Prilagodba modela transformatora za strojno prevođenje u južnoslavenskim dijalektima

Haramija, Gašpar

Undergraduate thesis / Završni rad

2024

UNIVERSITY OF ZAGREB

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

BACHELOR THESIS No. 1530

ADAPTING TRANSFORMER MODELS FOR MACHINE TRANSLATION IN SOUTH-SLAVIC DIALECTS

Gašpar Haramija

Zagreb, June 2024

UNIVERSITY OF ZAGREB

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

BACHELOR THESIS No. 1530

ADAPTING TRANSFORMER MODELS FOR MACHINE TRANSLATION IN SOUTH-SLAVIC DIALECTS

Gašpar Haramija

Zagreb, June 2024

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Zagreb, 04 March 2024

BACHELOR THESIS ASSIGNMENT No. 1530

Student: **Gašpar Haramija (0036538084)**

Study: Electrical Engineering and Information Technology and Computing

Module: Computing

Mentor: prof. Jan Šnajder

Title: **Adapting transformer models for machine translation in South-Slavic dialects**

Description:

Language models demonstrate a strong grasp of various languages and dialects, yet encounter challenges with dialects from less-represented languages due to limited training data availability. Given the hurdles in gathering data for these dialects, it becomes imperative to explore strategies for model adaptation to enhance their performance. This thesis aims to adapt the transformer model for machine translation in South Slavic dialects with the aim of enhancing translation quality. To achieve this goal, the research will utilize the COPA dataset encompassing both standard languages (Slovenian, Croatian, Serbian) and dialects (Cerkno, Torlak, Chakavian) to fine-tune transformer-based models for machine translation. Specifically, it will implement and evaluate various fine-tuning techniques to enhance translation quality. All references must be cited, and source code, documentation, and executables must be included with the thesis.

Submission date: 14 June 2024

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

ZAVRŠNI ZADATAK br. 1530

Pristupnik:	**Gašpar Haramija (0036538084)**
Studij:	Elektrotehnika i informacijska tehnologija i Računarstvo
Modul:	Računarstvo
Mentor:	prof. dr. sc. Jan Šnajder

Zadatak: **Prilagodba modela transformatora za strojno prevođenje u južnoslavenskim dijalektima**

Opis zadatka:

Jezični modeli pokazuju izvanredno razumijevanje jezika i dijalekata, ali pokazuju poteškoće s dijalektima manje zastupljenih jezika zbog ograničene dostupnosti podataka za predtreniranje. S obzirom na poteškoće u skupljanju podataka za takve dijalekte, potrebno je istražiti strategije prilagodbe modela radi poboljšanja njihovih performanci. Cilj završnoga rada prilagodba je modela transformatora za strojno prevođenje u južnoslavenskim dijalektima radi poboljšanja kvalitete prijevoda. Koristit će se skup podataka COPA na standardnim jezicima (slovenski, hrvatski, srpski) i dijalektima (cerkljanski, torlački, čakavski) za fino ugađanje modela nad modelima arhitekture transformatora koji su specijalizirani za strojno prevođenje. Implementirati postupak finog ugađanja transformatora na specifične dijalekte južnoslavenskih jezika za poboljšanje kvalitete prijevoda. Dodatno, provesti analizu učinkovitosti različitih pristupa finog ugađanja na kvalitetu prijevoda u južnoslavenskim dijalektima. Radu je potrebno priložiti izvorni kod, rezultate i programsku dokumentaciju te citirati korištenu literaturu.

Rok za predaju rada: 14. lipnja 2024.

I want to thank everyone who helped and supported me in completing this work. I’m especially grateful to my mentor Jan Šnajder and Ana Barić for their guidance and support.

Contents

1 Introduction

The field of machine translation has made significant improvements in recent years, largely due to advancements in neural network architectures such as transformer models. These models have revolutionized natural language processing by providing state-of-the-art results in various translation tasks. However, neural machine translation remains a challenging task when only a limited amount of data is available, which is the case with low-resource languages and dialects. Unlike standard languages, dialects often span over small geographical areas, with distinct micro-dialects spanning areas as small as single-digit square kilometers and they are mainly used in daily interactions and social media [1, 2, 3].

The biggest problem in facing dialect translation and translation for low-resource languages is data availability since many of these dialects are primarily used in conversational settings. Consequently, they do not have large-scale parallel corpora nor a big presence on the web [4, 5]. Various techniques such as back-translation and self-supervised learning can effectively reduce the reliance and need for large-scale parallel corpora [6, 3].

There are approximately 30 million speakers of South Slavic languages, mainly in the Balkans. The South Slavic languages constitute a dialect continuum – a series of language varieties across some geographical area such that neighboring varieties are mutually intelligible. There are three subgroups of South Slavic languages: Western, Eastern and Transitional.

The goal of this thesis is to adapt a model for machine translation of South Slavic languages, more precisely, Eastern South Slavic and Transitional South Slavic Dialects [8, 9]. Figure 1.1 shows a map of the South Slavic language area and Slavic dialects spoken

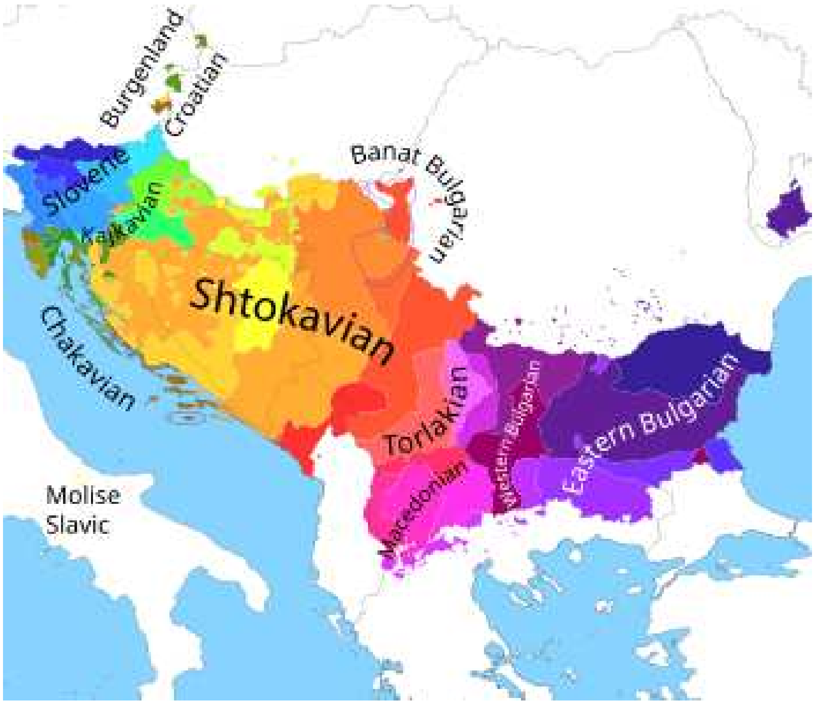

Figure 1.1: South Slavic dialect continuum [7]

there.

