DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
59

MAKSYM DEL

Multilingual and Multi-Domain Representational Patterns Across Transformer-Based Models

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

59

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
59

MAKSYM DEL

Multilingual and Multi-Domain Representational Patterns Across Transformer-Based Models

UNIVERSITY OF TARTU
Press
1632

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in Computer Science on October 1, 2024 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor

Prof.	Mark Fišel University of Tartu, Estonia

Opponents

Prof.	Anders Søgaard University of Copenhagen, Denmark
Assoc. Prof.	Mathias Creutz University of Helsinki, Finland

The public defense will take place on November 7 at 12:00 in Narva Rd. 18-1018.

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906 (print) ISSN 2806-2345 (pdf)
ISBN 978-9916-27-698-3 (print) ISBN 978-9916-27-699-0 (pdf)

University of Tartu Press
http://www.tyk.ee/

ABSTRACT

Artificial intelligence models often act like black boxes: they take data and produce predictions, but the internal processes remain unclear. This uncertainty makes it difficult to trust that these models are safe, fair, and reliable. Without a clear understanding of how they work, we cannot be sure they will not generate harmful, biased, or suboptimal results. For example, a model that performs well in one language can fail to maintain the same level of quality when prompted in another language. Although we can easily access the internal workings of the model - which consists of countless numerical values - the challenge is to make sense of how decisions are made.

Two types of models are particularly important in today's globalized world: multilingual models, which handle multiple languages, and multi-domain models, which operate on data across various topics and styles. This thesis aims to advance the interpretability of multilingual and multi-domain Transformer-based models, focusing on the evolution of their internal representation and manipulation of information across different languages and domains.

Our key contribution is establishing two core representational phenomena in Transformer-based models: multilingual abstraction and multi-domain specialization. Multilingual abstraction shows how sentence representations evolve from language-specific states to more generalized, language-agnostic states as the information flows through the layers of the model. Multi-domain specialization highlights the retention and enhancement of domain-specific features in sentence representations throughout layers of the model. Importantly, we show that these phenomena appear consistently across different model types and variants, training datasets, and architectural variations, suggesting a universal pattern in Transformer-based models.

While our central deliverables are qualitative insights into the workings of Transformer-based models, we also, as a means to our end, introduce a specialized methodology for comparing multilingual representations, resolve conflicting evidence and clashing literature, and present a practical application of our multi-domain findings to support the analysis.

We hope our findings about multilingual and multi-domain models can assist in improving the democratization, safety, fairness, and accessibility of AI technology, especially for underrepresented languages and domains.

CONTENTS

List of original publications **11**

1. Introduction **13**

2. Background **16**
2.1. Transformer Architecture 16
2.1.1. Encoder and Decoder Blocks 16
2.1.2. Self-Attention Mechanism 18
2.1.3. Multi-Head Attention 19
2.1.4. Position-wise Feed-Forward Networks 19
2.1.5. LayerNorm 20
2.2. Model Types 21
2.2.1. Language Modeling 21
2.2.2. Pretrained Language Models (PLMs) 22
2.2.3. Machine Translation (MT) 23
2.3. Representations Comparison Indexes 24
2.3.1. Canonical Correlation Analysis (CCA) 24
2.3.2. Projection Weighted CCA (PWCCA) 25
2.3.3. Singular Vector CCA (SVCCA) 25
2.3.4. Centered Kernel Alignment (CKA) 26

