Ye. Kanchura, PhD in Phil., As. Prof.
V. Korol, Student
Zhytomyr State Polytechnic University

TRANSLATION FROM THE PERSPECTIVE OF ARTIFICIAL INTELLIGENCE: HOW HUMANS AND MACHINES SEE TEXT DIFFERENTLY

One of the key innovations in the recent AI development is the advancement of large language models (LLMs) that demonstrated impressive results in a variety of natural language processing (NLP) tasks, such as contextual understanding of human language, aggregation of large textual data, and emulating different styles of speech. However, one prominent field that is still in early development is machine translation whereby an AI system aims to translate text from one language to another preserving context, realia and handling difficult-to-translate cases.

Historically, there were two main approaches used up to this date: statistical and deep learning (DL) translation. The statistical method was founded by Warren Weaver where the translator system would scan different tokens (words) from a variety of texts and try to translate them taking into account different meanings that it took on in its dataset. This method is commonly used to translate single words or phases and could allow its users to see broader meaning a lexeme can possess, but it ultimately fails to capture the dependencies and abstract relationships between tokens that constitute human thinking. One famous tool using statistical translation is ReversoContext.

The other approach was translation based on deep learning, oftentimes using recurrent neural networks (RNNs). As a branch of computer science, RNNs leverages concepts from a multitude of mathematical concepts, such as calculus, linear algebra and statistics, to figure out a set of numbers (weights) that when plugged into a linear summation would produce the expected result. Virtually all DL models that process text represent it as indexes in a dictionary, and as its final value they produce a number that corresponds to a word in an embedded dictionary which is then retrieved and returned. RNN translation handles its task considerably better than statistical translation, albeit it's only effective as long as the text complexity is low, and it fails to do more higher-level translation that involve conveying an indirect human thought, such as the politeness level, jokes, or poetry, primarily because it processed the text sequentially and was trying to predict the next most likely word, which couldn't consider non-linear relations between components of the sentence. Google Translate is one of the most well-known examples of this technology.

One of the most groundbreaking discoveries in the domain was the self-attention mechanism by A. Vaswani et al that took a different approach to how language was processed [1, 2]. The main innovation in self-attention was finding a way to represent the meaning of the entire text, referred to as the attention scores, for which the neural network could iterate through each token and query its contextual weight in the whole text. As the result, this technique allowed self-attention-based AI models to capture relations between words across distances in the text and build up the intertextual representation, which contributed greatly to its ability to process language in a way closer to how humans do. The most well-known model based on this discovery was a series of OpenAI GPT models that today power a large number of modern AI tools.

The success of Generative Pretrained Transformer (GPT) was in the fact that it could convey multiple dimensions of meaning of tokens into multiple separate units. This way, researchers can use mathematical tools to transform and extract meaningful concepts from the text. GPT, in particular, utilises 3 vectors to represent a single token:

- **value vector** encodes the lexical meaning of word;
- **key vector** encodes contextual shades how the token can be used;
- **query vector** encodes semantic properties of token in regards with others.

In conclusion, it would be possible to convey more nuanced and inclusive relations between language tokens by adding more vectors, such as evaluating subtle hints from language patterns, infer additional information, process behaviour patterns and ultimately, produce more accurate translations by querying and matching more explicit information about cultural context and seeking the relationships between words.

REFERENCES

1. Attention Is All You Need [Електронний ресурс] / [A. Vaswani, N. Shazeer, J. Uszkoreit та ін.]. – 2017. – Режим доступу до ресурсу: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

2. Olah C. Attention and Augmented Recurrent Neural Networks [Електронний ресурс] / C. Olah, S. Carter. – 2016. – Режим доступу до ресурсу: https://distill.pub/2016/augmented-rnns.

3. Weaver W. Translation / Warren Weaver. // MIT Press. – 1949. – С. 17–20.