ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Transformer-based Neural Network Machine Translation Model for the Kurdish Sorani Dialect

Soran Badawi

Language Center, Charmo Center for Scientific Research & Consulting, Charmo University, Chamchamal, Sulaimani, KRG, Iraq

ABSTRACT

The transformer model is one of the most recently developed models for translating texts into another language. The model uses the principle of attention mechanism, surpassing previous models, such as sequence-to-sequence, in terms of performance. It performed well with highly resourced English, French, and German languages. Using the model architecture, we investigate training the modified version of the model in a low-resourced language such as the Kurdish language. This paper presents the first-ever transformer-based neural machine translation model for the Kurdish language by utilizing vocabulary dictionary units that share vocabulary across the dataset. For this purpose, we combine all the existing parallel corpora of Kurdish – English by building a large corpus and training it on the proposed transformer model. The outcome indicated that the suggested transformer model works well with Kurdish texts by scoring (0.45) on bilingual evaluation understudy (BLEU). According to the BLEU standard, the score indicates a high-quality translation.

Index Terms: Machine translation, Transformers, Dialect, Kurdish language, Bilingual evaluation understudy

1. INTRODUCTION

Human language has a complex and irregular system that can pose significant issues for machine translation. The kind of morphemes, their implications, and their syntactic and semantic relations in the context is causing the natural language to be complex and abnormal. The complexity of these problems has led some to believe that human translation is infeasible for such tasks.

Historically, machine translation has experienced numerous changes. First, dictionary-based and rule-based translation methods were developed and provided translation services through the manual specification of rules and resources [1]. Following that, statistical translation emerged as the new model to diminish the role of a linguist and increase the emphasis on language dependency [2].

Luckily, the advancement of neural networks and artificial intelligence has primarily impacted many areas of science, including machine translation. As a result of the neural machine translation research, top-notch translations were produced for texts written in resourceful languages. Therefore, the need to achieve the same goal for low-resourced languages has become significant and the attempts to achieve that have increased [3]. Languages are considered less resourced when they lack human-constructed linguistic resources, substantial monolingual or parallel corpora, and general-purpose grammar are the sole sources available. The research industry has primarily ignored Kurdish dialects, which are practiced by 20–30 million people across four regions [4].

Access this article online

DOI: 10.21928/uhdjst.v7n1y2023.pp15-21

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Copyright © 2022 Badawi. This is an open access article distributed under the Creative Commons Attribution Non-Commercial No Derivatives License 4.0 (CC BY-NC-ND 4.0)

Corresponding author's e-mail: Soran Badawi, Language Center, Charmo Center for Scientific Research & Consulting, Charmo University, Chamchamal, Sulaimani, KRG, Iraq. E-mail: soran.sedeeq@charmouniversity.org

Received: 12-10-2022 Accepted: 24-12-2022 Published: 15-01-2023

This study presents a transformers-based model using the vocabulary dictionary concept. We collect the parallel corpora in the language and merge them to be a large corpus for training and report the results. The resources used for the task include the Tanzil corpus [5], TED corpus [6], KurdNet–the Kurdish wordnet [7], and the Auta corpus [8].

2. RELATED WORKS

Few studies have addressed the Kurdish language in the Machine Translation (MT) domain. The Apertium project is the first machine translation system for both Sorani and Kurmanji. The Apertium uses rules-based machine translation, which has developed various tools and resources for the Kurdish language, such as bilingual and morphological dictionaries, structural transfer rules, and grammar [4]. InKurdish1 is another attempt to construct a machine translation model for Kurdish. The system applies dictionary-based methods for translation. According to Taher *et al.* (2017), this method is ineffective in translating lengthy and idiomatic sentences. Finally, Ahemdi and Mansoud (2020) attempted to translate Kurdish texts using neural machine translation [4]. Their work was based on collecting the parallel datasets in the Kurdish language. They used different tokenization techniques for training the dataset. They eventually reported the Bilingual evaluation understudy (BLEU) achieved using each tokenizer. Regarding other low-resourced languages worldwide, Abbott and Martinus (2018) employed transformer models to translate texts from English to Setswana using the parallel Autshumato dataset [9]. The outcome of their work indicated that the transformer outperforms previous methods by 5.33 BLEU points. Moreover, Przystupa and Abdul-Mageed (2019) used transformer models with back-translation. Their results demonstrate that transformer models translate texts between Spanish–Portuguese and Czech–Polish [10]. Tapo *et al.* (2019) used neural machine translation to translate texts from Barbara's language to English and French. Their work mainly concentrated on the challenges when performing neural machine translation on a low-resourced language such as Barbara [11].

