UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Dmytro Pashchenko

Paragraph-Level Translation of Low-Resource Finno-Ugric Languages

Master's Thesis (30 ECTS)

Supervisors: Mark Fishel, Prof.
Elizaveta Yankovskaya, PhD

Tartu 2024

Paragraph-Level Translation of Low-Resource Finno-Ugric Languages

Abstract:
The emergence of massively multilingual neural machine translation models made it possible to efficiently translate many languages simultaneously, including those with extremely limited resources. The recent record holder, MADLAD-400, which spans over 400 languages, remains largely unexplored. In this work, we attempt to investigate the capabilities of MADLAD by fine-tuning it to translate four low-resource Finno-Ugric languages (Proper Karelian, Livvi, Ludian, and Veps, not included in MADLAD's collection) into Russian and back. Moreover, we explore the impact of paragraph-level translation on the model's performance, leveraging the document-level capabilities of MADLAD. We find that (1) the MADLAD-based system achieves results comparable to those of state-of-the-art models and discover that (2) the paragraph-level version of the system outperforms the sentence-level version by up to 3 BLEU points, significantly improving the consistency between sentences.

Keywords:
neural machine translation, paragraph-level translation, discourse-level phenomena, massively multilingual models, MADLAD-400

CERCS: P176, Artificial Intelligence

Vähese ressursiga soome-ugri keelte lõigutaseme tõlge

Lühikokkuvõte:
Massiliselt mitmekeelsete masintõlkemudelite teke võimaldas tõhusalt tõlkida paljusid keeli samaaegselt, sealhulgas neid, millel on piiratud hulk ressursse. Hiljutine rekordiomanik MADLAD-400, mis katab üle 400 keele, on suuresti uurimata. Käesolevas töös püüame uurida MADLADi võimekust, häälestades seda nelja väikese ressursiga soome-ugri keele (karjala, liivi, lüüdi ja vepsa, mis ei sisaldu MADLADis) tõlkimisele vene keelde ja tagasi. Lisaks uurime lõigutasandil tõlke mõju mudeli kvaliteedile, kasutades MADLADi dokumenditasemel tõlkimise võimekust. Leiame, et 1) MADLADi-põhine süsteem saavutab tipptasemel mudelitega võrreldavad tulemused ja avastame, et 2) süsteemi lõigutasandil versioon ületab lausetasemel versiooni kuni kolme BLEU punkti võrra, parandades oluliselt lausetevahelist kooskõla.

Võtmesõnad:
närvi masintõlge, lõigutaseme tõlge, diskursuse tasemel nähtused, massiliselt mitmekeelsed mudelid, MADLAD-400

CERCS: P176, Tehisintellekt

Acknowledgements

For invaluable support in writing this work, which goes beyond mere academic guidance and is nothing short of genuine, unpaid friendliness, I express my sincere gratitude to my supervisors, Mark Fishel and Elizaveta Yankovskaya. Thank you for believing in me and accepting me into the team. For assistance in working with Finno-Ugric languages and making expert translations into the Livvi language, I thank Ilia Moshnikov. I will surely run out of words before I finish describing the importance of friendships that have accompanied me throughout this academic year. To you, who drove away my fears and kept my soul alive, for Angst essen Seele auf. A million thanks to Nastia, Nina, Nikolai, and Lizi from Tartu. A million thanks to my friends in Ukraine, Nika, Vlad, Anya, and the entire community of cinephiles gathered around my beloved Cinemachat. I wish I could name you all, but there are too many of you already! The rest of my gratitude goes to my kindred souls, with whom I have had many a tacit conversation during my darkest days. Thank you for your existence, Giacomo, Arthur, and Emil. Last but not least, note to self:

What you are regarding as a gift is a problem for you to solve.

Contents

1 Introduction

Recent years have been marked by the advent of massively multilingual neural machine translation (NMT) systems capable of translating between many languages simultaneously. The most prominent examples include NLLB-200 (NLLB Team et al., 2022), which supports about 200 languages, and MADLAD-400 (Kudugunta et al., 2023), the current record holder translating over 400 languages. One of the fundamental benefits of using a single model to translate between multiple languages is the emergence of *transfer learning abilities* (Zoph et al., 2016). It turned out that a multilingual model can be trained to translate a new language on a very limited amount of data if it is first pre-trained on languages with a large number of resources (texts). This discovery was a breakthrough in NMT and caused a race to accumulate as many languages within one model as the model's memory allowed. The creators of NLLB set the goal to ensure that "no language was left behind," hence the model's name. NLLB and MADLAD bridged the gap between languages with large amounts of translated texts, such as English, and low-resource languages, such as Friulian or Urdu. The importance of this initiative cannot be overstated, as the availability of translation tools for endangered languages could prevent their extinction, reinvigorating their use.

The current state of massively multilingual NMT leaves significant gaps to be filled. For instance, most languages of the Finno-Ugric family, despite having tens of thousands of native speakers at the time of writing this work, are not supported by existing systems, including NLLB and MADLAD. Research was carried out on extending the capacity of NLLB to new Finno-Ugric languages and yielded positive results (Purason et al., 2024). In contrast, the capabilities of the recently introduced MADLAD model remain largely unexplored. We embark on the task of exploring the effectiveness of this model for translating low-resource Finno-Ugric languages using the example of Karelian and Veps. The Karelian language has several dialects: Proper Karelian, Livvi-Karelian (also called Livvi), and Ludian, which we shall treat as separate languages (but sometimes for brevity we shall call them all together Karelian). We chose Russian as the target language because of the availability of professionally translated data. Thus, we fine-tune the MADLAD model to translate Veps and Karelian into Russian.

Moreover, we investigate the features of MADLAD that could enhance the quality of translation. In particular, we were interested in the fact that MADLAD was partially trained on a large amount of document-level data, that is, on full texts rather than individual sentences. Most NMT systems are trained to translate only sentences for practical reasons. The thing is that the processing time of Transformer-based models (Vaswani et al., 2017), such as MADLAD and NLLB, scales quadratically with the input length. By translating texts sentence-by-sentence, the models no longer depend on the size of the input texts and significantly save time and memory consumption. In most cases, sentence-level translation is enough to achieve good translation scores, despite sacrificing the overall, *discourse-level* context. Yet, in any coherent text, sentences largely depend

on each other, for example, by means of pronominal anaphora (e.g., pronouns and their referents) or word repetitions (Bawden et al., 2018). Läubli et al. (2018), having asked people to evaluate human- and machine-made translation of documents, discovered that respondents overwhelmingly preferred humans to machines. This indicates the importance of considering an *extrasentential* context in machine translation.

In order to leverage MADLAD's document-level abilities without causing massive processing times, we split the input documents into paragraphs of 5 sentences. Five sentences seem enough to capture some topical discourse-level phenomena arising when translating from Karelian/Veps to Russian, in particular gender-neutral pronouns. In contrast, pronouns in Russian are gendered; therefore, translating them from Karelian or Veps requires additional context (information about the referent), sometimes spanning several preceding sentences. We explore the properties of MADLAD and its ability to deal with the discourse-level issues in a series of experiments.

We fine-tune two MADLAD-3B models (with 3 billion parameters): one for translating sentences and the other for translating paragraphs. For this, we manually collect document-level data and split it into paragraphs of 5 sentences. The same data, but divided into sentences, is used to fine-tune the sentence-level model. We evaluate the performance on two test sets. The first one comprises the first 250 rows of the FLORES benchmark, translated into Finno-Ugric languages. Although this is a sentence-level set, it is composed initially of excerpts from Wikipedia articles. We create the second *paragraph-level* test set by merging the sentences back into paragraphs. We measure the performance of the models on both datasets with metrics like BLEU and chrF++. To translate paragraphs with the sentence-level model, we translate them sequentially, sentence by sentence.

Then, in order to improve the quality of translation in a high-to-low-resource direction, we perform back-translation of monolingual corpora by our models. We pre-process the monolingual data similarly, creating its sentence-level and paragraph-level versions. Additionally, we test the zero-shot capabilities of the models on the task of translation into English and back. Finally, we compare the performance of our paragraph-level model with Neurotõlge, a sentence-level state-of-the-art model for translating Finno-Ugric languages.

We pose three main research questions:

1. Can MADLAD-400 be effectively extended to translate more low-resource languages?

2. Is it possible to fine-tune MADLAD-400 for high-quality translation of paragraph-level data?

3. Is paragraph-level translation better than sentence-level translation?

The work is structured as follows. Chapter 2 provides a general background to our research. In Chapter 3, we explain in precise detail our choices pertaining to splitting

data into paragraphs and evaluating sentence and paragraph translations, while also briefly reviewing the methods used in our work, such as back-translation and zero-shot translation. Chapter 4 details our data sources and how we processed the data to create training and test sets. In Chapter 5, we discuss our experiments and their technical setup. Finally, Chapter 6 presents the results of the experiments. In Chapter 6.1, we discover the benefits of the paragraph-level model and provide specific translation examples that demonstrate the model's ability to handle discourse-level phenomena. Then, in Chapters 6.2-6.3, we successfully perform the back-translation and discover the excellent zero-shot capabilities of MADLAD. Finally, in Chapter 6.4, we conduct the evaluation of the current state-of-the-art translator Neurotõlge on paragraphs and find that the performance of our paragraph-level model is more or less comparable with Neurotõlge depending on the translation direction.

2 Background

2.1 Neural Machine Translation

2.1.1 Introduction to Machine Translation

Machine Translation (MT) is one of the major branches of Natural Language Processing (NLP), which studies ways of translating natural languages using computers. So far, there have been three principal approaches to performing machine translation: rule-based machine translation (RBMT), statistical machine translation (SMT), and neural machine translation (NMT) — with the latter one currently dominating the field. The first machine translation systems, dating back to the 1950s, were rule-based: researchers had to develop an elaborate set of translation rules covering morphology, syntax, and semantics of languages of interest (Hutchins, 1995). Yet, given the intricacy inherent in natural languages, it has proven unrealistic to come up with an accurate and exhaustive set of translation rules accounting for all irregularities.

Since the end of the 1980s, RBMT has been supplanted by statistical (or corpus-based) methods that spared researchers the hard labor of rule creation (Brown et al., 1990). Statistical machine translation systems 'learned' to translate from a large set of human-translated examples (so-called parallel corpora). In particular, they learned to align words between source and target languages and place them in the correct order based on statistics retrieved from corpora. Although SMT vastly improved translation quality, the fluency of generated translations suffered.

The next breakthrough came with the rise of deep learning in the 2010s when a novel Neural Machine Translation (NMT) paradigm stepped into the spotlight (Tan et al., 2020). NMT has brought with it a number of improvements and simplifications and quickly replaced statistical and other approaches. First and foremost, it offered a single architecture, the neural net, to model the entire translation process. Second, while statistical models represented words as integer numbers (so-called tokens), neural models further mapped tokens to learnable real-value vectors (so-called embeddings). Thus, neural networks learned good data representations by themselves, eliminating the need for excessive feature engineering. Despite their simplicity, the NMT models have achieved state-of-the-art performance.

2.1.2 Encoder-Decoder Framework

Most neural machine translation models build on the encoder-decoder framework (Cho et al., 2014). This framework consists of four components: the embedding layers, the encoder and decoder networks, and the classification layer. Let us briefly describe each component.

The embedding layers map discrete symbols to continuous vectors representing learnable features. The encoder is a neural network that maps source embeddings

to continuous hidden representations, which accumulate contextual information. The decoder network resembles the encoder but behaves as a language model. It gradually generates target embeddings, conditioned on previously generated tokens and the encoder output. The classification layer predicts the distribution of target tokens based on the target embeddings.