3. Key Results **27**
3.1. Representations Comparison with ANC 27
3.1.1. Representations Comparison in Multilingual Models 28
3.1.2. Averaged Neuronwise Correlation 29
3.2. Result I: Multilingual Abstraction 31
3.2.1. Evidence 1: Representations Comparison 32
3.2.2. Evidence 2: Unsupervised Visualization 34
3.2.3. Evidence 3: Suggestive Evidence from Related Works . . . 35
3.2.4. Evidence 4: Paraphrase Analysis 36
3.3. Result II: Universality of Multilingual Abstraction 38
3.3.1. Evidence 1: Universality Across Architectural Variations . 38
3.3.2. Evidence 2: Universality Across Scale 40
3.3.3. Evidence 3: Universality Across Training Objectives 41
3.4. Result III: Multi-Domain Specialization 43
3.4.1. Evidence 1: t-SNE Visualization 43
3.4.2. Evidence 2: PCA Visualization 44
3.4.3. Evidence 3: k-means Clustering 45
3.5. Result IV: Universality of Multi-Domain Specialization 46
3.5.1. Evidence 1: Universality Across Training Objectives 46
3.5.2. Evidence 2: Document-level Representations 47

3.5.3. Evidence 3: Another Language Pair 48

4. Addressing Objections and Practical Application **50**
4.1. Multilingual Abstraction: Addressing Conflicting Literature . . . 50
4.1.1. Evidence 1: Cosine Similarity 51
4.1.2. Evidence 2: Probing Task Analysis 52
4.1.3. Evidence 3: Argument from Self-Attention 52
4.2. Multilingual Abstraction: Addressing Outliers 53
4.3. Multilingual Abstraction: Addressing Conflicting Measurements . 55
4.3.1. Evidence 1: Zero-shot Cross-lingual Transfer 57
4.3.2. Evidence 2: Per-Layer Matching Accuracy 57
4.3.3. Evidence 3: Results from SVCCA and PWCCA 58
4.4. Multi-Domain Specialization: Practical Application 60
4.4.1. Existing Framework 61
4.4.2. Revising the Existing Framework 62

5. Conclusion **63**

Bibliography **64**

Acknowledgements **73**

Sisukokkuvõte (Summary in Estonian) **74**

6. Publications **77**

Curriculum Vitae **124**

Elulookirjeldus (Curriculum Vitae in Estonian) **125**

LIST OF FIGURES

1. Transformer architecture. Image source: d2l.ai (Zhang et al. 2021). 17
2. Schematic of the cross-lingual representations comparison process. 28
3. ANC results for the mBERT and XLM-R models. 33
4. Multilingual abstraction in mBERT revealed by SVCCA, PWCCA, and CCA. 34
5. t-SNE visualization of sentence representation transformation in mBERT, highlighting multilingual abstraction. 35
6. Comparison of sentence representations for translations and paraphrases in BERT models, illustrating the abstraction process. . . . 37
7. ANC results illustrating multilingual abstraction across different normalization methods in XLM-R. 39
8. ANC analysis across XLM-R models of varying sizes, indicating scale-independent multilingual abstraction. 41
9. ANC cross-lingual representational similarity for the XGLM CLM-style models of different sizes. All models follow a similar Multilingual Abstraction pattern. We aggregate among en-fr, en-de, en-ru, and en-et pairs and show similarity average and spread. 42
10. t-SNE visualization of sentence representations across Transformer layers, color-coded by domain. Layer 0 represents fixed encoder embeddings, Layers 1–6 are encoder layers, Layer 7 represents fixed decoder embeddings, and Layers 8–13 are decoder layers. The model learns to distinguish between domains despite not being explicitly provided with any domain information. 44
11. PCA visualization showing sentence representations by domain across Transformer layers. 44
12. Cluster distribution of sentence representations in the EN-ET NMT model, indicating predominant domain-driven clustering. 45
13. Clustering of domains in the XLM-R model, illustrating multi-domain specialization. 47
14. Document clustering in the EN-ET NMT model, closely aligning with the original data domains. 47
15. Domain clustering in the DE-EN NMT model, strongly aligned with original data domains. 48
16. Different views on language representation in mBERT. The representations go from similar to dissimilar and thus diverge (left) vs from dissimilar to similar and thus converge (right). Adapted from Singh et al. (2019a) and Muller et al. (2021). The graph illustrates that two related works arrive at two opposing answers to the same question (convergence/divergence of language representations). As we debunk in this section, authors use different similarity indexes and sentence representation types which results in different patterns. 50