The primary objective of this thesis is to adapt the No Language Left Behind (NLLB) [10] model for machine translation of South Slavic dialects: (1) Cerkno dialect in Slovenia, (2) Chakavian dialect in Croatia, and (3) Torlakian dialect in Serbia.

This involves fine-tuning the NLLB model on South Slavic language data, applying the technique of back-translation to augment existing data to enhance model performance, and comparing the translation results of the fine-tuned model with those of an unmodified model.

To achieve these objectives, a series of experiments were conducted. Initially, the NLLB model was fine-tuned using a COPA dataset [11] specifically curated for South Slavic dialects. Also, a back-translation method was implemented to augment existing data and to further refine the translation quality. The performance of these adaptations was then evaluated and compared to the baseline NLLB model to determine their effectiveness.

The remainder of this thesis is organized as follows: Chapter 2 reviews related work, Chapter 3 covers the theoretical background, and Chapter 4 details the datasets. Chapter

5 outlines the methodology, while Chapter 6 presents the results and analysis. Finally, Chapter 7 concludes the thesis with a summary and future research directions.

2 Related Work

Over the past decade, machine translation, the automated conversion of written text from one natural language to another, has undergone a significant transformation. Traditionally dominated by statistical machine translation, which heavily relied on diverse count-based models, the field has now transitioned to neural machine translation. This approach, powered by deep learning techniques, has emerged as the dominant paradigm, marking a substantial shift in translation research methodologies [12].

In the field of neural machine translation, various strategies have emerged to address the challenges posed by low-resource languages. These include techniques like transfer learning [13], back-translation [6], and multilingual machine translation [10], each offering unique solutions to improve translation quality.

Additionally, researchers are exploring how to translate dialects, like Arabic [1] and various German dialects [4] among many others. They focus on finding ways to deal with the lack of data and the complexity of language structure in these dialects. Also, they describe the challenge of evaluation: due to minimal changes in the dialectal orthography, the exact word matching implemented in the BLEU metric often fails. This problem has also been detected for morphologically rich languages such as South Slavic languages [4, 8].

Moreover, considerable research has been conducted on translation among closely related South Slavic languages. Specifically, attention has been directed towards the shared traits of South-western Slavic languages such as Slovenian, Croatian, and Serbian, as well as the accumulation of parallel and monolingual datasets [9]. This research has demonstrated that employing back-translation yields promising outcomes, improving model performance [9].

In addition, there is a focus on Croatian dialects including Kajkavian, Shtokavian, and Chakavian. This research primarily delves into unsupervised neural machine translation between the standard language and dialects, both ways. Notably, this work contributes to the aggregation of monolingual data valuable for dialectal studies [3].

Also, important for this thesis is the research and model called "No Language Left Behind" (NLLB) [10]. This project aimed to ensure high-quality translation for over 200 languages, many of which are low-resourced. They did this by making new datasets and models to help improve translation quality for languages with fewer resources. While the model covers South Slavic languages, it doesn't include the smaller dialects in this area, making it a suitable choice for comparing the aforementioned machine translation techniques.

3 Theoretical Background

This chapter aims to introduce key theoretical concepts essential for comprehending the thesis and its analysis. Beginning with an exploration of specific South Slavic dialects, the chapter proceeds to explain different methods used in machine translation, with a special emphasis on technologies used and how they can be tailored for specific tasks. Furthermore, it delves into data augmentation methods, particularly focusing on back-translation. Finally, the chapter concludes by discussing the evaluation techniques employed in this study.

3.1 South Slavic Languages

As shown in Figure 1.1, there's a variety of South Slavic dialects across the Balkan region. In this study, we focus on three dialects: Chakavian, Torlakian, and Cerkno. Each of these has many smaller micro-dialects, making it hard to pin down specific lexical, typological, and morphological rules for each one. The focus is on translating between the standard languages Croatian, Slovenian, and Serbian, and these three dialects. Here are the specific micro-dialects primarily looked at:

- the Cerkno dialect of Slovenian, spoken in the Slovenian Littoral region, specifically from the town of Idrija;

- the Chakavian dialect of Croatian from northern Adriatic, specifically from the town of Žminj;

- the Torlak dialect from southeastern Serbia, northeastern North Macedonia, and northwestern Bulgaria, specifically from the town of Lebane.

3.2 Machine Translation

There are approximately 7,000 languages spoken around the world. Some aspects of language seem to be common across all these languages, while others are more common among most of them. For instance, all languages have nouns and verbs, ways to ask questions, and ways to express agreement or disagreement. However, languages also have many differences, like how words are ordered, their vocabulary, and how they form words [14].

Understanding why languages differ like this can help us create better machine translation models. This is important because there's a high demand for translation systems. They're mainly used for accessing information, but lately, they're also being used for real-time communication between people [14].

The typical method used for machine translation is called the encoder-decoder network which is represented in Figure 3.1. These models are handy when we need to convert one sequence into another, especially when the output sequence depends on the whole input sequence. In machine translation, the words in the target language might not match the words in the source language in terms of number or order. This highlights the differences in language structure and vocabulary between languages and dialects.