During the inference stage, it is often the case that the most probable next token is not the most suitable one. Maximizing the token probability does not necessarily entail maximizing the probability of the final translation. However, keeping track of all possible translations is impractical due to the large hypothesis space. One way to deal with this problem is to keep track of a small number of best successors at each inference step. This strategy, called beam search (Tan et al., 2020), has been widely adopted in NMT.

2.1.3 Attention Mechanism and Transformer Model

The "Attention Is All You Need" paper, published in 2017, became a huge milestone in the field of NLP and beyond. Vaswani et al. (2017) introduced a new self-attention mechanism and a model, Transformer, that successfully employed it. Variations of the original Transformer model have become the basis of such state-of-the-art models as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2023), LLaMA (Touvron et al., 2023), etc., which are commonly known as large language models (LLM). Transformer follows the same encoder-decoder framework we discussed above. The difference is that both the encoder and decoder now consist of multiple layers of attention heads that apply the self-attention mechanism, which we shall discuss now.

The original attention mechanism helped to solve the problem of long-range dependencies between words from the source and target sentences (Bahdanau et al., 2014). As pre-Transformer encoders processed sentences sequentially, one token at a time, decoders tended to "forget" the first processed tokens. The attention mechanism had become a tool for directly providing encoders with necessary information about each word in a source sentence. At each generation step, each token in the input sequence was assigned a value vector and the attention network calculated the relevance of each value vector based on queries and keys. In the authors' words, decoders "emulated searching through a source sentence during decoding a translation."

Given a set of m query vectors $\mathbf{Q} \in \mathbb{R}^{m \times d_k}$, a set of n key vectors $\mathbf{K} \in \mathbb{R}^{n \times d_k}$, and associated value vectors $\mathbf{V} \in \mathbb{R}^{n \times d_v}$, attention can be calculated by the formula:

$$\text{Attention}(\mathbf{QKV}) = \text{softmax}(\frac{\mathbf{QK}^{\text{T}}}{\sqrt{d_k}})\mathbf{V}. \tag{1}$$

This is (Vaswani et al., 2017)'s version of attention, which they used in their breakthrough Transfomer model. They discovered that the attention mechanism could also be used in encoders which led to state-of-the-art performance. They called it self-attention,

because now the queries, keys, and values all came from the same sequence. The proposed Transformer model almost entirely relied on self-attention blocks, hence the article title.

Vaswani et al. (2017) also introduced a concept of multi-head attention. Both the encoder and decoder of Transformer were composed of six layers, each layer had eight parallel attention heads, and each head attended to a particular subspace of input vectors. This allowed the model to learn diverse data representations. Without going into further detail, we provide you with a scheme of the Transformer network (Figure 1), which is the basis of the MADLAD-400 model we shall use throughout our study.

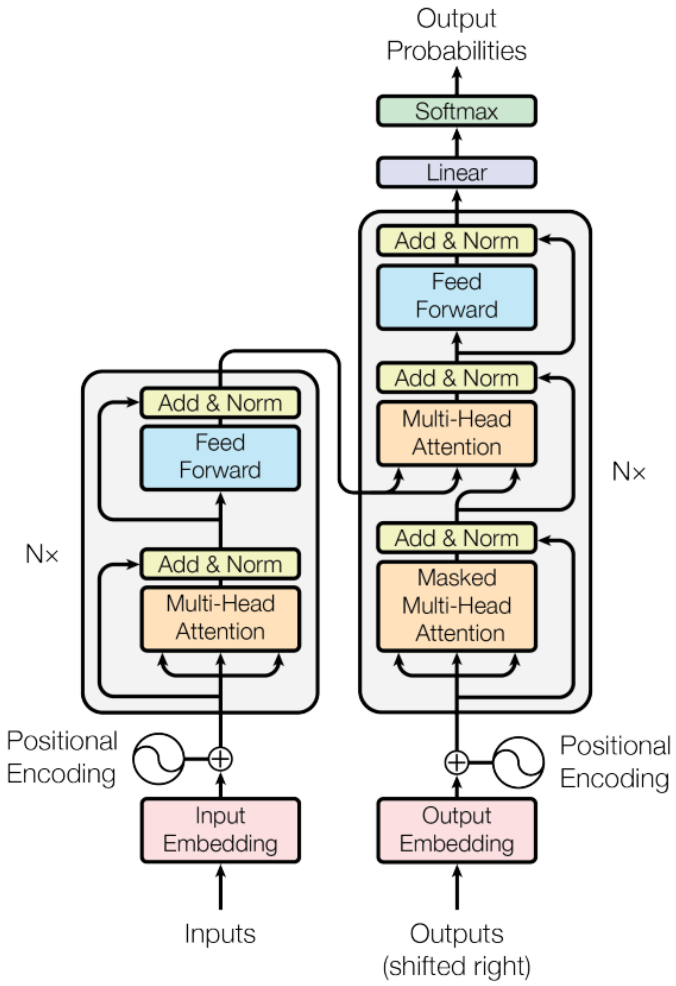

Figure 1. The Transformer-model architecture. (Vaswani et al., 2017)

2.2 Low-Resource NMT and Multilinguality

2.2.1 Transfer Learning for Low-Resource NMT

The encoder-decoder framework, while effective for translating between pairs of languages with large amounts of parallel data, faced significant challenges in low-resource settings. In these cases, such as translating from English to Urdu, NMT models were found inferior even to some statistical models. To address this, Zoph et al. (2016) proposed a novel method of transferring knowledge from high-resource to low-resource translation models called *transfer learning*. This method involved training an NMT model to translate a high-resource language pair and then moving its parameters to an identical model, which was then trained to translate a low-resource pair. The final system largely improved baseline results, which proved that internal representations learned when translating one language pair could be reused to learn to translate another language pair with much less training data.

2.2.2 Massively Multilingual Models

Expanding on the concept of transfer learning, Johnson et al. (2017) introduced a significant improvement. Instead of training separate models for each language pair, they used a single NMT model to translate multiple languages simultaneously. This was achieved by adding a target language token to input sequences. For instance, "`<2en> Je suis là.`" indicated that the French sentence "`Je suis là.`" was to be translated into English. This approach yielded three key benefits:

- **Scalability:** A single multilingual model could replace n^2 monolingual models needed to translate between n languages without losing quality.
- **Low-Resource NMT:** In a multilingual model, all parameters were implicitly shared between all pairs of languages. This enabled the model to generalize across languages. The authors observed that mixing data for high- and low-resource language pairs significantly improved the quality of low-resource translation.
- **Zero-Shot Translation:** Surprisingly, multilingual models showcased the ability to translate between pairs of languages they had never seen before. This phenomenon was called *zero-shot translation*. For example, a multilingual NMT model trained in Portuguese $\rightarrow$ English and English $\rightarrow$ Spanish directions could achieve a reasonable translation quality for Portuguese $\rightarrow$ Spanish.

Johnson et al. (2017) trained their multilingual model in less than five translation directions. Later research focused on scaling the multilingual models to as many languages as possible without losing the overall translation quality. Aharoni et al. (2019) trained a *massively multilingual* Transformer model that translated up to 59 low-resource

languages in all translation directions, outperforming the previous front-runners and bilingual models of similar capacity.

The current state-of-the-art massively multilingual model in terms of translation quality is NLLB-200 (NLLB Team et al., 2022). Researchers from Meta managed to break the 200-language barrier with unmatched results. NLLB is a Transformer-based model with number of parameters ranging from 600M (distilled version) to 54.5B (mixture of experts). This model was also accompanied by a novel FLORES-200 many-to-many translation benchmark covering the same 200 languages.

2.2.3 MADLAD-400

As of the time of writing this thesis, the record for the number of translation languages was set by the MADLAD-400 model by Google, which can translate 419 languages. Kudugunta et al. (2023) created a large document-level monolingual dataset with texts collected from the web and used freely available parallel corpora to train three encoder-decoder translation models along with one decoder-only language model. The translation models are Transformer-based and differ only in size; in particular, these are the 32-layer 3B-parameter model, 48-layer 7.2B-parameter model, and 32-layer 10.7B-parameter model (number of layers indicates the size of the encoder and decoder). We shall refer to these models as MADLADs, though the authors reserved this name for the dataset. Tested against the NLLB-54B model on such datasets as WMT and FLORES-200, MADLADs have shown comparable, though generally worse, results despite the much smaller number of parameters.

Now, let us overview the training procedure. MADLAD is trained using two objectives, machine translation and the MASS-style objective (Song et al., 2019), which are sampled with a 50% probability. The first objective consists in teaching the model to translate sentences based on the openly available parallel corpora. Each sentence was prepanded a <2xx> token to indicate the target language, where xx standed for language code. The context window was limited to 256 tokens. However, the parallel corpora encompassed only 157 languages with 4.1B sentence pairs. To make the system understand even more languages, the second objective was utilized. It used the brand-new monolingual corpora covering 419 languages to teach the model to recover masked tokens in documents. Under documents and document-level data, the authors mean complete, coherent texts, not segmented into sentences or paragraphs. Additionally, the authors obtained multiway (multilingual) data by automatically matching sentences from the monolingual corpora using the n-gram method (Freitag and Firat, 2020). This gave 11.9B more parallel examples across a total of 20 742 language pairs.

Thus, far from being a purely document-level model, MADLAD was primarily trained to translate sentences. Nevertheless, it still retains the ability to process documents. For example, the authors also did experiments with back-translating the document-level monolingual corpora and found it beneficial for the **en2xx** translation direction.

2.3 Discourse-Level Phenomena in NMT

Most machine translation models assume that texts can be translated sentence by sentence for practical reasons. The first SMT models, back in the 1990s, worked under strong assumptions of independence between sentences and even words in those sentences, intending to narrow down a large hypothesis space (Hardmeier, 2012). More precisely, they took into account at most two preceding words for each word in a target sentence — too little to capture the word's meaning. In contrast, recent attention-based NMT systems consider entire sentences; they encode the meaning of each source word as a weighted sum of the "meanings" of all other words in a sentence. These context-aware embeddings are then decoded into sentences in a target language, preserving the meaning of a source.

This approach has shown excellent results, making researchers reluctant to look beyond isolated sentences. Some even hurried to announce that NMT systems achieved parity with professional human translators. Läubli et al. (2018) disproved this claim by showing that people decisively preferred professional translations over machine ones when asked to evaluate entire texts (documents) instead of isolated sentences. This observation indicates that MT systems still omit the essential part of words meanings by translating sentences in isolation. The inclusion of extrasentential context is crucial as long-range dependencies between words stretch beyond the scope of a single sentence.

Currently, attempts have been made to incorporate context in the attention-based models' scope by changing their architecture. Researchers offered such methods as hierarchical attention (Miculicich et al., 2018) or memory networks (Maruf and Haffari, 2018). However, the most straightforward strategies, like passing an entire text to the model, proved the most effective in capturing context. Sun et al. (2022) trained the Transformer model on full texts, albeit splitting them into parts to vary input lengths. This approach demonstrated a big leap in translation quality.

Bawden et al. (2018) identify three major linguistic phenomena that are problematic for MT and occur at the level of discourse (that is, on the scale of entire texts): coreference, lexical cohesion, and lexical disambiguation. Below, we review all three phenomena in the case of Estonian-English translation. We give examples of alternative translations, highlighting the right one in bold, and underline the disambiguating context.

Coreference resolution consists of gendering anaphoric pronouns, which replace previously introduced noun phrases to avoid repetition. Estonian pronouns are gender neutral, meaning that the same word (*tema/ta*) may stand for both males and females, and extra context is required to disambiguate the translation. In the following example (1), the ambiguous pronoun *ta* cannot be accurately translated without the antecedent; moreover, it will most likely be translated as *she* due to the gender statistics of kindergarten workers in the real world (which models' training corpora usually reflects).

(1) Mark tormab tööle. Ta töötab lasteaias õpetajana.

Mark rushes to work. **He**/she works as a kindergarten teacher.

