evolve, providing increasingly sophisticated and integrated solutions to a wide range of problems and applications.

# 

# Summary translation

Обробка природної мови (NLP) є важливою галуззю штучного інтелекту, яка спрямована на створення систем, які можуть розуміти, інтерпретувати та створювати природну мову так само, як люди. Вступ до цієї галузі визнає постійний розвиток і важливість мовних моделей як важливого компонента сучасних систем НЛП.

Основи обробки природної мови включають такі елементи, як морфологічний аналіз, синтаксичний аналіз, семантичний аналіз і розуміння контексту та наміру. Історія та розвиток мовного моделювання показує еволюцію від простих статистичних підходів до складніших нейронних мереж і перетворювачів, особливо таких моделей, як BERT і GPT, які значно покращили нашу здатність розуміти та створювати природну мову.

Сучасні мовні моделі, що використовують архітектури глибокого навчання та перетворювачі, відкривають нові горизонти для завдань обробки природної мови. Ці моделі застосовуються в різних сферах: від машинного перекладу та віртуальних помічників до аналізу настроїв, автоматичного створення тексту та креативних індустрій. Ці моделі стали невід’ємною частиною повсякденного життя, покращуючи інтерфейс людини та машини та відкриваючи нові можливості для спілкування та творчості.

Загалом сфера обробки природної мови та моделювання мови досягла значного прогресу та, як очікується, продовжить розвиватися, забезпечуючи все більш складні та інтегровані рішення для широкого кола проблем і застосувань.

# References

1. Jurafsky, Dan, and James H. Martin. Speech and Language Processing. Pearson. 2021.
2. 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.
3. Language Processing and Python. (n.d.). Natural Language Toolkit, April 2023 [Електронний ресурс] – Режим доступу<https://www.nltk.org/book/ch01.html>
4. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems, 27.
5. Natural Language Processing | IBM Cloud Education. (n.d.). IBM, April 2023 [Електронний ресурс] – Режим доступу <https://www.ibm.com/cloud/learn/natural-language-processing>
6. Chowdhury, G. G. (2003). Natural language processing. Annual Review of Information Science and Technology, 37(1), 51-89.