2.1. Dataset

We used a collection of four parallel datasets. The first one is Tanzil, a group of Quran translations compiled by the Tanzil project8. The corpus has one Sorani translation aligned with 11 translations, totaling 92,354 parallel texts with 3.15M vocabularies on the Sorani Kurdish side and 2.36M on the English side. The corpus is available in translation memory exchange (TMX), where aligned verses are offered [4].

The second corpus, the TED corpus [6], is the collection of subtitles from TED Talks, a sequence of top-notch talks on different genres, "Technology, entertainment, and design." The only Kurdish dialect for which these subtitles are translated is the Sorani dialect. Even though there are only 2358 parallel sentences, the TED collection has translations in a broader, more comprehensive range of subjects than Tanzil.

The third corpus is WordNet [12], a lexical-semantic tool exploited for various Natural Language Processing (NLP) tasks like information extraction and word disambiguation. WordNet offers concise definitions and uses examples for groupings of synonyms, also known as synsets, in addition to semantic links like synonymy, hyponymy, and meronymy. Kurdish WordNet [7] is based on a semi-automatic technique that focuses on creating a Kurdish alignment for base concepts, a critical subset of WordNet's central meanings. Four thousand six hundred and sixty-three definitions directly translated from the Princeton WordNet are included in the most recent version of KurdNet (version 3.0). We included this resource despite having fewer translated purposes than necessary for machine translation because it covers more domains.

The final corpus is Auta, comprising 229,222 pairs of physically aligned translations [8]. The corpus is gathered from different text genres and domains to construct more solid and real-world machine translation applications. The researchers built this corpus and published a portion of this corpus available to promote study in this area, which contains 100.000 normalized and cleaned texts ready to be experimented with using the trendy machine learning models (Table 1).

2.2. Transformer's Model Architecture

Most neural machine translation models follow an encoder-decoder structure [13]. The encoder consists of six identical layers with a multi-head self-attention mechanism and position-wise sublayers. These layers are fully connected to feed-forward networks. The encoder aims to map an input sequence of symbol representations starting from (x1;:; xn) to a sequence of continuous representations, which is z = (z1;:; zn) [14].

Table 1: Size of each Kurdish–English corpus

No	Corpus	Language	Size
1	Tanzil	Kurdish–English	92.354 texts
2	Ted	Kurdish–English	2.358 texts
3	KurdNet	Kurdish–English	4.663 texts
4	Auta	Kurdish–English	100.000 texts
Total			199.375 texts

Similarly, the decoder is constituted of a stack of six identical layers. Encoder layers consist of two sublayers each, and the decoder adds the third sublayer to carry out multi-head attention around the encoder`s output. In the same way as the encoder, layer normalization uses residual connections around each sub-layer. The self-attention sub-layer in the decoder stack has been adjusted to block positions from attending to the following positions, as shown in Fig. 1. The goal of the decoder is to produce a sign output sequence (y1;:; ym) one element at a time [14].

Moreover, the model uses the attention function to map a query and a set of key-value pairs to an output, where the question, keys, values, and production are all vectors. The sum of the weight values calculates the result. A compatibility function between the query and the relevant key determines each value's weight.