Lexical cohesion means respecting alignment or repetition. In the example below (2), the words *silly* and *stupid*, which convey the same Estonian word *rumal*, are not interchangeable in the discourse context.

(2) Ta ütles, et mu küsimus on rumal. Mis selles nii rumalat on?

He said that my question was silly. What is so **silly**/stupid about it?

Lexical disambiguation refers to cases in which the correct translation of a term depends on the general context of discourse. The example below (3) contains an Estonian word *tee* that can mean *tea* and *road*, but in the context of the previous sentence, it becomes clear that one is talking about a road.

(3) Linn on teisel pool, miks me pöörasime? Ma eelistan seda teed.

The city is on the other side, why did we turn? I prefer this **way**/tea.

We shall use these discourse-level phenomena to test the abilities of our models later in our study.

2.4 Low-Resource Finno-Ugric Languages

In their comprehensive endeavor to create a system capable of translating the least spoken, researched, and digitalized languages, the Meta researchers managed to bring together more than 200 languages, many of them being low-resource (NLLB Team et al., 2022). They called a language low-resource if there were less than 1 million sentences of publicly available translated data in this language at the moment of study. Based on the same definition, we call the languages we explore in our research, namely Karelian (and its dialects Proper Karelian, Livvi, and Ludian) and Veps, the low-resource ones. The main source of material in these languages, the VepKar[1] corpus (Boyko et al., 2022), contains nearly tens of thousands of sentences in each of them at the time of writing this work.

Karelian and Veps belong to the Finno-Ugric family with three major languages: Estonian, Finnish, and Hungarian. They are closer to Estonian and Finnish and are part of the so-called Balto-Finnic branch. The native speakers are primarily settled near the border between Russia and Finland. All of the languages have an established written form; books are translated into them, and journalistic articles are written. However, at the moment, there is only one system capable of translating these languages - Smugri (Purason et al., 2024), developed by the Tartu NLP research group.

What is essential for our study is that, in addition being low resource, the Karelian and Veps languages are characterized by remarkable linguistic phenomena (e.g., gender-neutral pronouns), making the translation of these languages a challenging task.

[1] http://dictorpus.krc.karelia.ru/en

3 Methodology

In this chapter, we describe the methods that underlie our experiments. We first discuss why we prefer using paragraphs instead of documents, justify our choice of paragraph length, and define our procedure for obtaining sentence- and paragraph-level data from available document-level sources (Chapter 3.1). We then explain why we prefer the sentence-level BLEU and chrF+ metrics for both sentence- and paragraph-level translation evaluation and describe how we apply them to paragraphs (Chapter 3.2). We conclude the chapter by explaining back-translation (Chapter 3.3) and zero-shot translation (Chapter 3.4) and present ways of accelerating model training by using multiple GPUs and gradient accumulation (Chapter 3.5).

3.1 Splitting Documents into Paragraphs

First, let us clarify the terms, that is, what we mean by documents and paragraphs. Deutsch et al. (2023), who, like us, investigate the creation of paragraph-level data, indicate the lack of agreement upon definitions of the terms in MT literature and propose their formulations, which we shall adopt in our study. Under the segment, we understand the input of any length to an NMT model. Then, the paragraph is a multi-sentence segment. Notice that we do not require it to be a self-sufficient unit of text in a traditional sense, referring to it instead as an arbitrary sequence of consecutive sentences. The document is an entire input text, coherent and self-sufficient, which can be broken down into sentences and paragraphs.

Now, we should explain why we prefer paragraphs to documents and so paragraph-level to document-level translation. With our primary task being to test whether including the extrasentential context improves the performance of MADLAD, we need to decide on how many sentences to use as the model's input. On the one hand, the more sentences we take from a document, the more likely the model is to capture the necessary context for translating each sentence. Indeed, some long-range dependencies between words span the entire document, e.g., the need to preserve the overall style or specific terminology. On the other hand, passing the document as a whole as the model's input may be impractical. Despite the fact that the MADLAD model does not have a limited context window and, therefore, can process inputs of any length, we find document-level translation problematic, and here is why:

- **Translation quality.** Koehn and Knowles (2017), outlining six main challenges for NMT, indicate that NMT systems perform poorly on long sequences. Even though the attention mechanism was designed to address this issue, experiments show that Transformers may still be outperformed by SMT systems in extreme cases. Varis and Bojar (2021) speculate that the problem boils down not to lengths

of sequences per se but to the fact that Transformers generalize badly to out-of-distribution input lengths. Thus, without restrictions on the input size, the model will still be subject to the generalization error.

- **Time and memory consumption.** The attention mechanism inside Transformers has quadratic computational complexity $\mathbf{O}(n^2)$ since the attention is calculated between each pair of tokens. Therefore, computation time and memory consumption increase quadratically with the input size. Shorter input sequences would ensure much faster model training.

Sun et al. (2022) tried to overcome the first issue by repeatedly dividing a document into k parts, $k \in \{1, 2, 4, 8, ...\}$, and giving the document to a model in parts, thus varying the input lengths. However, despite being superior in quality to the complex methods, requiring changes to the Transformer architecture, this approach does not amend large processing times. We overcome the two aforementioned issues at once by splitting documents into smaller paragraphs of fixed, reasonable length. Since MADLAD was trained on sequences whose length did not exceed 256 tokens, we set a similar length limit. We abandoned the idea of forming paragraphs from as many sentences as possible

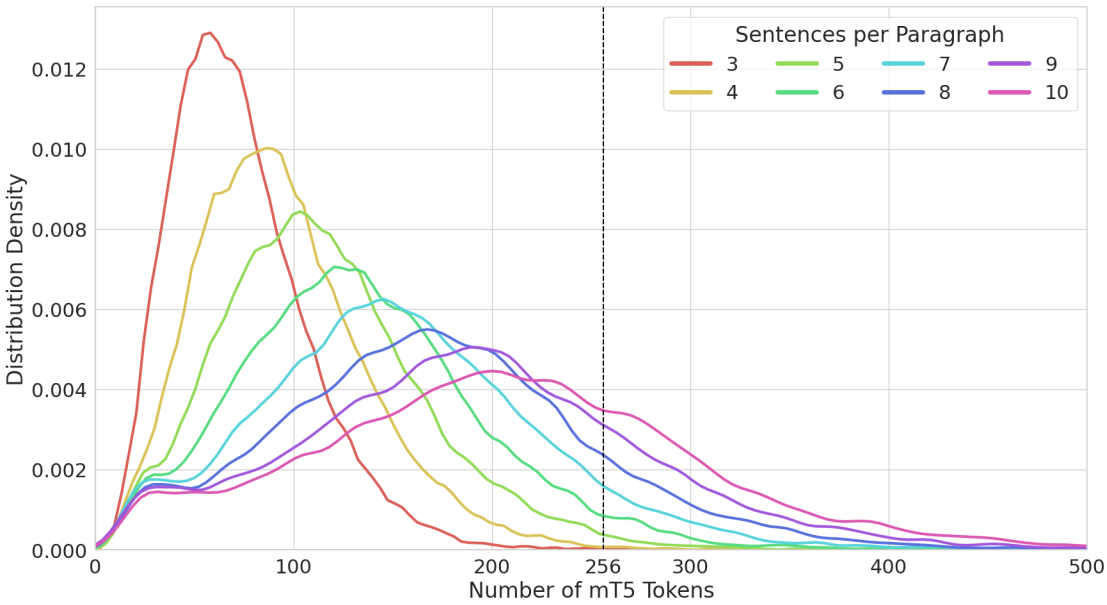

Figure 2. The distribution density of paragraph sizes as measured in mT5 (MADLAD) tokens depending on the number of included sentences. The source of sentences is VepKar parallel corpora. The dashed line indicates the maximum sequence length, exceeding which would lead to trimming the sequence.

to get close to the size limit, for this would have led to a low variance of data lengths. Instead, we combine a fixed number of sentences by the following procedure:

1. Split the document into sentences.
2. Merge the sentences into paragraphs, five sentences each.
3. If there are less than five sentences left, merge them into a paragraph as well.
4. If the paragraph length exceeds 256 tokens, trim the paragraph.

To determine the most suitable paragraph size, we compared distributions of paragraph lengths in parallel corpora as measured in tokens for paragraphs consisting of 3 through 10 sentences (Figure 2). The vertical line on the plot indicates the cut-off length. Our goal was to keep a sufficient number of sentences to capture context while also avoiding trimming, which corrupts examples. We eventually chose a paragraph size of five sentences. Its distribution spans a wide variety of sequence lengths with a mean length of 115. The number of trimmed examples remains negligibly small (221 paragraphs out of 17 890).

The monolingual corpora we used for back-translation have a different distribution density, which may result in unwanted trimmings. To address this, we modified the last step of the splitting procedure:

4*. If the paragraph length exceeds 256 tokens, split the paragraph in two.

5*. If the paragraph is still too long but consists of one sentence and cannot be split anymore, trim the paragraph.

For a detailed description of the final paragraphed dataset, consult the section dedicated to data.

3.2 Evaluating Sentence- and Paragraph-Level Translations

Evaluating translations is a tricky task, given that the same meaning can be conveyed in entirely different words, so intuitively, the good evaluator should operate on the semantics level. Surprisingly, syntax-level metrics, such as BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017), demonstrated good alignment with human judgments and became standard in NMT research. Yet, these metrics were designed to evaluate sentences, not paragraphs. The other alternatives could be, for example, training regression models to predict human scores (see COMET (Rei et al., 2020)) or prompting instruction-tuned LLMs to perform the evaluation (Kocmi and Federmann, 2023). Unfortunately, these options are also problematic in our resource-constrained setting: training models require gathering large amounts of human-evaluated data; at the same time, current LLMs (e.g., ChatGPT[2] and Llama (Touvron et al., 2023)) do not support Karelian and Veps. Finally,

[2] https://chatgpt.com/

in light of recent research, we preferred to use BLEU and chrF++ as paragraph-level translation metrics. We explain our reasons below.

BLEU is a standard metric for automatically evaluating machine translations, which demonstrates a high correlation with human judgments. It measures the quality of translation based on the number of coinciding sequences of n-tokens in the machine- and human-translated sentences. However, BLEU was designed as a sentence-level metric, and applying it to paragraphs could compromise correlation with human judgments due to error accumulation. Paradoxically enough, Deutsch et al. (2023) have proved the opposite: BLEU scores for paragraphs not only align with those of humans but also become more accurate as paragraph size increases. The authors explain this phenomenon as an effect of "averaging away" the noise in the metric scores. Those findings allowed us to adopt BLEU as a paragraph-level metric without the need to train custom scoring models. Notably, the authors also found that there is no difference between applying BLEU to an entire paragraph and averaging BLEU scores for sentences within a paragraph: both methods produce scores that closely match human judgments. Thus, we calculate the BLEU scores of paragraphs directly.

chrF (Popović, 2015) is an advanced but less popular version of BLEU. Unlike BLEU, it calculates the number of coinciding n-grams (sequences of n-characters) instead of tokens. chrF++, yet another improvement over BLEU, is, in fact, a middle ground between BLEU and chrF that takes into account both tokens and characters n-grams. Using character-level metrics such as chrF++ is essential in our study, since Karelian, Veps, and Russian are morphologically rich languages for which capturing matching word parts can be more telling than capturing only matching whole words. Even though Deutsch et al. (2023) did not test chrF/chrF++ in application to paragraphs, we still believe that the scores of these metrics exhibit the same behavior as the scores of their close analog BLEU.

3.3 Back-Translation

Unfortunately, a large part of the data one can obtain for a particular language comes without translation. This is so-called monolingual data, and there are several ways in which translation systems can benefit from it. One of them is back-translation (Sennrich et al., 2016). When a system becomes proficient enough, it can be used to translate monolingual corpora. The resultant synthetic data can then augment the original parallel corpora but in a flipped manner: generated examples become part of the source data, while high-quality source segments move to the target side. With this, one can significantly boost translation quality in the opposite direction.