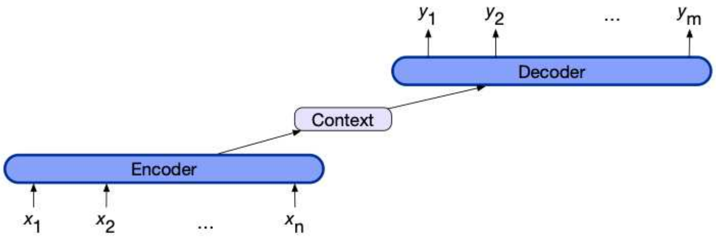

Figure 3.1: The encoder-decoder architecture [14]

Machine translation relies on supervised machine learning. During training, the system is provided with a collection of parallel sentences, enabling it to learn the mapping from source to target sentences. Instead of using complete words, sentences are divided into sequences of smaller units called tokens, which can include words, subwords, or individual characters. The encoder-decoder architecture consists of two main components: an encoder and a decoder as can be seen in Figure 3.1. The encoder processes the input words and generates an intermediate context representation. During decoding, the system utilizes this context to generate the output one word at a time.

Machine translation systems utilize a predefined vocabulary, where words are tokenized using algorithms designed for subword tokenization. This shared vocabulary encompasses both source and target languages. To construct this vocabulary, subword tokenization algorithms are applied to a corpus containing data from both languages. Modern systems often employ advanced tokenization algorithms like the wordpiece algorithm. For example, in Table 3.1,the sentence is tokenized using the wordpiece algorithm [15]:

Sentence	Jet makers feud over seat width with big orders at stake.
Tokenized Sentence	_Jet _makers _feud _over _seat _width _with _big _orders _at _stake

Table 3.1: Sentence and its tokenized form

3.3 Transformers and Large Language Models

In this section, the concept of transformer architecture is introduced, a common framework for building large language models and the algorithm behind most modern Natural Language Processing systems. The core of the transformer architecture is a mechanism called attention and self-attention. In this thesis, the NLLB model [10] is used and fine-tuned. Following this, we explore fine-tuning techniques and the relevant background information.

3.3.1 Transformer Architecture

The transformer architecture [16] represents a significant advancement in the field of Natural Language Processing, forming the foundation of many contemporary large language models. Its primary innovation lies in the attention mechanism, particularly self-attention, which enables the model to weigh the importance of different words in a sentence, regardless of their position [14].

As can be seen in Figure 3.2, when processing each item in the input, the model has access to all previous inputs, including the current one, but not to any future inputs. Additionally, the computations for each item are independent of one another. This

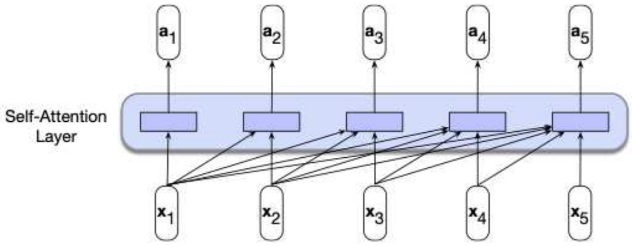

Figure 3.2: Information flow in causal self-attention model [14]

independence allows for easy parallelization of both forward inference and training processes.

Transformers consist of multiple transformer blocks, each being a multilayer network that maps sequences of input vectors to sequences of output vectors of the same length. These blocks combine simple linear layers, feedforward networks, and self-attention layers [14].

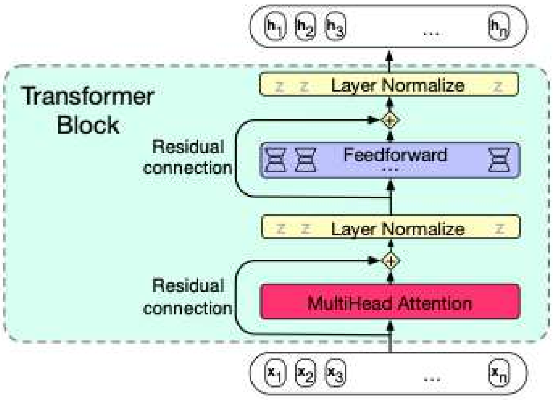

Figure 3.3: A transformer block showing all the layers [14]

The feedforward layer in transformers consists of position-wise networks, each being a fully connected two-layer network with one hidden layer. These networks are independent and can be computed in parallel. Residual connections allow information to bypass intermediate layers, improving learning and giving higher-level layers direct access to lower-layer information. Layer normalization improves training performance by normalizing summed vectors [14]. Dimensions of input and output vectors of transformer blocks are the same to allow stacking. Large language models use many stacked layers of these blocks.

The NLLB model is a specific implementation of the transformer architecture, designed to provide high-quality translation across more than 200 languages, many of which are low-resource. It aims to bridge the performance gap between low-resource and high-resource languages by leveraging extensive datasets and advanced training techniques [10].

The NLLB also has its tokenizer and vocabulary of 256,204 tokens. It breaks down words into smaller units called tokens, which can be as small as a character or as large as a word, to efficiently process text. Each token has a corresponding position in the vocabulary. The model uses a tokenizer to convert text into these token IDs and back. Each token is represented by a numerical vector known as an embedding, and the NLLB-200–600M model that is used in this thesis has 256,204 such embeddings, each being a 1024-dimensional vector. These embeddings are trained along with the neural network.

3.3.2 Fine Tuning

In general, fine-tuning involves taking a pretrained model and further training it, often by adding a neural network classifier that uses the model's top layer as input, to perform specific tasks like named entity recognition or question answering. The idea is that the pretraining phase equips the model with rich word representations, making it easier to adapt to the specific requirements of a downstream task. Fine tuning is a form of transfer learning, where knowledge gained from one task or domain is applied to solve a different task. In fine-tuning, applications are built on top of pretrained models by adding a small set of parameters specific to the application. This process involves using labeled data related to the application to train these additional parameters. Usually, the pretrained language model parameters are either frozen or only slightly adjusted during this training [14].

During the fine-tuning process, different settings for both training and optimization are used. An optimizer is responsible for adjusting the model's parameters to minimize the difference between the predicted outputs and the actual targets:

- **Learning Rate (lr):** Determines how quickly the model learns. A lower learning rate means slower but potentially more stable learning;

- **Clip Threshold:** Prevents the model from making large parameter updates during training;

- **Weight Decay:** A regularization technique that prevents the model from overfitting.

Different training configurations influence the training process:

- **Batch Size:** Defines how many examples the model processes at once during training, larger batch size requires more memory;

- **Total Training Steps:** Specifies the total number of training iterations the model will go through.