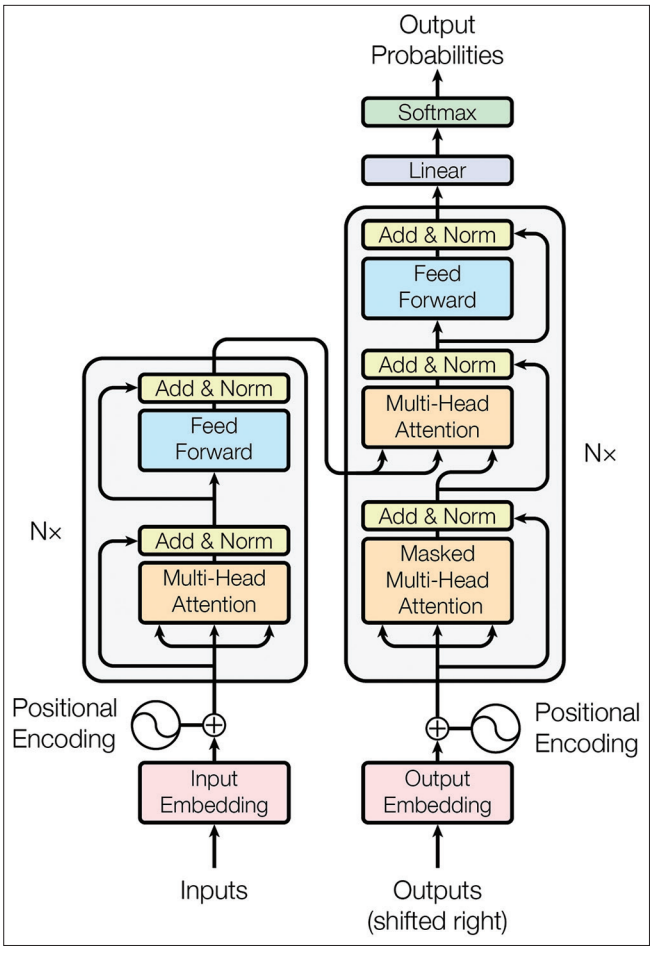

Fig. 1. The transformer model architecture [15].

As is shown, the transformer operates multi-head attention on three different stages:

1. Encoder-decoder attention lets every decoder position focus on every input post.
2. The encoder has layers for the self-attention. All the keys, values, and queries in a self-attention layer originate from the same source, in this case, the encoder's output from the previous layer.
3. The decoder's self-attention layers enable each location to pay attention to all postings below and above.

A fully connected feed-forward network is implemented to each position separately and uniformly in each layer of the encoder and decoder. Two linear transformations and a ReLU (Rectified Linear Unit) activation make up this process [15].

The decoder output is transformed to project next-token probabilities using the SoftMax function. The embedding is utilized to convert the input and output tokens to vectors of the dimension model.

The positional encodings to the input embedding at the bottoms of the encoder and decoder stacks. Since the positional encodings and the embeddings share the same dimension model, both can be added. Positional encodings come in a variety of discovered and fixed forms [16].

2.3. The Proposed Model Architecture

The decoder and encoder for the proposed transformer-based Neural Machine Translation (NMT) have a stack of six layers, as shown in Fig. 2. Every layer has two sublayers: The position-wise feed-forward sub-layer and the multi-head attention sub-layer (FFN). The encoder and decoder in the proposed Transformer NMT model architecture for Kurdish texts produce variable-length sequences using an attention model and feed-forward net. Multi-head attention is the foundation for how attention operates across multiple tiers. The mapping of an input sequence of symbol representations, X = (x1, x2..., xnenc)T to an intermediate vector. Given the intermediate vector, the decoder generates the output sequence (target sentence) Y = (y1, y2..., yndec)T. The convolutional or recurrent structures are absent from the transformer design. At the first layer of both the encoder and the decoder, the positional encodings computed by the Equations below are summed to the input embeddings.

1) PE(pos, 2i) = sin(pos100002i/dmodel)
2) PE(pos, 2i+1) = cos(pos100002i/dmodel)