3.4 Zero-Shot Translation

The zero-shot translation (Johnson et al., 2017) is a surprising benefit of incorporating multiple languages in a single model. This is a special case of transfer learning when a system trained in one set of translation directions gets the ability to translate in directions it has never seen. In this study, we want to test whether the MADLAD model, with its knowledge of the English → Russian translation, can generalize to the English → Karelian/Veps translation when introduced to Russian → Karelian/Veps examples.

3.5 Scaling Neural Machine Translation

Training state-of-the-art large language models with billions of parameters can require several days, weeks, or even months. Ott et al. (2018) propose using reduced precision and large batches to speed up training on a multi-GPU machine. They observe the 5x speedup of training on 8 GPUs. Let us briefly summarize their methods of scaling, which we also used to speed up our models.

Half-Precision Training. Modern GPUs allow computations with half-precision floating point (FP), which are several times faster than full-precision operations. Although the authors used the FP16 format, in our study, we adopted a slightly different BF16 (Kalamkar et al., 2019) precision, representing a much more extensive numerical range. The primary rationale behind this decision is that Google employed the BF16 format to train MADLAD-400, and the model relies on it. Switching to FP16 precision leads to errors in calculations.

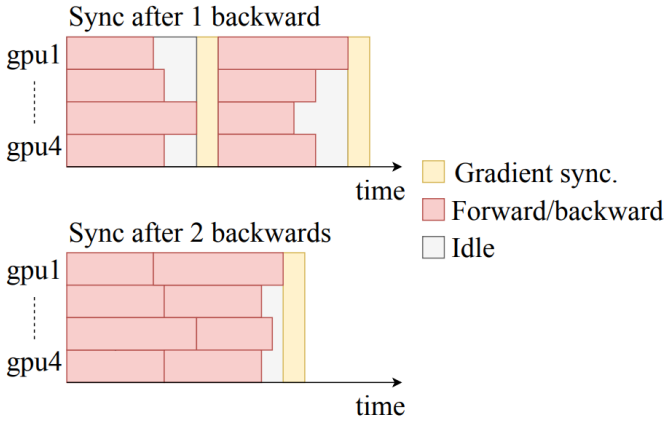

Figure 3. The illustration of gradient accumulation; training is being sped up by: (i) reducing communication between workers, and (ii) saving idle time by reducing variance in workload between GPUs. (Ott et al., 2018)

Training with Larger Batches. The authors note that increasing the batch size leads to faster model convergence in terms of the number of weight updates (i.e., optimization steps). One can distribute the training so that each worker independently calculates the gradients for its own mini-batch and then synchronizes with other workers during the weight update phase when all the calculated gradients are averaged. This approach is equivalent to training on a single large batch, the size of which is equal to the sum of the mini-batch sizes. Thus, larger batches not only enable faster convergence but also allow for parallelization.

Parallel Training. As described above, parallel training is possible by distributing a large batch across multiple GPUs. However, the time required to communicate the computed gradients increases with the number of GPU nodes involved since the computation time varies for each node: the first finishing node remains idle, waiting for other nodes. The authors mitigate this problem by allowing each node to iterate over multiple mini-batches and accumulate the gradients. This technique, called gradient accumulation (Figure 3), helps to reduce communication and workload variance. Additionally, the authors conduct gradient synchronization in parallel with back-propagation, further improving training speed.

4 Data

A crucial but challenging task when working with low-resource languages is finding sources of high-quality translated texts written in those languages. Moreover, as we are interested in a document-level translation, it is essential that the texts consist of multiple sentences and be coherent. We utilize data from four sources that satisfy these conditions: VepKar[3], Wikipedia, Omamedia[4], and FLORES-200 (NLLB Team et al., 2022). On a separate note, we should mention why the existing corpora, particularly the dataset used for training Smugri (Purason et al., 2024), the state-of-the-art model for translating Finno-Ugric languages, does not suit our purposes. Despite originating from documents, the Smugri dataset consists of sentences in random order, which renders it impossible to recover the original pieces of text. We kindly borrow the aforementioned data sources from Smugri but **rebuild the dataset from scratch**.

Another challenge is splitting the documents into chunks of text that fit the context window - previously, we agreed on the heuristic that such a chuck must include up to five sentences. Some sources provide texts already broken down into sentences, while others require segmentation. Provided with isolated sentences, we merge them again into paragraphs of size 5. Where the number of sentences is not divisible by 5, we take the remainder as a separate paragraph.

In this chapter, we review the corpora in descending order of their share in the final dataset and define our approaches to preprocessing for each corpus. Table 1 presents the composition of the final dataset. In this table, as in all subsequent ones, we use language codes: *krl* stands for Proper Karelian, *lud* for Ludian, *olo* for Livvi, and *vep* for Veps.

data source	krl		lud		olo		vep	
	mono	para	mono	para	mono	para	mono	para
vepkar-sent	45.4	32.3	5.9	7.9	36.0	22.2	38.3	20.4
vepkar-par	9.6	6.9	1.2	1.6	7.6	4.8	8.1	4.5
wikipedia-sent	-	-	-	-	28.4	-	99.8	-
wikipedia-par	-	-	-	-	7.7	-	24.1	-
omamedia-sent	-	-	-	-	11.4	-	6.3	-
omamedia-par	-	-	-	-	2.8	-	1.5	-
total-sent	45.4	32.3	5.9	7.9	75.8	22.2	144.4	20.4
total-par	9.6	6.9	1.2	1.6	18.1	4.8	33.7	4.5

Table 1. The distribution of sentence-level (sent) and paragraph-level (par) parallel data (para) and monolingual data (mono) by language in the final dataset. Quantities are given in thousands, rounded to the nearest tenth.

[3] http://dictorpus.krc.karelia.ru/en

[4] https://omamedia.ru/en/

4.1 VepKar

Corpus Description. The open corpus of Veps and Karelian languages, VepKar (Boyko et al., 2022), comprises texts written in Karelian Proper, Livvi, Ludian, and Veps as well as their translations into Russian. Not only does it meet the above requirements, it also provides us with a smorgasbord of data sources. The corpora encompass samples of texts written in an archaic literary style (e.g., Bible, folklore) as well as modern ones using ordinary language (e.g., journalistic articles, subtitles). Near half of them are supplemented with professional translations into Russian. VepKar is maintained by researchers of the Karelian Research Centre.

Preprocessing Steps. VepKar stores its data on a website where source texts and translations are marked up in such a way that each sentence (or several) of the source corresponds to a sentence (or several) of the translation. It allowed us to avoid segmentation and parse texts as separate sentences. We discarded the samples where the markup was corrupt, i.e., where the number of source sentences did not correspond to the number of target sentences. This way, we retrieved a total of 5913 texts/translations, which amounted to 82 739 sentences or 20 590 paragraphs up to 5 sentences in length. The ratio of total corpora sizes for Karelian Proper, Livvi, Ludian, and Veps is close to 7:6:1:6.

4.2 Wikipedia

Corpus Description. Among the languages being examined, only two are present on Wikipedia: Livvi-Karelian and Veps. As of 01/03/2024, there are 4697 articles in Livvi and 6970 in Veps. Even though there are no guarantees on data quality, as Wikipedia is maintained by volunteers, we judged the articles to be mainly well-written upon consultation with a professional linguist.

Preprocessing Steps. We used the Wikipedia HuggingFace dataset[5], which was collected on the fly from a fresh Wikipedia dump (the one dated 01/03/2024 at the time of experiments). The script performs some useful preliminary data cleaning, such as stripping markdown and removing unwanted sections (e.g., references). Nevertheless, we had to do additional work to complete the cleaning, with particular attention to respecting the integrity of the documents. First, we introduced some article-level changes, specifically:

- excluded articles dedicated to bibliographies and literary lists;
- deleted standard[6] appendices from every article, in particular, sections "See also," "References," "Bibliography," "Further reading," "External links";

[5] `https://huggingface.co/datasets/wikipedia`

[6] `https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Layout`

- removed section titles (mostly incomplete sentences that could be disposed of without compromising the integrity of an article);
- eliminated defects of previous data processing: removed HTML legacy characters, fixed parsing artifacts, and more.

Next, we had to split the articles into sentences. To this end, we used nltk's[7] utility `sent_tokenize` for Estonian. Despite being trained for another language, `sent_tokenize` managed to split texts in Livvi and Veps reasonably well due to their proximity to Estonian. However, the resulting segmentation contained some common errors: for example, when the tokenizer mistakenly considered an abbreviation to be the end of a sentence and split the latter in two. To solve this problem, we collected a list of abbreviations for Livvi and Veps and prevented splitting where they occurred. To deal with other types of mis-segmentation, we appended sentences starting with a lowercase letter to the previous ones.

Finally, we filtered sentences according to the following criteria:

- Sentences must include letters from the respective alphabets of Livvi or Veps (the criterion is motivated by the fact that articles sometimes included large amounts of text in Russian and (much less) other languages).
- The number of Cyrillic characters must not exceed the number of Latin ones (the sentences filtered out, upon investigation, often happened to be book/music/film titles in Russian).
- Sentences must be composed of at least three words (with minor exceptions, opposite cases were the artifacts of imperfect sentence splitting).

We did not conduct deduplication of sentences as this could disrupt the narrative flow of articles. Upon investigation, the most repeated sentences turned out to be general (e.g., "The climate is Russian continental.").

4.3 Omamedia

Corpus Description. Omamedia is an online resource that distributes news in the Livvi-Karelian, Finnish, and Veps languages among others. It aggregates articles from seven publishers, covering various topics such as sports, politics, and economics. These publishers are part of the "Periodika Publishing House," an autonomous institution of the Republic of Karelia.

Preprocessing Steps. The articles required segmentation into sentences. For this, we again used nltk's `sent_tokenize` for Estonian. It worked out fine, and further corrections were needed. In total, we got 1572 articles in Livvi and Veps, divided into 17 685 sentences and 4337 paragraphs, which became part of our monolingual data.

[7] `https://www.nltk.org/`

4.4 Smugri FLORES

Corpus Description. In addition to training data, a strong evaluation benchmark is needed to assess translation quality. The current state-of-the-art benchmark for multilingual translation is FLORES-200 (NLLB Team et al., 2022). The dataset comprises 3001 sentences sampled from English-language Wikimedia projects (viz. Wikinews, Wikijunior, and Wikivoyage) and professionally translated into more than 200 languages. Unfortunately, there is little to no data for low-resource Finno-Ugric languages. Purason et al. (2024) expanded the benchmark by translating the first 250 rows of FLORES into 9 (later 17) Finno-Ugric languages with the help of native/fluent speakers. This Smugri FLORES test set, with sentences in Karelian Proper, Livvi, Ludian, and Veps, was used to evaluate our models.

Preprocessing Steps. As we decided to test each model on documents, our test set had to be document-level. Fortunately, FLORES-200 is not a set of random, unrelated sentences but a collection of short excerpts from Wikipedia, where the sentences are sequential. All we had to do was isolate these paragraphs. When their length exceeded the maximum allowable, we manually divided them into smaller paragraphs in such a way as to avoid incurring a significant loss of context. Thus, the original 250 sentences turned into 87 paragraphs.

5 Experiments

5.1 Fine-Tuning MADLAD to Translate Finno-Ugric Languages

To investigate the effect of paragraphs on the quality of translation of Karelian and Veps, we fine-tune two MADLAD models: one on sentence-level data and the other on paragraph-level data. We translate the languages into Russian and vice versa, using the same data for opposite directions. As a base MADLAD model, we pick the version with 3B parameters. We split the original data into training and validation sets. We randomly select 400 paragraphs from the paragraphed data, preserving the proportion of languages, to create a balanced paragraph-level validation set. Then, we decompose them into sentences to construct a sentence-level validation set amounting to 1818 sentences. This way, we ensure that both models see identical data when training and guarantee the fairness of evaluation. The training set consists of 34 980 paragraphs or 161 774 sentences. The detailed view of the composition of training and validation sets is presented in Table 2.