3.4 Data Augmentation

Data augmentation is a statistical technique used to enhance training data by generating synthetic data from existing natural data. Most of the world's languages lack large parallel training texts. This scarcity presents a significant challenge for achieving quality translations in lesser-resourced languages. Two common approaches to address data sparsity in machine translation are back-translation and multilingual models [14].

The most prevalent data augmentation method in machine translation is back-translation. While parallel corpora in low-resource languages may be limited, larger monolingual corpora are often available. The back-translation process involves creating synthetic parallel datasets using monolingual corpora in the target language.

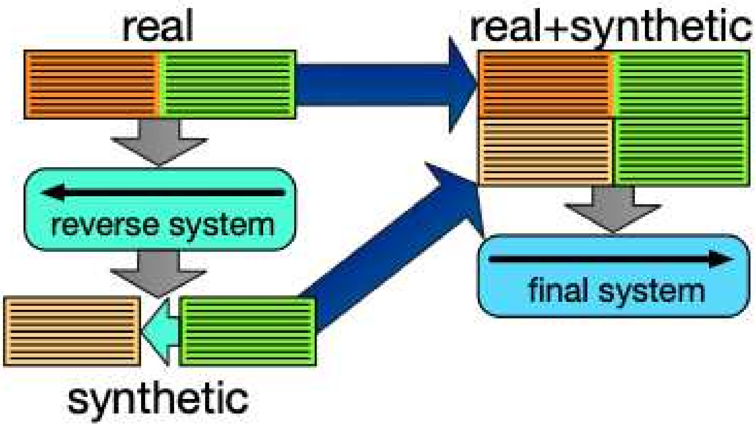

Figure 3.4: Creating a synthetic parallel corpus through back-translation [6]

Here are the steps for the back translation method [14, 6]:

1. **Train a reverse machine translation system** – using available parallel texts in the source and target languages, train a machine translation system to translate from the target language back to the source language;

2. **Generate synthetic dataset** – use this train machine translation system to translate the monolingual target data into the source language;

3. **Augment training data** – add these synthetic parallel texts to the original training data and retrain the machine translation model.

Studies indicate that back translation is a highly effective technique for improving machine translation [5]. There is also evidence to suggest that this process can be iteratively applied to further enhance performance [6].

3.5 Evaluation

Human evaluations of machine translation are thorough but resource-intensive. Conducting human evaluations can be time-consuming, often taking months to complete, and requires human labor that cannot be recycled for other tasks. Consequently, automatic metrics are commonly employed for their convenience. While automatic metrics are not as precise as human evaluation, they serve as useful tools for testing potential system enhancements [14, 17]. In this work, two such metrics will be used and explained: BLEU and chrF.

3.5.1 chrF metric

The chrF metric [18], short for character F-score, is a straightforward and robust method for evaluating machine translation quality. It assesses each machine translation target sentence based on the overlap of character n-grams with the corresponding human translation.

To calculate chrF, a parameter k is specified to determine the length of character n-grams considered. It computes the average precision (chrP) and recalls (chrR) for each n-gram length (from unigram to k-gram) as follows [18]:

- chrP: The percentage of character n-grams in the machine translation hypothesis that also appears in the reference translation, averaged across all n-gram lengths;

- chrR: The percentage of character n-grams in the reference translation that is found in the MT hypothesis, averaged across all n-gram lengths.

Using a weighting parameter β, the metric combines chrP and chrR to calculate the F-score. A common choice is to set $\beta = 2$, giving more weight to recall [18]:

$$\text{chrF}_\beta = \frac{(1+\beta^2)\cdot \text{chrP}\cdot \text{chrR}}{\beta^2 \cdot \text{chrP} + \text{chrR}}$$

For $\beta = 2$, it simplifies to:

$$\text{chrF}_2 = \frac{5\cdot \text{chrP}\cdot \text{chrR}}{4 \cdot \text{chrP} + \text{chrR}}$$

ChrF is a simple and effective evaluation method that correlates very well with human judgments in many languages [19].

3.5.2 BLEU metric

Before chrF, another commonly used overlap metric in machine translation evaluation was BLEU (BiLingual Evaluation Understudy) [17]. Unlike chrF, BLEU is a word-based metric that focuses solely on precision, without combining precision and recall [17]. The BLEU score for a corpus of candidate translation sentences is calculated based on the n-gram word precision across all sentences, along with a penalty computed over the entire corpus.

Due to its word-based nature, BLEU is highly sensitive to word tokenization, making it challenging to compare systems that rely on different tokenization standards. Additionally, BLEU may not perform as effectively in languages with complex morphology [14].

To calculate BLEU, a precision-based approach is used to compare n-grams of the machine translation hypothesis with n-grams of the reference translations. The BLEU score considers n-gram precision for up to 4-grams by default and includes a brevity penalty to penalize translations that are too short, where n-gram precision is the proportion of n-grams in the machine translation hypothesis that also appear in the reference translations [17].

To prevent very short translations from achieving high precision, BLEU includes a brevity penalty (BP). The brevity penalty is calculated as follows

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases}$$

where c is the length of the candidate translation, and r is the length of the reference translation.

The BLEU score combines n-gram precision with the brevity penalty:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where p_n is the precision of n-grams of length n, N is the maximum n-gram length, and w_n is the weight assigned to n-gram precision.

4 Datasets

This chapter presents the datasets used in this study, which are critical for fine-tuning and evaluating the NLLB model for specific South Slavic dialects. Three distinct datasets are employed. The COPA Dataset is used for fine-tuning the NLLB model, focusing on specific South Slavic dialects and standard languages. The Monolingual Dataset consists of texts in three Croatian dialects and is later used in the process of back-translation. The Parallel Evaluation Dataset, also in Croatian dialects and standard language, is used to evaluate the performance of various models. The following sections provide detailed descriptions of these datasets, including their sources, characteristics, and how they are applied in this research.