17. Cosine similarities of sentence representations under three different pooling strategies. The legend indicates whether similarity is measured between parallel sentences ("parallel") or arbitrary pairs of sentences ("random"). 51
18. Accuracy of matching the closest sentence vector to the source sentence across pooling strategies averaged over four languages. . . . 52
19. Multilingual abstraction across language pairs in mBERT and XLM-R, with outliers highlighted. 54
20. Layer-specific language similarities, identifying Urdu, Hindi, Swahili, and Thai as outliers. 54
21. Linguistic clustering based on CKA distances, demonstrating language groupings and isolating outliers. 55
22. Counter-intuitive CKA (dis)similarity of XLM-Normformer layers. CKA index shows drastic dissimilarity for layers 6-12. 56
23. Layer-wise sentence matching accuracy in XLM-Normformer. . . 58
24. SVCCA similarity analysis for XLM-Normformer layers, showing a multilingual abstraction pattern. 59
25. PWCCA similarity analysis for XLM-Normformer layers, demonstrating a discrepancy from CKA findings. 59
26. Comparison of the existing automatic domains framework (left) and our proposed revision (right). 61

LIST OF TABLES

1. Model details for XLM-R models we study. *type*: training objective of the model, *#params*: number of parameters, *l*: number of layers, *n*: number of hidden units (neurons at each layer), *#lgs*: number of languages used in pertaining. 40
2. Model details for XGLM models we study. *type*: training objective of the model, *#params*: number of parameters, *l*: number of layers, *n*: number of hidden units (neurons at each layer), *#lgs*: number of languages used in pertaining. 42
3. Accuracy of XLM-Roberta Base Transformers pre-trained with different normalization schemes and fine-tuned on the English portion of the XNLI sentence classification task. The models show similar zero-shot cross-lingual transfer performance. 57

LIST OF ORIGINAL PUBLICATIONS

Publications included in the thesis

I **Del**, **Maksym** and Fishel, Mark (Nov. 2022). "Cross-lingual Similarity of Multilingual Representations Revisited". In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online only: Association for Computational Linguistics, pp. 185–195. URL: https://aclanthology.org/2022.aacl-main.15.

II **Del**, **Maksym** and Fishel, Mark (2021a). "Similarity of Sentence Representations in Multilingual LMs: Resolving Conflicting Literature and a Case Study of Baltic Languages". In: *Baltic Journal of Modern Computing* 10. URL: https://api.semanticscholar.org/CorpusID:249921326.

III **Del**, **Maksym**, Korotkova, Elizaveta, and Fishel, Mark (Nov. 2021b). "Translation Transformers Rediscover Inherent Data Domains". In: *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics, pp. 599–613. URL: https://aclanthology.org/2021.wmt-1.65.

Publications not included in the thesis

IV Luhtaru, Agnes, Purason, Taido, Vainikko, Martin, **Del**, **Maksym**, and Fishel, Mark (2024). *To Err Is Human, but Llamas Can Learn It Too*. arXiv: 2403.05493 [cs.CL]. URL: https://arxiv.org/abs/2403.05493.

V **Del**, **Maksym** and Fishel, Mark (July 2023). "True Detective: A Deep Abductive Reasoning Benchmark Undoable for GPT-3 and Challenging for GPT-4". In: *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*. Toronto, Canada: Association for Computational Linguistics, pp. 314–322. URL: https://aclanthology.org/2023.starsem-1.28.

VI Korotkova, Elizaveta, Luhtaru, Agnes, **Del**, **Maksym**, Liin, Krista, Deksne, Daiga, and Fishel, Mark (2019). *Grammatical Error Correction and Style Transfer via Zero-shot Monolingual Translation*. arXiv: 1903.11283 [cs.CL]. URL: https://arxiv.org/abs/1903.11283.

VII **Del**, **Maksym**, Tättar, Andre, and Fishel, Mark (Oct. 2018). "Phrase-based Unsupervised Machine Translation with Compositional Phrase Embeddings". In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computa-

tional Linguistics, pp. 361–367. URL: https://aclanthology.org/W18-6407.