Typically, in order to teach large language models to translate new languages, one must expand their native vocabulary with new language tokens. In the case of MADLAD, the only language tokens are the ones prepended to the beginning of input segments, indicating the target language. These tokens take the form <2xx>, where xx stands for target language code. For instance, the sequence "<2en> Je suis là." indicates that the French sentence "Je suis là." needs to be translated into English. A language token should be manually prepended to the user's input sequence. This feature of MADLAD allows for the extension of the model into new languages without expanding the model's vocabulary. As long as MADLAD treats a target language indicator as an inherent element of text, we assume that there is no need to make it into a separate token — the model will recognize it anyway. Thus, we prepend four language indicators to the input sequences: <2krl> for Proper Karelian, <2lud> for Ludic Karelian, <2olo> for Livvi-Karelian, and <2vep> for Veps. The codes are taken from the ISO 639-3[8] code set. As for the Russian language, MADLAD encodes it as <2ru>.

[8] https://iso639-3.sil.org/code_tables/639/data

set	krl-ru	lud-ru	olo-ru	vep-ru	ru-krl	ru-lud	ru-olo	ru-vep
train-sent	31 583	7681	21 734	19 889	31 583	7681	21 734	19 889
valid-sent	685	185	499	449	685	185	499	449
train-par	6803	1603	4701	4383	6803	1603	4701	4383
valid-par	150	39	110	101	150	39	110	101

Table 2. The distribution of sentence-level (sent) and paragraph-level (par) parallel data by language for training and validation.

5.1.1 Fine-Tuning Details

Let us now briefly discuss the setup of fine-tuning. The codebase of MADLAD-400 relies on the t5x[9] framework of Google's own development, which automatizes the process of building and training large language models. However, this framework is not yet widely used, and the number of helpful resources is very limited. Instead, we opted for a HuggingFace[10] version of the model. The HuggingFace Transformers library (Wolf et al., 2020) provides a convenient set of tools for operating with pre-trained Transformer-based models, including MADLAD-400.

Using HuggingFace, we fine-tune both studied models for 10 epochs under equal conditions. We set the hyperparameters of Seq2SeqTrainingArguments to their default values with the following exceptions:

- We limit the generation length to 256 tokens.
- Following the MADLAD paper, we set up an inverse square root scheduler with 300 warmup steps.
- We distribute fine-tuning across 8 GPUs with 16 gradient accumulation steps. To approximately equalize the number of optimization steps for both models, we adjust the batch size depending on the total amount of data: 8 examples per node for paragraph-level data and 32 examples per node for sentence-level data. Thus, we get 2730 optimization steps for the paragraph-trained model and 3160 steps for the sentence-trained model.

We perform fine-tuning on the LUMI[11] supercomputer with AMD Instinct MI250X GPUs.

5.1.2 Evaluation Details

From this point onwards, we adhere to the principle that each model is trained, evaluated, and used exclusively on its respective kind of data, whether sentence- or paragraph-level. The only exception is the final evaluation, where our goal is to comprehensively examine the behavior of the two models in all possible scenarios. We use both models to translate the sentences from the Smugri FLORES benchmark and the paragraphs from its paragraphed version. It means that apart from the foreseen scenario, we want to see how well the paragraph-trained model handles sentences and the sentence-trained model generalizes to paragraphs. For generation, we use the standard beam size of 5. We evaluate the translations with BLEU and chrF++. We use the SacreBLEU (Post, 2018) implementation of the metrics. When calculating chrF++, we count only word bigrams.

[9] https://github.com/google-research/t5x

[10] https://huggingface.co/docs/transformers/en/model_doc/madlad-400

[11] https://lumi-supercomputer.eu/

5.2 Leveraging Back-Translation to Improve High-to-Low-Resource Translation

We make use of our models to bootstrap themselves through back-translation. We managed to collect a fairly large amount of monolingual data for each of the low-resource languages. This means that we can translate the obtained sentences and paragraphs into Russian using our models and then make the resulting parallel sets part of the overall training corpora. It is crucial that we keep the original monolingual segments on the target side so that the models can learn to construct sentences and paragraphs properly. Therefore, we should flip the translated dataset: Russian would become the source language, while Proper Karelian, Livvi, Ludian, and Veps would be the targets. We leave the validation set the same as for the previous experiment, modifying only the training set. The final distribution of data by language can be seen below in Table 3.

Inference (translation) is performed with a standard beam size of 5. Fine-tuning begins with the base MADLAD model and lasts for 9 epochs, retaining the same configurations as for the previous experiment. Translation quality is measured with BLEU and chrF++ using the same procedure as described previously. However, since we are primarily interested in the translation of paragraphs rather than the translation of individual sentences, this time we use for evaluation only the paragraphed version of the Smugri FLORES dataset.

data set		krl-ru	lud-ru	olo-ru	vep-ru	ru-krl	ru-lud	ru-olo	ru-vep
train-sent	para	31.6	7.7	21.7	19.9	31.6	7.7	21.7	19.9
	bt	-	-	-	-	45.4	5.9	75.8	144.4
	total	31.6	7.7	21.7	19.9	76.9	13.6	97.6	164.3
valid-sent		685	185	499	449	685	185	499	449
train-par	para	6.8	1.6	4.7	4.4	6.8	1.6	4.7	4.4
	bt	-	-	-	-	9.6	1.2	18.1	33.7
	total	6.8	1.6	4.7	4.4	16.4	2.8	22.8	38.1
valid-par		150	39	110	101	150	39	110	101

Table 3. The distribution of sentence-level (sent) and paragraph-level (par) parallel data (para) and back-translated data (bt) by language for training and validation. Quantities of training data are given in thousands, rounded to the nearest tenth; quantities of validation data are given as is.

5.3 Testing Zero-Shot Capabilities on the Task of English Translation

The phenomenon of transfer leaning allows us to achieve good translation quality even between languages that a model has never seen together during training. In other words, when training a multilingual model to translate in certain directions, we simultaneously enable the model to improve the quality of translation in unseen directions in virtue of sharing hidden cross-lingual representations. In this experiment, we want to explore MADLAD's ability to translate Karelian and Veps in a zero-shot manner, that is, without showing the model any examples. To this end, we pair these languages with English. Since MADLAD was primarily trained on English data and shows better results for English translation, we consider English a good basis for testing zero-shot capabilities of the model. Again, we want to compare the translation at the sentence and paragraph level, so we take the English sentences from Smugri FLORES and combine them into paragraphs in a similar way to the Finno-Ugric data. As investigated models, we choose the base MADLAD model and sentence- and paragraph-level MADLADs fine-tuned with back-translation. We perform translation in both directions (high-to-low-resource and vice versa) using the beam size of 5. We expect that the knowledge of Karelian and Veps to Russian and Russian to English translation will transitively evoke the emergent ability to translate from Karelian and Veps to English. And vice versa.

6 Results

Before we begin to describe the results, it is important to note that the BLEU and chrF++ metrics may sometimes contradict each other. For example, when comparing the evaluation scores of two models on the same dataset, BLEU may show improvement, while chrF++ may show deterioration of quality. We explain this by different specificity of the metrics. Karelian, Veps, and Russian are morphologically rich languages wherein words can have multiple slightly different forms, composed, for example, by joining prefixes, suffixes, or endings. chrF++ takes this into account by calculating character n-grams. A higher chrF++ score indicates that a model better aligned the morphological structures between the languages under evaluation. A higher BLEU score, in turn, suggests that a model spelled more words correctly. Despite their differences, both metrics are highly correlated with human judgments and, therefore, were used in our work.

6.1 Fine-Tuning MADLAD to Translate Finno-Ugric Languages

We fine-tuned two MADLAD models in an identical setting to perform two different tasks: translating sentences and translating paragraphs. In the future, we shall refer to them as the sentence-level (SL) model and the paragraph-level (PL) model. The models saw strictly equal amounts of data. Translation occurred between Russian (a high-resource language) on one side and Proper Karelian, Livvi-Karelian, Ludian, and Veps (low-resource languages) on the other in both directions. After 10 epochs, when the training loss stopped changing significantly (within a thousandth of a unit), we stopped the fine-tuning and chose the best performers based on the BLEU scores on the validation set. For both models, the last epoch was best. Finally, we measured the BLEU and chrF++ scores of the models on sentence- and paragraph-level versions of Smugri FLORES. To use the sentence-level model on paragraphs, we translated them as separate sentences and merged them back. Sentence-level metrics BLEU and chrF++ were applied to paragraphs without compromising the quality of evaluation (we discussed this in detail in the chapter dedicated to methodology). Additionally, we evaluated the performance of the base MADLAD-3B model on both datasets; paragraphs were passed directly to the model without being split into sentences. For each of the models, we provide the evaluation scores on Smugri FLORES and its paragraphed version in Tables 4 and 5, respectively. BLEU and chrF++ scores come separated by a slash. Henceforth, we discourage the reader from directly comparing the scores from different tables because the metrics become no longer comparable when the dataset changes.

In the following three subchapters, we first analyze the evaluation scores of the models when translating sentences, then we examine the scores when translating paragraphs and discover that the paragraph-level model works better than the sentence-level model, and finally, we draw specific examples to demonstrate why this is the case.

6.1.1 Results of Evaluation on the Sentence-Level Dataset

We shall start our analysis by addressing the fine-tuning results on the Smugri FLORES dataset, presented in Table 4. First of all, we note that the base MADLAD model already had some initial ability to translate Karelian and Veps into Russian with an average quality of 7.0 BLEU or 26.3 chrF++, although it probably had not seen a single example of this during pre-training (MADLAD's developers filtered out all data that their LandId model classified beyond the 419 languages included). This phenomenon is especially noticeable in the Proper-Karelian–Russian translation, whose scores are more than twice as high as the others in the BLEU equivalent. We speculate that this gap may be explained by the closer linguistic proximity of Proper Karelian and Russian compared to other language pairs. Thus, MADLAD exhibited decent zero-shot capabilities when a high-resource language was on the target side. As for the high-to-low-resource translation, it had not yet been possible, which is noticeable from the scores. The untrained MADLAD is unaware of the languages into which we want it to translate.

In the course of fine-tuning, MADLAD's performance improved significantly. The average quality of translation *from* Karelian and Veps rose to 19.0 BLEU or 43.9 chrF++, while MADLAD's skills to translate *into* Karelian and Veps developed from none to 4.8 BLEU or 30.4 chrF++ on average. The sentence-level model achieved the highest scores for translation from the Proper Karelian and Livvi languages. Note that the BLEU and chrF++ metrics signal different best translation directions, although the scores conflict by a small margin. We addressed this phenomenon in the preface to this chapter. Generally, the metric scores for each pair of languages are distributed proportionately to the amount of data for this pair in the parallel corpora. The high-to-low-resource translations have lower scores because the model has never seen sentences in Karelian

FLORES				
direction	MADLAD-base	MADLAD-sent	MADLAD-para	$\Delta_{para-sent}$
krl-ru	11.73 / 33.54	19.69 / 45.70	21.56 / 47.01	+1.87 / +1.31
lud-ru	5.03 / 23.83	16.74 / 42.08	18.54 / 43.20	+1.80 / +1.12
olo-ru	5.88 / 24.62	20.44 / 45.35	21.74 / 46.50	+1.30 / +1.15
vep-ru	5.23 / 23.41	19.12 / 42.54	20.52 / 44.03	+1.40 / +1.49
ru-krl	0.73 / 3.43	6.87 / 35.65	7.79 / 37.23	+0.92 / +1.58
ru-lud	0.27 / 2.41	2.07 / 25.31	2.82 / 27.53	+0.75 / +2.22
ru-olo	0.89 / 2.87	4.67 / 29.59	5.04 / 30.44	+0.37 / +0.85
ru-vep	0.77 / 3.18	5.73 / 31.10	6.34 / 31.44	+0.61 / +0.34

Table 4. BLEU and chrF++ scores (separated by slash) of base MADLAD, sentence-level MADLAD, and paragraph-level MADLAD evaluated on the Smugri FLORES benchmark.

and Veps, as opposed to sentences in Russian, which constituted a considerable part of the pre-training data.