4.1 COPA Dataset

The Choice Of Plausible Alternatives (COPA) dataset [11] is a significant resource in natural language processing research, specifically designed to assess commonsense causal reasoning. Inspired by large-scale question sets used in previous studies, the COPA dataset consists of one thousand English-language questions. Each question is presented in a forced-choice format, providing a premise and two plausible causes or effects [11]. The task is to select the alternative that is more plausible than the other. This dataset serves as an important benchmark for evaluating models' ability to understand and reason about everyday scenarios and causal relationships.

Building on the foundation of the original COPA dataset, there have been ongoing efforts to translate it into various languages, making it accessible for a broader range of linguistic research. One such initiative is the translation of the COPA dataset for the Shared Task at the VarDial Workshop 2024, specifically targeting South Slavic dialects [2]. This translation effort aims to adapt the dataset for evaluating commonsense reasoning in these dialects. Although the COPA dataset is primarily intended for commonsense reasoning tasks, it also offers valuable parallel sentence pairs in South Slavic dialects. This parallel data can be effectively used in machine translation research. For this study, the COPA dataset is used to fine-tune the NLLB model for specific South Slavic dialects.

The specific COPA dataset used in this research includes 500 triplets of sentences, each consisting of a premise, an alternative 1, and an alternative 2. Sentences in the triplet are contextually connected, providing a rich source of parallel sentences. For this study, only the texts from the original COPA dataset will be extracted. These triplets will then be separated, transforming the dataset from 500 entries into 1500 individual sentences. Consequently, this will result in 1500 distinct sentences in South Slavic standard languages and dialects.

Language	**Translation**
English	The man turned on the faucet. Water flowed from the spout.
Slovenian	Moški je odprl pipo. Iz ustja pipe je pritekla voda.
Cerkno dialect	Dic je adparu pipa. Iz pipe je partjekla uoda.
Croatian	Muškarac je otvorio slavinu. Voda je potekla iz mlaznice.
Chakavian dialect	Muški je otpra špino. Oda je počela teć z mlaznici.
Serbian (transliterated)	Čovek je otvorio slavinu. Voda je tekla iz slavine.
Torlak dialect (transliterated)	Čovek odvrnuja slavinu. Voda ističala od slavinu.

Table 4.1: Example sentences from the COPA dataset in different South Slavic languages and dialects. [2]

In the dataset presented in Table 4.1, there are examples from South Slavic standard languages and dialects. Specifically, the table includes sentences in Serbian, Slovenian, and Croatian, as well as the Cerkno, Chakavian, and Torlak dialects. The table provides an example of a premise and the correct alternative for that premise, illustrating how the sentences within a triplet can be contextually connected.

Furthermore, it is important to note that for Serbian and Torlak, there are two versions of each sentence: one in the Cyrillic alphabet and one transliterated into the Latin alphabet. The inclusion of these two versions is crucial because the NLLB model used in this research is pretrained in Serbian Cyrillic, making it essential to have the sentences in this script for optimal performance. The transliterated versions are provided for reference and to give a broader understanding of the dataset.

4.2 Monolingual Dataset

To further this research, a monolingual dataset is collected from previous work aimed at studying translation between modern Croatian and its dialects [3]. The Croatian language is characterized by three main dialect groups: Shtokavian, Kajkavian, and Chakavian, each encompassing more specific local dialects. These groups are geographically distributed across Croatia. The cited work shows the importance of monolingual data for low-resource languages like Croatian. The Croatian language is primarily based on the Shtokavian dialect but also includes Kajkavian and Chakavian dialects. These dialects are named after their pronouns: "što", "kaj", and "ča".

All the resources that were found in Kajkavian, Shtokavian, and Chakavian are from before the standardization of the Croatian language, consisting mostly of old poems and dramas. The researchers gathered data from various sources [3]. From these sources, the researchers extracted individual sentences, resulting in a cleaned dataset of over 53,000 sentences.

Dialect	**Sentence**
Chakavian	Ostavila san ga bome. Kajiš stegni, strah pritegni, zube stisni, pa zakorači. Lipu obilatu primalitnju kišu koja je razveselila zemju i jude.
Kajkavian	Em nikaj ni slajše Ne čuje se rajše Neg dobri i dragi naš kaj. Zakaj se srdite gospon Matek.
Shtokavian	Kad je baka ovo čula zamisli se teško. I što da vam kažem. Pusti da se do kraja za grijehe pokajem.

Table 4.2: Examples from the monolingual dataset with respective dialects

Table 4.2 shows examples from the monolingual dataset, examples of different Croatian dialects present in the dataset, including Chakavian, Kajkavian, and Shtokavian. The Chakavian dialect is the primary focus of interest in this dataset. However, additional effort will be necessary to filter out Kajkavian and Shtokavian sentences to prioritize the Chakavian dialect for further analysis and processing.

4.3 Parallel Evaluation Dataset

The final dataset, though the smallest, serves for additional testing of the performance of models. Acquired through the same research on Croatian dialects mentioned earlier [3], this dataset comprises 60 sentences in Croatian dialect. While monolingual data sufficed for training, there is a need for parallel data for testing. In the mentioned research they collaborated with a human Croatian specialist to create parallel sentences in Croatian dialects.

Language	Sentence
Croatian	Po lijepom vremenu boso se išlo, a sandale su nosile kada je padala kiša.
Chakavian	Po lipom vrimenu boso se išlo, a čačule nosile kade je kišilo.
Croatian	Možeš li se sjetiti mama, tih jutra, tako plavih?
Kajkavian	Zmisliš se, mama, tih jutrah tak plavih?
Croatian	Vezao me za nju i dao mi žestok napor.
Chakavian	Pri njer me jur sveza i žestok trud zada.
Croatian	Proklet bio, tko te rodio, da te nije k sreći pustio.
Chakavian	Proklet bia, ko te je rodia, da te nije u sriću puštia.

Table 4.3: Examples from the parallel dataset with respective dialects

Table 4.3 demonstrates differenet Croatian dialects, including Chakavian, Kajkavian, and standard Croatian, present in the dataset. Further efforts will be required to manually select sentences in the Chakavian dialect for use in the research of this work.

5 Methodology

The primary focus of this research is to explore and improve the capabilities of the multilingual NLLB model for translating South Slavic dialects. This chapter will discuss the methodologies used in this process. Initially, the preprocessing steps undertaken to prepare the datasets for training and evaluation are explained. Also, the following techniques used are detailed: baseline model evaluation, fine-tuning with specific datasets, and back-translation for data augmentation.