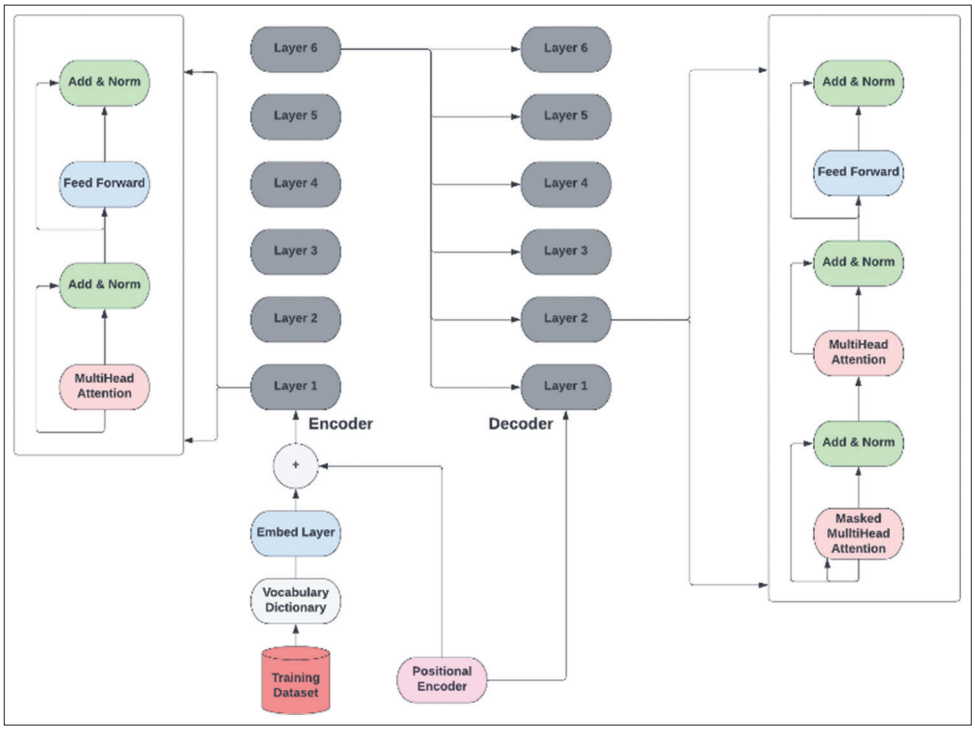

Fig. 2. The Proposed model architecture.

Where pos stands for position, i is considered the dimension, and d is the dimension of the intermediate representation. Every encoding layer has a position-wise feed-forward sub-layer and a multi-head attention sub-layer. A residual connection method [17] and a layer normalization unit (LayerNorm) [13] are used around each sub-layer to facilitate training and enhance performance. In contrast to the encoder, every layer of the decoder has three sub-layers: a multi-head attention sublayer, a position-wise feed-forward sub-layer, and so on. Encoder-decoder multi-head attention sub-layer is inserted in between them.

3. METHODOLOGY

The proposed model uses the concept word dictionary inside the dataset to find the equivalence meaning of each word. Therefore, in the preprocessing stage, we only tokenized the cleaned texts, which converted the sentences into lists of words. Following that, we converted them into an extensive dictionary of words which has Kurdish words and their English meanings. Next, we fed the dictionary to our proposed transformer's model. As shown in Fig. 3. We used the batch size of 20 and trained on 100 epochs.

At first, we tried to train the model on the central processing unit (CPU); since the amount of data was huge for the CPU, the model trained for days without providing any results, and numerous ram crashes forced the computer to reboot and restart the process again. However, we tried to train the model for one epoch and compare its result with graphics processing unit (GPU). As it is shown in Table 2.

As shown in Table 2, training one epoch on the CPU lasted 3 h and 37 min, while it lasted <5 min for GPU. Because 100 epochs are enormous to be trained on CPU and to avoid Ram crashes, we trained the model on Google Colab Pro, a monthly subscription program that gives you higher Ram and GPU. The whole training and test process lasted 5 h. The complete code and the training program are publicly available at/https://github.com/mbrow309/MachineTranslationUsingTransformers/blob/master/KurdishMTTransformers.ipynb/.

4. RESULTS AND DISCUSSION

We train the system on 100 epochs since introducing the MT module at higher values will help guarantee a good BLEU score. Neural networks are usually trained over several

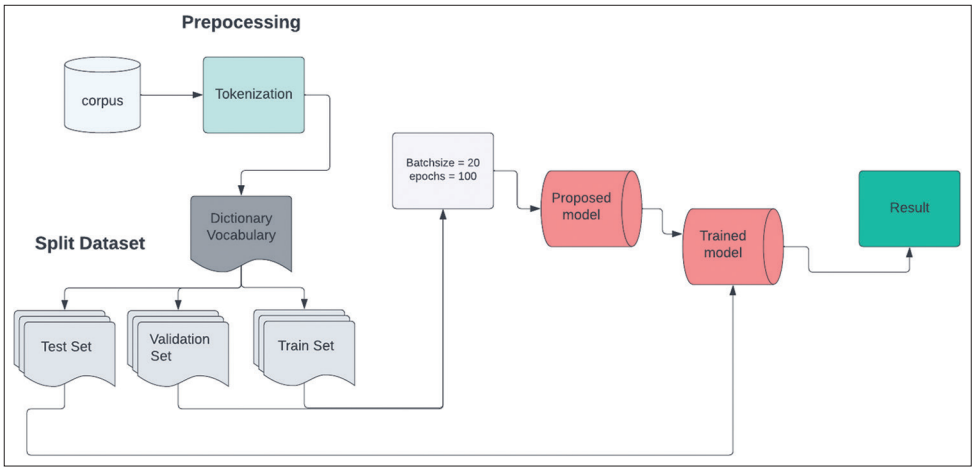

Fig. 3. An outline of the methodology.