An unexpected observation is that the paragraph-level model improved the sentence-level scores of both metrics in all translation directions. The average improvement goes to 1.6 BLEU or 1.3 chrF++ in the low-to-high-resource direction and 0.7 BLEU or 1.2 chrF++ in the high-to-low-resource direction. This paradoxically indicates that the PL model can translate separate sentences better than the identical SL model, even in the absence of extrasentential context. We explain this by the assumption that the PL model better understands long-range dependencies that can arise not only within paragraphs but also within individual sentences; in virtue of seeing more examples of, say, pronominal anaphora in paragraphs, the PL model did a better job of indicating it in sentences. Otherwise, the paragraph-level model is subject to the same phenomena that we pointed out for the SL model.

6.1.2 Results of Evaluation on the Paragraph-Level Dataset

Let us now discuss Table 5. Although not trained to translate Karelian and Veps into Russian nor to translate paragraphs, the base MADLAD model demonstrated the initial ability to perform both tasks simultaneously with an average quality of 10.4 BLEU or 33.5 chrF++. These scores were further improved by fine-tuning to an average of 19.5 BLEU or 46.3 chrF++. As for the translation into Karelian and Veps, the model learned to perform it with a quality of 5.1 BLEU or 33.1 chrF++ on average. The translation leadership is distributed the same way as in the case of the previous sentence-level dataset translation.

The most important results of our study are displayed in the last column of Table 5.

FLORES (paragraphed)				
direction	MADLAD-base	MADLAD-sent	MADLAD-para	$\Delta_{para-sent}$
krl-ru	14.55 / 40.22	20.22 / 48.15	21.66 / 49.28	+1.44 / +1.13
lud-ru	8.78 / 31.06	17.36 / 44.47	19.91 / 46.76	+2.55 / +2.29
olo-ru	9.88 / 32.02	20.93 / 47.58	23.02 / 49.34	+2.09 / +1.76
vep-ru	8.55 / 30.85	19.62 / 45.17	21.09 / 46.79	+1.47 / +1.62
ru-krl	0.61 / 3.31	7.16 / 38.55	8.53 / 39.37	+1.37 / +0.82
ru-lud	0.32 / 2.41	2.28 / 27.98	3.24 / 29.37	+0.96 / +1.39
ru-olo	0.63 / 2.78	4.96 / 32.25	5.40 / 33.12	+0.44 / +0.87
ru-vep	0.34 / 3.18	6.08 / 33.86	6.14 / 32.60	+0.06 / -1.26

Table 5. BLEU and chrF++ scores (separated by slash) of base MADLAD, sentence-level MADLAD, and paragraph-level MADLAD evaluated on the Smugri FLORES benchmark divided into paragraphs.

Here, we can see that the paragraph-level model significantly improved the scores of the sentence-level model when evaluating paragraphs. The average improvement is 1.9 BLEU or 1.7 chrF++ in the low-to-high-resource direction and 0.7 BLEU or 0.4 chrF++ in the high-to-low-resource direction. We interpret these results as evidence that translating sentences individually is less effective than translating them together, even when the context is only five sentences long.

The only case when translation quality drops noticeably when moving from the sentence-level to the paragraph-level system is translation from Russian to Veps. The BLEU score indicates a slight improvement, whereas the chrF++ score detects a decrease of 1.26 points. Without human assessment, we cannot say for sure the reason for this.

6.1.3 Case Study of Paragraph Translation

To see firsthand how the paragraph-level model is superior to the sentence-level model, we addressed the discourse-level phenomena such as coreference, lexical cohesion, and lexical disambiguation, presenting a major obstacle for sentence-level NMT. The first phenomenon inevitably arises when translating from Karelian or Veps to Russian since, unlike Russian, these languages have gender-neutral pronouns. Thus, the model must guess what gender one is talking about (male, female, or neuter) based on the available context. A single sentence often does not provide sufficient context for solving this task. We came up with three tricky paragraphs in Russian (of which we are fluent speakers) and asked a professional linguist to translate them into Livvi. We envisioned that the paragraphs in Livvi would contain a gender-neutral pronoun "häi," gendering which in Russian would require an extrasentential context. We then used our models to translate the paragraphs from Livvi to Russian. The SL model translated the paragraphs sentence by sentence, while the PL model translated them fully. The results are reported in Table 6.

In order for the reader to understand what we are talking about, we accompany all Russian-language paragraphs with their translations into English, which we made ourselves. In English, as in Russian, pronouns are not gender-neutral, which allows us to preserve the essence of the examples. We also note that since both models produced grammatically correct outputs, we had no problem translating them into English while maintaining the original structure and meaning. All of the provided examples are oriented toward revealing gender bias: out of context, a model is forced to guess the appropriate gender of a pronoun based on the statistics of the dataset on which it was trained.

In the first example (1), the sentence-level model, noticing that the second sentence is about a teacher in a kindergarten, chooses the pronoun "she," although from the previous sentence, it becomes clear that one is talking about a man. The paragraph-level model successfully passes the test using the correct pronoun "he." The second example (2) also manages to trick the sentence-level model into using the incorrect pronoun "he" when discussing a builder. The paragraph-level model, in the presence of the context, figures out that the gender-neutral pronoun "häi" refers to a woman named Masha and utilizes

N		Paragraphs
(1)	Source	Andrei ei voinnuh nosta aijalleh da myöhästyy ruadoh. Hänel pidäy olla aijalleh ruavos, sendäh gu häi ruadau kazvattajannu päivykois.
	Reference	Андрей проспал и опаздывает на работу. Ему надо быть вовремя на работе, так как он работает воспитателем в детском саду. [Andrey overslept and is late for work. He needs to be at work on time, since he works as a teacher in a kindergarten.]
	SL-translation	Андрей не смог встать вовремя и опоздал на работу. Ей надо быть на работе вовремя, потому что она работает воспитателем в дневной школе. [Andrey could not get up on time and was late for work. She needs to be at work on time because she works as a teacher at a day school]
	PL-translation	Андрей не мог встать вовремя и опоздал на работу. Ему нужно быть вовремя на работе, потому что он работает воспитателем в детском саду. [Andrey could not get up on time and was late for work. He needs to be on time at work because he works as a kindergarten teacher.]
(2)	Source	Mašal on tänäpäi lebopäivy da ehtäl häi lähtöy teatrah. Häi tahtos puaksumbah kävvä teatrah, ga häi ruadau rakendajannu da ehtypuoleh äijäl väzyy.
	Reference	У Маши сегодня выходной, и вечером она идёт в театр. Она хотела бы чаще ходить в театр, но она работает строителем и к вечеру очень устаёт. [Masha has a day off today, and in the evening she is going to the theater. She would like to go to the theater more often, but she works as a builder and gets very tired in the evening.]
	SL-translation	У Маши сегодня рождественский праздник, и вечером она пойдёт в театр. Он хотел бы чаще ходить в театр, но работает строителем и вечером очень устает. [Masha has a Christmas holiday today, and in the evening she will go to the theater. He would like to go to the theater more often, but he works as a builder and gets very tired in the evening.]
	PL-translation	У Маши сегодня день отдыха, и вечером она отправляется в театр. Она хотела бы чаще ходить в театр, но работает строителем и вечером очень устает. [Masha has a rest day today, and in the evening she goes to the theater. She would like to go to the theater more often, but she works as a builder and gets very tired in the evening.]
(3)	Source	Maks kirjutti omale susiedale Ol'gale, ongo hänel kois jauhuo. Häi tahtoi pastua piiruan, ga unohti, gu hänel loppih jauho.
	Reference	Макс написал своей соседке Ольге, есть ли у неё дома мука. Он хотел испечь пирог, но забыл, что у него закончилась мука. [Max wrote to his neighbor Olga if she had any flour at home. He wanted to bake a cake, but forgot that he had run out of flour.]
	SL-translation	Макс написал своей соседке Ольге, есть ли у нее дома мука. Она хотела испечь пирог, но забыла, что у нее кончилась мука. [Max wrote to his neighbor Olga if she had any flour at home. She wanted to bake a cake, but forgot that she had run out of flour.]
	PL-translation	Макс написал своей соседке Ольге, есть ли у нее дома мука. Она хотела испечь пирог, но забыла, что у нее кончилась мука. [Max wrote to his neighbor Olga if she had any flour at home. She wanted to bake a cake, but forgot that she had run out of flour.]

Table 6. The translations of paragraphs from Livvi-Karelian (Source) to Russian made by the sentence-level model (SL-translation), the paragraph-level model (PL-translation), and a professional linguist (Reference). Each translation in Russian is supplemented with a literal English translation made by us, coming in square brackets. In each sentence, we underline pronouns with a straight line and their referents with a curvy line.

the appropriate pronoun in Russian. However, the third example (3) proved problematic for both models. It is more sophisticated than the preceding examples in that it has two actors: Max (male) and Olga (female). The sentence-level model misfires again, mistaking the "häi," who wants to bake a pie, for a woman. Surprisingly enough, the paragraph-level model makes the same mistake, probably getting confused by two actors

from the first sentence.

From these examples, it becomes partially clear why the PL model outperforms the SL model. We did not consider more general discourse-level phenomena, such as lexical cohesion or disambiguation, because our tests revealed that both models made identical mistakes when dealing with them. This observation indicates that either more training data with more diverse texts is needed or the paragraphs are too short to capture the phenomena.

6.2 Leveraging Back-Translation to Improve High-to-Low-Resource Translation

We conducted back-translation to improve the scores of translation from Russian to low-resource Karelian and Veps. We used the synthetic data generated by the models from monolingual corpora to fine-tune them. The fine-tuning ran for 9 epochs under the same settings as the original fine-tuning, with each model receiving the data it generated mixed with the original parallel data. Then again, the sentence-level and paragraph-level models were evaluated on the paragraphed version of Smugri FLORES using BLEU and chrF++. We skip sentence-level evaluation the original Smugri FLORES dataset because testing the performance of the paragraph-level model on sentences is beyond our primary interest (we bring the results to Appendix I). The results of evaluation on the paragraphed set are presented in Table 7. In this chapter, we observe a significant boost in high-to-low-resource translation quality but discover that the difference between sentence-level and paragraph-level translations becomes less pronounced. We speculate that this is due to the rareness of discourse-level phenomena in the test set, which become even less influential for translation into Karelian and Veps.

6.2.1 Improvements over the Initial Results

The results of evaluation on the paragraphed version of Smugri FLORES indicate a considerable increase in high-to-low-resource translation quality. During the previous experiment, where we used only parallel corpora, the best model to translate sentences from Smugri FLORES (paragraphed) into Karelian and Veps was the paragraph-level MADLAD, which reached the average score of 5.8 BLEU or 33.6 chrF++. The new sentence-level model fine-tuned with synthetic data beat the result with the average of 9.5 BLEU or 40.4 chrF++.

The PL model still dominates the low-to-high-resource translation, although the scores occasionally dropped compared to its previous results. We attribute this to the fact that, having received a lot of synthetic data for translation in one direction, the model partially forgot how to translate in the other direction. Surprisingly enough, the scores of the SL model increased. A partial explanation for this leap may be that Wikipedia-based FLORES ceased to be an out-of-domain benchmark after fine-tuning the model

on Wikipedia data. Hence, the boost in translation quality. Yet, in this case, why such a boost did not affect the paragraph-level model remains an open question.