5.1 Data Preprocessing

In the data preprocessing stage, several steps were undertaken to prepare the datasets for training and evaluation. For the COPA dataset, sentences were extracted and three distinct datasets were created, each consisting of 1,500 entries. These newly created datasets are parallel corpora of the following: Croatian-Chakavian, Serbian-Torlak, and Slovenian-Cerkno. All of them were split into training and test data. Due to the limited size of the corpora, only 10% were used for test data. The remaining 90% were used for training data. The same test set was used for all evaluations in the following experiments, ensuring consistency and comparability across different methodologies and models.

The original monolingual dataset, which initially contained 53,000 sentences, was reduced to 39,000 sentences following the preprocessing phase. This reduction was achieved through a heuristic approach that focused on identifying and retaining sentences in different dialects present within the dataset. The primary objective was to maximize the proportion of Chakavian sentences compared to Kajkavian and Shtokavian ones. This was primarily accomplished by examining the usage of distinct pronouns in each dialect, such as "kaj", "što", and "ća". Additionally, all sentences shorter than 4 words or longer than 25 words were excluded from the dataset. Some sentences were also handpicked and removed manually due to their lack of semantic meaning. The entire dataset was first utilized in the process of back-translation, translating from the dialect to the standard language and then back to the dialect. This process generated new parallel corpora, which were subsequently used for fine-tuning the Croatian-Chakavian translation model.

The parallel evaluation dataset, which initially contained 60 sentences, was reduced to 44 sentences by removing those in dialects other than Chakavian. This selection process relied on the author's native knowledge of the Croatian language. This dataset was used exclusively for the evaluation of the models, as it provides a contextually distinct set of sentences compared to the other datasets.

5.2 Baseline

In this section, the focus is on the baseline model. The baseline used is the NLLB model without any additional configurations. A significant challenge is that the model's vocabulary does not include any South Slavic dialects. The issue is solved by using a language classifier to detect the language of each sentence in the training sets of the dialect corpora. Based on these results, the languages with the most sentences are selected for translation.

The following approaches were used in the dialect-to-standard language translations and vice versa:

- Chakavian-Croatian: Approximately half of the sentences are recognized as Slovenian and the other half as Croatian. When translating the test set, the model randomly chooses between translating from Croatian or Slovenian for the Chakavian dialect;

- Cerkno-Slovenian: Similar to Chakavian, approximately half of the sentences are recognized as Slovenian and the other half as Croatian. The model randomly chooses between translating from Croatian or Slovenian for the Cerkno dialect;

- Torlak-Serbian: The majority of sentences are recognized as Serbian, so the translation for the entire test dataset was accordingly set to Serbian.

5.3 Individual Models for the Dialects

In this section, the process of fine-tuning the NLLB model for each specific South Slavic dialect is detailed. A primary challenge in this process is the absence of South Slavic dialects in the model's vocabulary. Additionally, the model's tokenizer does not include language tags for these dialects.

Before adding language tags, an analysis was made to determine if additional tokens needed to be incorporated into the model's vocabulary. This analysis involved examining the average number of tokens per word.

Language/Dialect	Average Number of Tokens per Word
Croatian	1.63
Chakavian	1.52
Slovenian	1.57
Cerkno	1.63
Serbian	1.69
Torlak	1.78

Table 5.1: Average Number of Tokens per Word for Each Language/Dialect

As shown in Table 5.1, the average number of tokens per word for each dialect is similar to that of the standard languages included in the NLLB model. This similarity suggests that the translation quality of the fine-tuned model would be adequate without the need to extend the vocabulary. Consequently, language tags for each dialect were created and added to the model's tokenizer. By adding a new tag, its embedding is initialized for each dialect to its respective standard language.

The fine-tuning process for each dialect involved optimizing the model using the Adafactor optimizer. The configuration details in Table 5.2 describe the fine-tuning process applied uniformly across all dialects.

By applying configurations in Table 5.2 across all dialects, the fine-tuning process was standardized. This consistent approach helped in achieving comparable improvement in the translation quality for each South Slavic dialect.

Optimizer Configuration	
Learning Rate (lr)	1e-4
Clip Threshold	1.0
Weight Decay	1e-3
Training Configuration	
Batch Size	16
Maximum Sequence Length	128
Total Training Steps	1500

Table 5.2: Configuration details for individual models for the dialects

5.4 Back-translation

The next step involves using the fine-tuned model for back-translation. The trained model, designed for translating between Chakavian and Croatian, is applied to a collection of Chakavian sentences that have been preprocessed. These sentences are first translated into Croatian and then back into Chakavian, creating new sets of parallel sentences. These newly generated sentences are then used to further train the model using configurations in Table 5.3.

Optimizer Configuration	
Learning Rate (lr)	1e-4
Clip Threshold	1.0
Weight Decay	1e-3
Training Configuration	
Batch Size	32
Maximum Sequence Length	128
Total Training Steps	15000

Table 5.3: Configuration details for back-translation model

Through this process, a new model is refined on back-translated data specifically for translating between Chakavian and Croatian. The performance of this model will be compared and evaluated with other Chakavian-Croatian models.

6 Results and Analysis

This section presents the results of experiments conducted to evaluate machine translation models. The aim is to analyze translation quality achieved by both baseline and fine-tuned models, using evaluation metrics such as BLEU and chrF2. The evaluation involves comparing baseline models with fine-tuned models. This comparison helps assess the impact of fine-tuning on translation accuracy. Additionally, the results of the back-translation process specifically for the Chakavian dialect are analyzed.