Table 2: The difference between training on GPU and CPU

Machine type	Loss	BLEU	Time
CPU	5.140	0.0183	3 h and 37 min
GPU	5.109	0.0191	4 min

GPU: Graphics processing unit, CPU: Central processing unit, BLEU: Bilingual evaluation understudy

Table 3: Samples of the produced translation texts

Kurdish	من تۆم خۆشدەوێت
English	I love you, and I love you.
Kurdish	من دەچم بۆ بازار
English	I am going to go to the marketing game.
Kurdish	ڕۆژێک لە ڕۆژان نەخۆشیەکی ترسناک بوو بە هەڕشە
English	Once upon a time, there was a dread disease.
Kurdish	تیوە لە کوێ بوون
English	Where were you, like yesterday?

epochs. Epochs refer to cycles through a training dataset [18]. It is important to note that we tried to train the module on each dataset, and due to the low amount of datasets, the module yielded a significantly lower BLEU score. However, merging the datasets did a perfect job. As shown in Fig. 4, the amount of BLEU improves significantly per 10 epochs and finally reaches the ideal score.

We fed the module some unseen texts to translate. Overall, the module did an excellent job of translating the texts. Below are samples of translated texts shown by the module.

The model does a relatively good job of translating unseen texts. Even though the translation results from Table 3 show some cases of word repetition and some cases of producing ungrammatical sentences, particularly in the final test example. The issue is substantially related to having a low amount of data. Therefore, if the model is trained on much larger datasets, the translation results would be more accurate and flawless.

In the next phase of our work, we intend to investigate the ergative case of our model by feeding it examples that have ergative and compare our model`s translation with the latest Google translation for the Kurdish language. In Kurdish, the word order is subject-object-verb with tense-aspect-modality markings [19]. As a split-ergative language, Sorani Kurdish marks transitive verbs in the past tenses differently from nominative verbs [20]. For ergative-absolute alignment, Sorani Kurdish uses different pronominal enclitics [4]. To clarify further, we have included a few examples in Sorani Kurdish below. The bold suffix is used for patient marking in Example 1 in the past tense, which uses the pronominal enclitic = man as an agentive marker.

1. Kurdish/منداڵەکانمان هێنان
Transcribe/mndalakanman hênan.

Translation/we brought the children.

2. Kurdish/هێنامانن
Transcribe/hêna**man**in

Translation/we brought them.

3. Kurdish/دەچنە باخەکامان

Transcribe/deçine baxakaman

Translation/they are going to our garden.

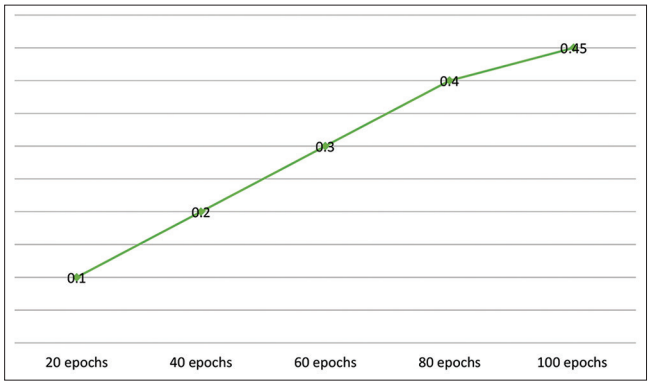

Fig. 4. The value bilingual evaluation understudy value per 20 epochs.

Table 4: Comparison between translating ergative sentences using Google translator and transformer-based model

Google translator	Our model	Kurdish text
Sutandmian brought it	They burnt me A two into difficulties	سوتاندمیان هێنایانی
They took them to the market.	They took me to the market.	بردمیان بۆ بازار
They brought them home.	They took me home	هێنامیان بۆ ماڵ
I saw them in the market.	I saw them at the market.	بینیانم لە بازار
I love you.	I love you.	من تۆم خۆشدەوێت

As shown, the pronominal (man) is translated differently in different examples. In examples 1 and 2, "man" was the subject, and its equivalence is "we." While in example 3, it functions as a possessive pronoun, and it is "our." During machine translation, this creates significant issues when the model tries to align the two languages.

The examples in Table 4 show that our model performs well in the ergative situation for Kurdish texts. According to the results, the Google translator faces issues when the sentence contains the pronominal enclitics (m), and it functions as the object of the sentence. This is because our corpus includes many natural language texts that include such pronominal pronouns, particularly in the Tanzil corpus. Thus, our model would easily detect the pronominal enclitics and their alignment inside the texts.