FLORES (paragraphed)				
direction	MADLAD-base	MADLAD-sent	MADLAD-para	$\Delta_{para-sent}$
krl-ru	14.55 / 40.22	21.20 / 48.97	21.69 / 49.35	+0.49 / +0.38
lud-ru	8.78 / 31.06	17.85 / 44.86	19.61 / 46.14	+1.76 / +1.28
olo-ru	9.88 / 32.02	21.67 / 47.90	22.67 / 48.90	+1.00 / +1.00
vep-ru	8.55 / 30.85	20.29 / 45.72	20.59 / 46.50	+0.30 / +0.78
ru-krl	0.61 / 3.31	12.84 / 45.72	12.46 / 45.98	-0.38 / +0.26
ru-lud	0.32 / 2.41	4.21 / 33.54	4.32 / 33.42	+0.11 / -0.12
ru-olo	0.63 / 2.78	8.75 / 38.67	8.78 / 38.89	+0.03 / +0.22
ru-vep	0.34 / 3.18	12.22 / 43.56	12.03 / 43.12	-0.19 / -0.44

Table 7. BLEU and chrF++ scores (separated by slash) of base MADLAD, sentence-level MADLAD, and paragraph-level MADLAD fine-tuned **with back-translation** as evaluated on the Smugri FLORES benchmark divided into paragraphs.

6.2.2 Sentence- vs. Paragraph-Level Translation Dynamics

The paragraph-level model continues to outperform the sentence-level model in the task of translation into Russian, even though the difference grew smaller compared to the previous results. The PL model was on average 1.9 BLEU or 1.7 chrF++ ahead of the SL model as evaluated on the paragraphed dataset; now, the difference dropped to 0.9 BLEU or 0.9 chrF++. It means that the SL model slowly catches up with the PL model as the dataset size increases. We attribute this to the fact that there are too few discourse-level phenomena in the dataset to make the paragraph-level model strongly stand out. As stated in the previous experiment, the PL model translated sentences better than the SL model since the former had better knowledge of long-range dependencies between words. Now, the SL model seems to be catching up with the knowledge. However, it is still incapable of outperforming the PL model in low-to-high-resource translation directions.

As for the high-to-low-resource translation, the average scores of the two models became virtually equal. This can be considered evidence that extrasentential context plays an increasingly lesser role in translation from Russian into Karelian and Veps. Indeed, the most urgent discourse-level obstacle appearing when translating Finno-Ugric languages is a pronominal anaphora in the form of gender-neutral pronouns. In contrast, pronouns in Russian are gendered, so the problem does not arise when translating in the reverse direction. The other discourse-level phenomena, such as lexical cohesion and disambiguation, are much less common and do not play a significant role in translation of such a small test set.

6.3 Testing Zero-Shot Capabilities on the Task of English Translation

We translated the paragraphed Smugri FLORES dataset from Karelian and Veps to English and vice versa with three models: the original MADLAD model and sentence- and paragraph-level MADLADs. Then, we measured the scores of the translations with BLEU and chrF++, which we report in Table 8. Again, we skip the results of evaluation on the sentence-level FLORES and bring them to Appendix I.

Let us first discuss zero-shot capabilities of the original MADLAD model. MADLAD demonstrated an excellent quality of translation from Karelian and Veps to English with scores up to 21.12 BLEU or 49.04 chrF++ in the case of Proper Karelian. Even though the model likely has never seen a single sentence in these Finno-Ugric languages, it effortlessly reached the translation quality comparable with that for translation into Russian, which we achieved through meticulous fine-tuning.

After fine-tuning the MADLADs to translate from Karelian and Veps to Russian and back, we witnessed a substantial leap in their translation capabilities into English. The scores of translation into Karelian and Veps, which previously fluctuated around zero, skyrocketed to 15.4 BLEU or 47.2 chrF++ in the case of Proper Karelian translation by the sentence-level model. The low-to-high-resource translation direction improved as well, with the paragraph-level model reaching 28.0 BLEU or 55.3 chrF++ for translation from Proper Karelian.

Comparing the SL and PL models, we noticed the most prominent differences in low-to-high-resource translation directions. The average score of SL-MADLAD, amounting to 19.0 BLEU or 47.5 chrF++, was increased by the PL-MADLAD to 19.9 BLEU or 48.3 chrF++. Moreover, the largest gap, which occurred in translation from

FLORES (paragraphed)				
direction	MADLAD-base	MADLAD-sent	MADLAD-para	$\Delta_{para-sent}$
krl-en	21.12 / 49.04	25.25 / 53.84	27.97 / 55.33	+2.72 / +1.49
lud-en	12.93 / 39.50	19.23 / 47.37	19.91 / 48.36	+0.68 / +0.99
olo-en	10.41 / 36.67	18.02 / 46.38	18.35 / 46.54	+0.33 / +0.16
vep-en	9.01 / 33.55	13.34 / 42.28	13.44 / 43.14	+0.10 / +0.86
en-krl	0.73 / 3.73	15.40 / 47.62	13.84 / 46.28	-1.56 / -1.34
en-lud	0.30 / 2.61	3.54 / 31.12	3.65 / 30.67	+0.11 / -0.45
en-olo	0.41 / 2.76	6.49 / 35.39	6.29 / 34.26	-0.20 / -1.13
en-vep	0.34 / 3.08	7.25 / 37.34	7.24 / 37.12	-0.01 / -0.22

Table 8. BLEU and chrF++ scores (separated by slash) of base MADLAD, sentence-level MADLAD, and paragraph-level MADLAD fine-tuned **with back-translation** as evaluated on the Smugri FLORES benchmark divided into paragraphs.

Proper Karelian to English, was +2.7 BLEU or +1.5 chrF++. This may indicate that the paragraph-level system successfully transferred its knowledge of how to handle discourse-level phenomena in Russian translation to English translation. However, we obtained the opposite results in the case of high-to-low-resource translation. The SL model outperformed the PL model with an average score of 0.4 BLEU or 0.8 chrF++. Yet, it is premature to talk about the decisive superiority of the sentence-level model over the paragraph-level one, as most of the differences in their scores are ambiguously small. The only substantial disparity came from translation into Proper Karelian and amounted to 1.6 BLEU or 1.3 chrF++.

Again, we tend to explain the asymmetry in scores, which arises when switching from one translation direction to another, by the lack of discourse-level phenomena in the test set. While gender-neutral pronouns present a major obstacle for sentence-level models translating from Finno-Ugric languages to English, such a common obstacle is wanting in the reverse direction. Language-agnostic discourse-level phenomena, such as lexical cohesion or disambiguation, are too rare to be captured by the small test set of up to 5 sentences per example.

6.4 Comparison with Neurotõlge

The current state-of-the-art NMT model for translating Finno-Ugric languages is Neurotõlge[12], developed by researchers at the University of Tartu. This model supports 20 low-resource Finno-Ugric languages, including Proper Karelian, Livvi, Ludian, and Veps, and 9 high-resource languages, including Russian. Neurotõlge is available via its free API[13]. The current version of Neurotõlge is based on the NLLB-200 model fine-tuned on the Smugri dataset. Neurotõlge is a sentence-level model; thus, it ignores discourse-level phenomena. Upon testing Neurotõlge with the examples from Chapter 6.1.3, we indeed discovered that it makes mistakes when translating gender-neutral pronouns in Karelian and Veps into Russian.

We used Neurotõlge and our paragraph-level MADLAD model fine-tuned with back-translation to translate the paragraphs from the Smugri FLORES dataset. We passed inputs to Neurotõlge as entire paragraphs, without breaking them into sentences. Finally, we evaluated the translations with BLEU and chrF++. The results are presented in Table 9.

Our model lags behind Neurotõlge in translation into Russian, with some differences being relatively small (e.g., 1.5 BLEU or 1.0 chrF+ for Proper-Karelian-to-Russian translation) and others being quite significant (5.7 BLEU or 4.9 chrF++ for Veps-Russian translation). The superiority of Neurotõlge can be attributed to differences in the training data (although we shared many data sources with Smugri), using a different model,

[12] `https://neurotolge.ee/`

[13] `https://neurotolge.ee/api`

FLORES (paragraphed)			
direction	Neurotõlge	MADLAD-para	Δ
krl-ru	23.20 / 50.32	21.69 / 49.35	-1.51 / -0.97
lud-ru	21.63 / 48.27	19.61 / 46.14	-2.02 / -2.13
olo-ru	26.68 / 51.76	22.67 / 48.90	-4.01 / -2.86
vep-ru	26.68 / 51.42	20.59 / 46.50	-5.69 / -4.92
ru-krl	10.69 / 43.65	12.46 / 45.98	+1.77 / +2.33
ru-lud	3.44 / 31.14	4.32 / 33.42	+0.88 / +2.28
ru-olo	7.07 / 36.33	8.78 / 38.89	+1.71 / +2.56
ru-vep	12.30 / 43.00	12.03 / 43.12	-0.27 / +0.12

Table 9. BLEU and chrF++ scores (separated by slash) of Neurotõlge and paragraph-level MADLAD fine-tuned **with back-translation** as evaluated on the Smugri FLORES benchmark divided into paragraphs.

or the fact that the model was fine-tuned for a much greater number of translation directions, benefiting the translation into Russian. On the other hand, our model managed to outperform Neurotõlge in translation into Proper Karelian, Livvi, and Ludian within a range of 1-2 BLEU/chrF++. The scores for Veps-Russian translation are ambiguous but roughly indicate the parity of the translation quality. Even though the results of our model yield to those of Neurotõlge, it still possesses the key benefit of capturing discourse-level phenomena.

7 Discussion

The experiments we did and the results we got can be divided into two branches: one is exploring the possibilities of MADLAD-400 as a massively multilingual system in application to translating new (in our case, Finno-Ugric) low-resource languages, and the other is investigating the benefits of paragraph-level translation that MADLAD allows in virtue of its pre-pretraining procedure. In this chapter, we summarize our findings along each branch, point out weaknesses, and provide directions for future research.

7.1 Extending MADLAD-400 to More Languages

MADLAD-400 demonstrated a decent zero-shot ability to translate Karelian and Veps, languages beyond the 400 included, into Russian. Further fine-tuning greatly improved the translation scores despite the simplicity of the procedure we followed: prepending a target language code to examples without expanding the model's vocabulary with new target language tokens. MADLAD delegates the responsibility of managing target language tokens to the users. This approach proved simple yet effective, making MADLAD stand out among other multilingual models, such as NLLB, which manage special tokens internally and therefore require the overhead work of expanding its vocabulary. MADLAD showed excellent zero-shotting abilities for English translation, successfully transferring learned knowledge to new translation directions. Compared with the current state-of-the-art translator of Finno-Ugric languages, NLLB-based sentence-level Neurotõlge, our final paragraph-level MADLAD model demonstrated more or less comparable quality of translation into Russian and improved Neurotõlge's scores for translation into Karelian and Veps. Even if MADLAD's scores yield to that of Neurotõlge in some translation directions, the key benefit of our model remains its ability to handle discourse-level phenomena inherent in Karelian, Veps, and other Finno-Ugric languages.

7.2 On Benefits and Challenges of Paragraph-Level Translation

Our decision to make paragraphs of five sentences in length was a rather arbitrary middle ground between the desire to include as much extrasentential context as possible and the reluctance to cause large computation times. One of the directions for further research is investigating the influence of different paragraph sizes on the model's performance. Will two sentences per paragraph be enough to capture most long-range dependencies, or will including more than five sentences in a paragraph be a significant obstacle in fine-tuning the model? We leave these questions to future research and move on to our findings.