Translation Direction	**Model**	**BLEU**	**chrF2**
Chakavian-Croatian	Baseline	*6.68*	*31.20*
	Fine-Tuned	*34.02*	*55.76*
Cerkno-Slovenian	Baseline	*6.70*	*27.17*
	Fine-Tuned	*19.77*	*41.91*
Torlak-Serbian	Baseline	*15.22*	*39.97*
	Fine-Tuned	*24.77*	*49.93*
Croatian-Chakavian	Baseline	*3.49*	*25.51*
	Fine-Tuned	*24.97*	*51.54*
Slovenian-Cerkno	Baseline	*2.37*	*22.80*
	Fine-Tuned	*21.35*	*45.36*
Serbian-Torlak	Baseline	*8.44*	*40.45*
	Fine-Tuned	*24.22*	*50.53*

Table 6.1: Performance of baseline and fine-tuned models for South Slavic dialects

Table 6.1 shows the performance of both baseline and fine-tuned models for South Slavic languages, assessed through BLEU and chrF2 metrics, each ranging from 0 to 100, where higher scores are better. Notably, the BLEU scores for baseline models across all language directions remain below 10, except for Torlak-Serbian, suggesting inadequacy at the word level. However, chrF2 scores for baseline models range from 22.80 (Slovenian-Cerkno) to 40.45 (Serbian-Torlak), showing slightly improved performance at the character level.

Additionally, analysis reveals that baseline models exhibit better BLEU and chrF2 scores for dialect-to-standard language translations compared to the reverse direction. On average, these scores are higher by 4.77 in BLEU and 3.19 in chrF2 for dialects translated to standard language. This finding suggests that the NLLB model, without additional configurations, shows better translation from dialects to standard language. Moreover, the notably better results for Torlak-Serbian and Serbian-Torlak may indicate more linguistic similarity between the dialect and standard language. Also, it needs to be considered that these translations primarily involve Serbian as both source and target languages, unlike Cerkno and Chakavian, where Slovenian and Croatian were used, respectively.

In further analyzing Table 6.1, when considering the fine-tuned models, chrF2 metrics are generally outperforming BLEU metrics. There's a noticeable improvement in performance for the standard language-to-dialect direction compared to the dialect-to-standard language direction. However, a distinct difference is observed for Chakavian and Croatian, where the dialect-to-language direction exhibits better performance in both BLEU and chrF2 scores.

When comparing baseline and fine-tuned models, the fine-tuned models consistently demonstrate superior performance across all metrics. On average, the fine-tuned models outperform the baseline models by 17.70 points in BLEU scores and 17.99 points in chrF2 scores. Particularly, the fine-tuned models achieve a chrF2 score of 49.17 on the character level, indicating adequate translation quality. However, the BLEU score averages 24.85, suggesting challenges in translating South Slavic languages and dialects due to their rich morphology.

Model	**Chakavian-Croatian**		**Croatian-Chakavian**	
	BLEU	**chrF2**	**BLEU**	**chrF2**
Baseline	*6.68*	*31.20*	*3.49*	*25.51*
Fine-Tuned	*34.02*	*55.76*	*24.97*	*51.54*
Back-translation	*39.10*	*61.43*	*41.35*	*62.61*

Table 6.2: Performance of models for Chakavian-Croatian and Croatian-Chakavian translation

Table 6.2 shows the performance of the back-translation model for Chakavian-Croatian and Croatian-Chakavian language directions, showcasing superior performance compared to both the baseline and fine-tuned models on the same test set. For the Chakavian-Croatian direction, the BLEU score is 39.10, and the chrF2 score is 61.43. This performance surpasses the fine-tuned model by 5.08 points in the BLEU score and 32.42 points in the chrF2 score while outperforming the baseline model by 6.67 points in the BLEU score and 28.23 points in the chrF2 score. For the Croatian-Chakavian direction, the BLEU score is 41.35, and the chrF2 score is 62.61. This performance demonstrates a significant improvement over the fine-tuned model by 16.38 points in the BLEU score and 11.07 points in the chrF2 score. These results favor the back-translation model, showing the highest quality translation.

Translation Direction	**BLEU**	**CHRF**
Chakavian-Croatian	*7.49*	*30.83*
Croatian-Chakavian	*2.32*	*25.87*

Table 6.3: Performance of back-translation model on Parallel evaluation dataset for Chakavian-Croatian and Croatian-Chakavian translation

Table 6.3 presents the performance of the back-translation model on the Parallel Evaluation Dataset for Chakavian-Croatian and Croatian-Chakavian translation directions, with BLEU scores of 7.49 and 2.32, respectively. These results are comparable to the baseline model results, showing no significant improvement. This suggests potential issues in the methodology of this study, where the domain specificity of the monolingual and COPA datasets, or the Parallel Evaluation Dataset's domain specificity compared to other datasets, may pose challenges. Also, the evaluation dataset's small size, having only 44 sentences, can be a problem.

Model	**Output**	**Expected Output**
Baseline	I njen sin je u pustinji. Upalio sam svijeću.	Njezin je sin pao s kreveta. San je težak za županu.
Fine-Tuning	Čuo je njezin miris. Je da šoldi na criekvo.	Osjetio je miris njezina parfema. Je da šoldi criekve.
Back-Translation	Skočili su po krevetu. Je poštiva pravila svojih roditelja.	Skakala su po krevetu. Je poštivala pravila svojeh roditelji.

Table 6.4: Translation performance comparison for Croatian and Chakavian

Table 6.4 illustrates the performance of various models across different sentence pairs. Examples from the baseline models highlight instances where the semantic meaning of words is often overlooked, a limitation less prevalent in other models. However, challenges persist across all models, particularly due to the colloquialisms in dialects, which pose difficulties for accurate translation. Additionally, the table highlights how the morphological complexity of language and dialects significantly impacts translation quality. These subtle linguistic differences can lead to discrepancies in BLEU scores, which operate at the word level, while chrF2 scores are less affected.

Furthermore, the results highlight how the back-translation model performs well in one scenario but poorly in another, indicating limitations in its generalization capability. However, despite these challenges, the results of the back-translation model are in favor of using this technique, particularly in the context of low-resource languages where access to parallel datasets is limited or even non-existent.

To enhance this research, having monolingual data for other South Slavic dialects could make significant improvements. Consequently, by having multiple back-translation models for more dialects, clearer conclusions could be drawn. Also, expanding the size of parallel datasets would better the training of fine-tuned models for specific dialects and languages, potentially leading to more accurate translations. Future research could explore the development of multilingual models encompassing other South Slavic dialects, offering a comprehensive approach to translation within this linguistic domain.

7 Conclusion

This work presents several experiments aimed at improving machine translation for low-resourced languages, with a particular focus on South Slavic dialects: Chakavian, Cerkno, and Torlak. Dialect translation encounters two primary challenges: the lack of parallel corpora and the conversational nature of dialects, which leads to inconsistencies in linguistic rules.