5. CONCLUSION

The transformer model is a unique and highly functional model to translate texts from one language to another. Undoubtedly, the Kurdish language suffers from a lack of resources, particularly in the field of NLP. The lack of a translation model is also part of the problem. The work undertaken in this paper demonstrates that the Kurdish language responds well to the newly developed and proposed neural machine translation model. It is worth noting that the existence of large corpora with more than 1 million data can actively work well and improve the model's score to near-perfect translation. Fortunately, the results acquired from this work can open many gates for the future researchers to dive deeply into the transformer model and modified in a way that can work specifically for the language. Finally, the transformer model's layers remain intact, and the training and process started this way as the model modification, particularly on the layers left for future researchers.

REFERENCES

[1] S. Tripathi and J. K. Sarkhel. "Approaches to machine translation". *Annals of Library and Information Studies*, vol. 57, pp. 383-393, 2010.

[2] P. Koehn. "*Statistical Machine Translation*". Cambridge University Press, Cambridge. 2009.

[3] L. Bentivogli, A. Bisazza, M. Cettolo and M. Federicoa. "Neural versus phrase-based mt quality: An in-depth analysis on english-german and english-french". *Computer Speech and Language,* vol. 49, pp. 52-70, 2019.

[4] S. Ahmadi and M. Masoud. "Towards Machine Translation for the Kurdish Language". *arXiv preprint arXiv:2010.06041,* 2020.

[5] J. Tiedemann. "Parallel data, tools and interfaces in OPUS". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12*). European Language Resources Association (ELRA), Istanbul, Turkey. pp. 2214-2218, 2012.

[6] M. Cettolo, C. Girardi and M. Federico. "Wit3: Web inventory of transcribed and translated talks". In: *Conference of European Association for Machine Translation*. 2012.

[7] P. Aliabadi, M. S. Ahmadi, S. Salavati and K. S. Esmaili. "Towards building kurdnet, the kurdish wordnet". In: *Proceedings of the Seventh Global Wordnet Conference*. University of Tartu Press, Tartu, Estonia. 2014.

[8] Z. Amini, M. Mohammadamini, H. Hosseini, M. Mansouri and D. Jaff. "Central Kurdish Machine Translation: First Large Scale Parallel Corpus and Experiments". *arXiv preprint arXiv:2106.09325,* 2021.

[9] L. Martinus and J. Z. Abbott. "A Focus on Neural Machine Translation for African Languages". *arXiv preprint arXiv:1906.05685,* 2019.

[10] M. Przystupa and M. Abdul-Mageed. "Neural machine translation of low-resource and similar languages with backtranslation". In: *Proceedings of the Fourth Conference on Machine Translation*. vol. 3. Association for Computational Linguistics, Florence, Italy. 2019.

[11] A. A. Tapo, B. Coulibaly, S. Diarra, C. Homan, J. Kreutzer, S. Luger, A. Nagashima, M. Zampieri and M. Leventhal. "Neural Machine Translation for Extremely Low-Resource African Languages: A Case Study on Bambara". *arXiv preprint arXiv:2011.05284,* 2019.

[12] G. A. Miller. "*WordNet: An Electronic Lexical Database*". MIT Press,

Massachusetts, United States. 1998.

[13] J. L. Ba, J. R. Kiros and G. E. Hinton. “Layer Normalization”. *arXiv preprint arXiv:1607.06450,* 2016.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin. “Attention is all you need”. In: *Conference on Advances in Neural Information Processing Systems.* 2017.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov. “Dropout: A simple way to prevent neural networks from overfitting”. *The Journal of Machine Learning Research*, vol. 15, pp. 1929-1958, 2014.

[16] J. Gehring, M. Auli, D. Grangier, D. Yarats and Y. N. Dauphin. “Convolutional sequence to sequence learning”. In: *Proceedings of the 34th International Conference on Machine Learning* (PMLR). 2017.

[17] M. Shafiq and Z. Gu, “Deep Residual Learning for Image Recognition: A Survey,” *Applied Sciences*, vol. 12, no. 18, p. 8972, 2022.

[18] L. N. Smith. “Cyclical learning rates for training neural networks”. In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Santa Rosa, CA, USA. 2017.

[19] G. Hai and Y. Matras. “Kurdish linguistics: A brief overview”. *STUF-Language Typology and Universals*, vol. 55, pp. 3-14, 2002.

[20] M. R. Manzini, L. M. Savoia and L. Franco. “Ergative case, aspect and person splits: Two case studies”. *Acta Linguistica Hungarica*, vol. 52, pp. 297-351, 2015.