# 

# Glossary

1. Artificial Intelligence (AI) - the simulation of human intelligence in machines programmed to think and learn.
2. Natural Language Processing (NLP) - a field of AI that focuses on the interaction between computers and human language.
3. Language Model - a statistical model that determines the probability of a sequence of words.
4. Deep Learning - a subset of machine learning involving neural networks with many layers.
5. Neural Network - a series of algorithms modeled after the human brain, designed to recognize patterns.
6. Transformer - a type of neural network architecture used primarily in the field of NLP.
7. BERT (Bidirectional Encoder Representations from Transformers) - a transformer-based model designed to understand the context of a word in search queries.
8. GPT (Generative Pretrained Transformer) - a type of language model that generates text based on input it has received.
9. Tokenization - the process of converting text into smaller units called tokens, which can be words, characters, or subwords.
10. Embedding - a representation of text in a high-dimensional space where similar words have a similar encoding.
11. Contextual Embedding - word embeddings that take into account the context in which a word appears.
12. Sequence to Sequence Model - a type of model in deep learning that transforms a given sequence of elements into another sequence.
13. Attention Mechanism - a component of neural networks that allows the model to focus on different parts of the input sequentially.
14. Machine Translation - the use of software to translate text or speech from one language to another.
15. Sentiment Analysis - the process of computationally identifying and categorizing opinions expressed in a piece of text.
16. Chatbot - a software application used to conduct an online chat conversation via text or text-to-speech.
17. Autocomplete - a feature that predicts the rest of a word or phrase while a user is typing.
18. Named Entity Recognition (NER) - a process in NLP for identifying and classifying key elements in text.
19. Part-of-Speech Tagging - the process of marking up a word in a text as corresponding to a particular part of speech.
20. Syntax Tree - a tree representation of the syntactic structure of a sentence.
21. Corpus (plural: corpora) - a large and structured set of texts used in linguistics and NLP.
22. Pretraining - the process of training a model on a large dataset before fine-tuning it on a specific task.
23. Fine-tuning - the process of taking a pre-trained model and adapting it to a specific task with additional training.
24. Generative Model - a model that can generate new data instances.
25. Discriminative Model - a model that predicts a label for a given instance.
26. Backpropagation - a method used in artificial neural networks to calculate the gradient of the loss function.
27. Gradient Descent - an optimization algorithm used to minimize some function by iteratively moving towards the steepest descent.
28. Overfitting - a modeling error in deep learning when a function is too closely fit to a limited set of data points.
29. Underfitting - a situation where a model cannot capture the underlying trend of the data.
30. Hyperparameter - a parameter whose value is set before the learning process begins.
31. Regularization - a technique used to prevent overfitting by adding additional information to a model.
32. Loss Function - a method of evaluating how well an algorithm models the given data.
33. Activation Function - a function applied to each neuron in the neural network, determining whether it should be activated or not.
34. Recurrent Neural Network (RNN) - a class of neural networks where connections between nodes form a directed graph along a temporal sequence.
35. Long Short-Term Memory (LSTM) - a type of RNN capable of learning long-term dependencies.
36. Convolutional Neural Network (CNN) - a deep learning algorithm commonly used for image recognition.
37. Transfer Learning - the practice of reusing a pre-trained model on a new problem.
38. Supervised Learning - a type of machine learning where the model is trained on labeled data.
39. Unsupervised Learning - a type of machine learning where the model is trained on unlabeled data.
40. Semi-supervised Learning - a class of machine learning tasks that involve some labeled data but mostly unlabeled data.
41. Reinforcement Learning - a type of machine learning where an agent learns to behave in an environment by performing actions and seeing the results.
42. Anomaly Detection - the identification of items, events, or observations that do not conform to an expected pattern.
43. Natural Language Understanding (NLU) - a subfield of NLP focused on machine reading comprehension.
44. Natural Language Generation (NLG) - a subfield of NLP focused on generating natural language from data.
45. Question Answering - a computer science discipline within AI and NLP concerned with building systems that automatically answer questions posed by humans.
46. Speech Recognition - the process of converting spoken words into text.
47. Text-to-Speech (TTS) - a type of assistive technology that reads digital text aloud.
48. Word Embedding - a learned representation for text where words with the same meaning have a similar representation.
49. Cross-lingual Transfer - the application of models trained in one language to other languages.
50. Bias in AI - systematic and unfair discrimination against certain individuals or groups.
51. Ethical AI - the field of study concerned with ensuring that artificial intelligence systems act as ethically as possible.
52. Federated Learning - a machine learning approach where the model is trained across multiple decentralized devices or servers holding local data samples, without exchanging them.
53. Autoregressive Model - A type of statistical model that bases its predictions on previous values in the data sequence.
54. Beam Search - A heuristic search algorithm that explores a graph by expanding the most promising node in a limited set.
55. Corpus Linguistics - The study of language as expressed in samples (corpora) or "real world" text.
56. Dependency Parsing - The process of analyzing the grammatical structure of a sentence by identifying its dependent and governing words.
57. Entity Linking - The task of disambiguating mentions of entities in text to a known entity in a database.
58. Gating Mechanisms - Techniques used in recurrent neural networks to control the flow of information.
59. Heuristic - A technique designed for solving a problem more quickly when classic methods are too slow, or for finding an approximate solution when classic methods fail to find any exact solution.
60. Language Identification - The task of determining the language that a piece of text is written in.
61. Lexical Semantics - The study of how and what the words of a language denote.
62. Morphological Analysis - The study of the structure and formation of words.
63. N-gram - A contiguous sequence of n items from a given sample of text or speech.
64. Perplexity - A measurement of how well a probability model predicts a sample.
65. POS Tagging - See Part-of-Speech Tagging (already included in the initial list).
66. Pragmatics - The branch of linguistics dealing with language in use and the contexts in which it is used, including such matters as deixis, taking turns in conversation, text organization, presupposition, and implicature.
67. Semantic Parsing - The process of converting a natural language sentence into a logical form: a machine-understandable representation of its meaning.
68. Stemming - The process of reducing inflected (or sometimes derived) words to their word stem, base or root form.
69. Stop Words - Commonly used words (such as "the", "a", "an", "in") which a search engine has been programmed to ignore.
70. Syntax - The arrangement of words and phrases to create well-formed sentences in a language.