Three different approaches based on the NLLB model were explored: the baseline model, fine-tuned model, and back-translation model, all using the COPA dataset comprising 1500 parallel sentences in South Slavic languages and dialects. Also, a monolingual dataset of little over 50,000 sentences was used in the back-translation process, augmenting the dataset for fine-tuning.

Individual models were trained for each of the three South Slavic dialects using corresponding COPA datasets. The fine-tuned models outperformed the baseline models by 17.70 points in BLEU scores and 17.99 points in chrF2 scores. The back-translation model demonstrated superior performance, outperforming the fine-tuned models by 6-16 BLEU points and 11-28 chrF2 points.

However, results from the parallel evaluation dataset, independent of COPA and monolingual data, were less promising. The performance of the back-translation model on this dataset did not show promising results, potentially influenced by dataset size and contextual differences compared to other datasets.

It's important to recognize that there's room for improving machine translation for South Slavic dialects. Future work could prioritize expanding dataset sizes and making new ones. Additionally, future research could explore new techniques and extend them to other dialects within the South Slavic language family.

References

[1] H. Sajjad, A. Abdelali, N. Durrani, and F. Dalvi, “Arabench: Benchmarking Dialectal Arabic-English Machine Translation,” in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020. https://doi.org/10.18653/v1/2020.coling-main.447

[2] D. V. 2024, “Dialect-COPA VarDial 2024,” n.d. [Online]. Available: https://sites.google.com/view/vardial-2024/shared-tasks/dialect-copa

[3] B. Penkova, M. Mitreska, K. Ristov, K. Mishev, and M. Simjanoska, “Learning Translation Model to Translate Croatian Dialects to Modern Croatian Language,” in *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, 2023. https://doi.org/10.23919/mipro57284.2023.10159848

[4] L. Lambrecht, F. Schneider, and A. Waibel, “Machine Translation from Standard German to Alemannic Dialects,” in *Proceedings of SIGUL2022 @LREC2022*, 2022, pp. 129–136.

[5] J. Maillard, C. Gao, E. Kalbassi, K. R. Sadagopan, V. Goswami, P. Koehn, A. Fan, and F. Guzman, “Small Data, Big Impact: Leveraging Minimal Data for Effective Machine Translation,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023. https://doi.org/10.18653/v1/2023.acl-long.154

[6] V. C. Hoang, P. Koehn, G. Haffari, and T. Cohn, “Iterative Back-Translation for Neural Machine Translation,” in *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, 2018. https://doi.org/10.18653/v1/w18-2703

[7] G. Golob, “Map of South Slavic Dialects,” https://commons.wikimedia.org/w/index.php?curid=131242888, 2023, cC BY-SA 4.0. [Online]. Available: https://commons.wikimedia.org/w/index.php?curid=131242888

[8] S. Kordić, “Jezik I Nacionalizam (Language and Nationalism),” *SSRN Electronic Journal*, 2010. https://doi.org/10.2139/ssrn.3467646

[9] M. Popović and A. Poncelas, “Neural Machine Translation between Similar South-Slavic Languages,” ADAPT Centre, School of Computing, Dublin City University, Ireland, Tech. Rep., 2020.

[10] N. Team, M. R. Costa-Jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, and J. ... Wang, “No Language Left Behind: Scaling Human-Centered Machine Translation,” Computer Software, 2022.

[11] M. Roemmele, C. A. Bejan, and A. S. Gordon, “Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning,” in *Logical Formalizations of Commonsense Reasoning — Papers from the AAAI 2011 Spring Symposium (SS-11-06)*, 2011.

[12] F. Stahlberg, “Neural Machine Translation: A Review,” *Journal of Artificial Intelligence Research*, vol. 69, pp. 343–418, 2020.

[13] B. Zoph, D. Yuret, J. May, and K. Knight, “Transfer Learning for Low-Resource Neural Machine Translation,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016. https://doi.org/10.18653/v1/d16-1163

[14] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Stanford University, 2024.

[15] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young,

J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. S. Corrado, M. Hughes, and J. Dean, "Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation," 2016.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010. [Online]. Available: https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001. https://doi.org/10.3115/1073083.1073135

[18] M. Popović, "CHRF: Character N-gram F-score for Automatic MT Evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 2015. https://doi.org/10.18653/v1/w15-3049

[19] T. Kocmi, C. Federmann, R. Grundkiewicz, M. Junczys-Dowmunt, H. Matsushita, and A. Menezes, "To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation," Computer Software, 2021.

Abstract

This work investigates methods to enhance machine translation for low-resourced South Slavic dialects, focusing on Chakavian, Cerkno, and Torlak dialects. Using the No Language Left Behind (NLLB) model, this study explores three approaches that were based on the COPA dataset: baseline, fine-tuned, and back-translation model. Fine-tuned models surpassed baseline ones by 17.70 BLEU and 17.99 chrF2 points, while the back-translation model, which was based on the technique of back-translation for augmenting the original dataset, outperformed fine-tuned models by 6-16 BLEU and 11-28 chrF2 points. However, results on an additional parallel dataset showed limited generalization potential.

Keywords: natural language processing; machine translation; low-resource languages; dialect

Sažetak

U ovom radu istražuju se metode za poboljšanje strojnog prevođenja za južnoslavenske dijalekte i jezike s malim resursima, fokusirajući se na čakavski, cerljanski i torlački dijalekt. Koristeći model No Language Left Behind (NLLB) model, ovaj rad istražuje tri pristupa koji su se temeljili na skupu podataka COPA: osnovni, fino ugađani i model obrnutog prijevoda. Fino ugađani modeli nadmašili su osnovne za 17.70 BLEU i 17.99 chrF2 bodova, dok je model obrnutog prijevoda, koji se temeljio na tehnici obrnutog prijevoda za proširenje izvornog skupa podataka, nadmašio fino ugađane modele za 6-16 BLEU i 11-28 chrF2 bodova. Međutim, rezultati na dodatnom paralelnom skupu podataka pokazali su ograničeni potencijal generalizacije.

Ključne riječi: obrada prirodnog jezika; strojno prevođenje; jezici s malim resursima; dijalekt