The fine-tuning experiments revealed the increase in translation quality when MADLAD translated sentences together as a paragraph instead of translating each one individually. This way, the paragraph-level model could capture long-range dependencies between words, of which we drew examples in Section 6.1.3. However, we encountered

some difficulties that prevent us from definitively declaring the superiority of paragraph-level translation. The difference between the sentence-level and paragraph-level systems tends to become less pronounced when the two systems get enough training, especially in the case of translation from Russian to Karelian and Veps. The latter observation can be explained by the fact that gender-neutral pronouns, a major obstacle for sentence-level models when translating into Russian, are no longer present in the opposite direction. Other discourse-level phenomena, such as lexical cohesion or disambiguation, are too rare to be captured by a small sentence-oriented test set that FLORES is, even though we were able to extract paragraphs from it. On the other hand, conflicting metric scores indicate that further human evaluation of the translation results is needed. After all, alignment between sentences, whether grammatical or stylistic, is assuredly a crucial part of translating texts on the human level. Therefore, there is a need to create a diverse, high-quality paragraph-level translation benchmark that captures the common discourse-level phenomena.

8 Conclusion

Throughout this work, we conducted a series of experiments aimed at exploring MADLAD-400's ability to translate unfamiliar languages, specifically four Finno-Ugric languages with an extremely small number of resources - Proper Karelian, Livvi, Ludian, and Veps. At the same time, taking advantage of the fact that MADLAD was partially pre-trained on documents, we examined the differences between the version of this model fine-tuned on sentences and the one fine-tuned on small paragraphs of 5 sentences. In the Introduction, we stated three main research questions, and now we shall answer them.

Can MADLAD-400 be effectively extended to translate more low-resource languages? Yes. Even without fine-tuning, MADLAD-400 has demonstrated decent skills in translating from Karelian and Veps, languages it has never seen, with an average score of 10.4 BLEU or 33.5 chrF++ on the paragraphed Smugri FLORES test set. Fine-tuning MADLAD on sentences with back-translation increased the scores to 20.2 BLEU or 46.9 chrF++. The quality of translation from Russian to low-resource Karelian and Veps rose from none to 9.5 BLEU or 40.4 chrF++.

Is it possible to fine-tune MADLAD-400 for high-quality translation of paragraph-level data? Yes. Our MADLAD-based paragraph-level model turned out to be comparable to Neurotõlge, a state-of-the-art model for translating Finno-Ugric languages. MADLAD is inferior in translating in low-to-high-resource direction within 5 BLEU points but superior in high-to-low-resource direction within 3 BLEU points, as evaluated on the paragraphed Smugri FLORES test set.

Is paragraph-level translation better than sentence-level translation? For the specific task of translating from Karelian and Veps into Russian and vice versa using MADLAD-400, paragraph-level translation outperformed sentence-level translation with a difference of up to 3 BLEU/chrF++ points on the paragraphed Smugri FLORES test set. A case study of translations revealed that the paragraph-level model successfully avoided gender bias in simple cases by utilizing the discourse-level context. The sentence-level model, as expected, failed all tests. Thus, despite learning with short paragraphs only, the paragraph-level model was able to handle some discourse-level phenomena, considerably affecting the quality of translation.

However, the differences between the models slightly leveled out as the amount of data increased after back-translation. The metrics scores for translation into Karelian and Veps became almost equal but ambiguous, requiring further human evaluation. Yet, the paragraph-level model retained its dominance in translation into Russian within 2 BLEU/chrF++ points. We attribute this performance asymmetry to discourse-level phenomena playing a much larger role in translation from Karelian and Veps to Russian than in the opposite direction. Another consideration is that FLORES was not supposed to be used as a paragraph-level test set and needed to be bigger (we obtained only 87 paragraphs). We indicate the need to develop a high-quality paragraph-level test set, capturing various discourse-level phenomena. We leave it to the future work.

Contribution. In a separate paragraph, we would like to highlight our developments in creating a sentence- and paragraph-level dataset for Karelian and Veps that will hopefully simplify subsequent research in this area. We are refraining from publishing the sentence- and paragraph-level datasets we assembled but instead sharing scripts that will allow one to reproduce our results. On our GitHub[14] page, we make public the scripts for parsing, normalizing, and segmenting into sentences the data from VepKar, Omamedia, and Wikipedia. On top of that, we add a script for merging sentences into paragraphs along the procedure described in Chapter 3.1. We hope these tools will contribute to future research.

[14] https://github.com/ozzwoy/Finno-Ugric-Data-Scripts

References

Aharoni, R., Johnson, M., and Firat, O. (2019). Massively multilingual neural machine translation. In Burstein, J., Doran, C., and Solorio, T., editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In Walker, M., Ji, H., and Stent, A., editors, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Boyko, T., Zaitseva, N., Krizhanovskaya, N., Krizhanovsky, A., Novak, I., Pellinen, N., and Rodionova, A. (2022). The open corpus of the veps and karelian languages: Overview and applications. KnE Social Sciences.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. Computational Linguistics, 16(2):79–85.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Moschitti, A., Pang, B., and Daelemans, W., editors, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2023). Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240):1–113.

Deutsch, D., Juraska, J., Finkelstein, M., and Freitag, M. (2023). Training and meta-evaluating machine translation evaluation metrics at the paragraph level. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, Proceedings of the Eighth Conference on Machine Translation, pages 996–1013, Singapore. Association for Computational Linguistics.

Freitag, M. and Firat, O. (2020). Complete multilingual neural machine translation. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., and Negri, M., editors, Proceedings of the Fifth Conference on Machine Translation, pages 550–560, Online. Association for Computational Linguistics.

Hardmeier, C. (2012). Discourse in statistical machine translation: A survey and a case study. Discours, (11).

Hutchins, W. J. (1995). Machine translation: A brief history. In KOERNER, E. and ASHER, R., editors, Concise History of the Language Sciences, pages 431–445. Pergamon, Amsterdam.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351.

Kalamkar, D. D., Mudigere, D., Mellempudi, N., Das, D., Banerjee, K., Avancha, S., Vooturi, D. T., Jammalamadaka, N., Huang, J., Yuen, H., Yang, J., Park, J., Heinecke, A., Georganas, E., Srinivasan, S. M., Kundu, A., Smelyanskiy, M., Kaul, B., and Dubey, P. K. (2019). A study of bfloat16 for deep learning training. ArXiv, abs/1905.12322.

Kocmi, T. and Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. In Nurminen, M., Brenner, J., Koponen, M., Latomaa, S., Mikhailov, M., Schierl, F., Ranasinghe, T., Vanmassenhove, E., Vidal, S. A., Aranberri, N., Nunziatini, M., Escartín, C. P., Forcada, M., Popovic, M., Scarton, C., and Moniz, H., editors, Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In Luong, T., Birch, A., Neubig, G., and Finch, A., editors, Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver. Association for Computational Linguistics.

Kudugunta, S., Caswell, I., Zhang, B., Garcia, X., Xin, D., Kusupati, A., Stella, R., Bapna, A., and Firat, O. (2023). Madlad-400: A multilingual and document-level large audited dataset. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, Advances in Neural Information Processing Systems, volume 36, pages 67284–67296. Curran Associates, Inc.

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? a case for document-level evaluation. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Maruf, S. and Haffari, G. (2018). Document context neural machine translation with memory networks. In Gurevych, I. and Miyao, Y., editors, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Miculicich, L., Ram, D., Pappas, N., and Henderson, J. (2018). Document-level neural machine translation with hierarchical attention networks. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Mejia-Gonzalez, G., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F.,

Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No language left behind: Scaling human-centered machine translation.

Ott, M., Edunov, S., Grangier, D., and Auli, M. (2018). Scaling neural machine translation. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, Proceedings of the Third Conference on Machine Translation: Research Papers, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Hokamp, C., Huck, M., Logacheva, V., and Pecina, P., editors, Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Popović, M. (2017). chrF++: words helping character n-grams. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., and Kreutzer, J., editors, Proceedings of the Second Conference on Machine Translation, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Purason, T., Ivanov, A., Yankovskaya, L., and Fishel, M. (2024). SMUGRI-MT - machine translation system for low-resource finno-ugric languages. In EAMT 2024 Products and Projects track.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language

Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In Erk, K. and Smith, N. A., editors, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2019). MASS: Masked sequence to sequence pre-training for language generation. In Chaudhuri, K. and Salakhutdinov, R., editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 5926–5936. PMLR.

Sun, Z., Wang, M., Zhou, H., Zhao, C., Huang, S., Chen, J., and Li, L. (2022). Rethinking document-level neural machine translation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, Findings of the Association for Computational Linguistics: ACL 2022, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., and Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. AI Open, 1:5–21.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models. ArXiv, abs/2302.13971.

Varis, D. and Bojar, O. (2021). Sequence length is a domain: Length-based overfitting in transformer models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Liu, Q. and Schlangen, D., editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In Su, J., Duh, K., and Carreras, X., editors, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Appendix

I. Results of Evaluation on the Sentence-Level Smugri FLORES Dataset

FLORES				
direction	MADLAD-base	MADLAD-sent	MADLAD-para	$\Delta_{para-sent}$
krl-ru	14.55 / 40.22	20.75 / 46.60	21.54 / 46.82	+0.79 / +0.22
lud-ru	8.78 / 31.06	17.36 / 42.45	18.82 / 43.10	+1.46 / +0.65
olo-ru	9.88 / 32.02	21.20 / 45.72	21.84 / 46.33	+0.64 / +0.61
vep-ru	8.55 / 30.85	19.75 / 43.11	20.11 / 43.61	+0.36 / +0.50
ru-krl	0.61 / 3.31	12.50 / 43.11	12.37 / 43.74	-0.13 / +0.63
ru-lud	0.32 / 2.41	4.13 / 31.13	4.00 / 31.57	-0.13 / +0.44
ru-olo	0.63 / 2.78	8.54 / 36.11	9.37 / 37.28	+0.83 / +1.17
ru-vep	0.34 / 3.18	11.66 / 41.11	10.99 / 40.53	-0.67 / -0.58

Table 10. BLEU and chrF++ scores (separated by slash) of base MADLAD, sentence-level MADLAD, and paragraph-level MADLAD fine-tuned **with back translation** as evaluated on the Smugri FLORES benchmark.

FLORES				
direction	MADLAD-base	MADLAD-sent	MADLAD-para	$\Delta_{para-sent}$
krl-en	19.85 / 46.39	24.32 / 51.31	24.58 / 51.34	+0.26 / +0.03
lud-en	10.81 / 33.85	18.41 / 44.72	17.87 / 44.55	-0.54 / -0.17
olo-en	9.43 / 32.67	17.43 / 43.44	16.43 / 42.67	-1.00 / -0.77
vep-en	5.77 / 27.60	12.48 / 38.92	12.27 / 38.61	-0.21 / -0.31
en-krl	0.96 / 4.62	15.13 / 45.11	12.49 / 43.09	-0.64 / -1.02
en-lud	0.32 / 2.63	3.39 / 28.66	2.98 / 27.82	-0.41 / -0.84
en-olo	0.59 / 2.91	6.07 / 32.23	6.03 / 31.82	-0.04 / -0.41
ru-vep	0.46 / 3.41	6.64 / 34.38	6.37 / 34.20	-0.27 / -0.18

Table 11. BLEU and chrF++ scores (separated by slash) of base MADLAD, sentence-level MADLAD, and paragraph-level MADLAD fine-tuned **with back translation** as evaluated on the Smugri FLORES benchmark.

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Dmytro Pashchenko**,
(author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Paragraph-Level Translation of Low-Resource Finno-Ugric Languages,
   (title of thesis)

   supervised by Mark Fishel and Elizaveta Yankovskaya.
   (supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Dmytro Pashchenko
15/05/2024