VIII Rikters, Matīss, Amrhein, Chantal, **Del**, **Maksym**, and Fishel, Mark (Sept. 2017). "C-3MA: Tartu-Riga-Zurich Translation Systems for WMT17". In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 382–388. URL: https://aclanthology.org/W17-4738.

Author's contribution to the publications

The author's contributions to Publications I, II, IV, and V include coming up with the research ideas and hypotheses, designing and conducting experiments, and writing the papers.

In Publication III, the author performed the interpretability part of the work and came up with the idea for the practical application, while the coauthor focused on the experimental design, analysis of results, and conducting experimental and ablation studies.

In Publication VI, the author performed part of the model evaluations and outlined the related work. In Publication VII, the author performed and described the experiments for the Ukrainian language. In Publication VIII, the author implemented and described the experiments on coverage penalties.

1. INTRODUCTION

Modern Transformer-based models are reshaping our interactions with the digital world and advancing the field of Artificial Intelligence (AI) research (Anthropic 2023; OpenAI 2023; Touvron et al. 2023). These neural models are typically first trained and then used for inference. Most Transformer (Vaswani et al. 2017) models consist of weights and biases arranged in a stack of several layers. After training, these weights and biases take their final values, and layers process information to produce outputs according to the model's performance. However, despite the effectiveness of these models in making predictions, understanding their inner workings remains a challenge. Although we can access the numerical values of the model's weights, biases, and layer outputs, comprehending how these components interact dynamically during inference, process input text, and form predictions is still not fully resolved. Advances in model interpretability can enhance AI technology's safety, fairness, and accessibility by detecting and reducing biases and errors in critical applications, making models more useful and secure globally.

This thesis focuses on interpreting multilingual and multi-domain Transformer-based models. Multilingual models handle input text in multiple languages, while multi-domain models perform effectively across various specialized domains. In this thesis, a ***multilingual model*** is the single Transformer-based neural network trained on a dataset of mixed languages, just as the monolingual model would be trained (without architectural multilingualism-specific modifications). Similarly, a ***multi-domain model*** is a single Transformer-based neural network trained on a dataset of mixed domains, just as the single-domain model would be trained (without architectural modifications specific to multiple domains).

Understanding how these models internally represent and manipulate multilingual and multi-domain information can improve technology for underrepresented languages and domains, contributing to the democratization and globalization of AI.

The primary contributions of this thesis lie in the qualitative findings related to the interpretability of multilingual and multi-domain Transformer-based models. Unlike typical computer science research, which often focuses on solving practical applications or improving performance metrics, the insights into neural network inner workings are the end deliverables here. This exploration into the mechanisms of neural models is similar to research in fields like neuroscience or biology, where understanding underlying systems is valued on its own, while practical applications belong to different fields like medicine and biotech. Although we present some practical and methodological contributions (e.g., Section 4.4), they are complementary elements that enrich the narrative and exemplify the applications of some findings. This approach shifts the focus from application-driven outcomes to a principle-driven understanding of AI systems.

In interpreting the model, one can aim for a low-level understanding by focus-

ing on interactions of fine-grained elements such as (linear combinations of) neurons and weights (Bricken et al. 2023) or a higher-level mechanistic picture that explains how larger elements like whole sentence representations are processed by Transformer layers (Voss et al. 2021). This thesis aims at the latter analysis: a higher-level comparative description of how sentence representations from multiple languages and domains develop as they flow through the model's layers. The methodology includes techniques such as representational similarity analysis (Kornblith et al. 2019), representation visualization (F.R.S. 1901; Maaten et al. 2008), k-means and hierarchical clustering (Lloyd 1982; Murtagh et al. 2012), and probing (Gupta et al. 2015; Köhn 2015).

This thesis defines two core phenomena identified in our analysis: ***multilingual abstraction*** and ***multi-domain specialization***. We introduce these terms to refer to this thesis's two central observed phenomena. By specialization, we mean that the network creates different representations for different domains/languages, and by abstraction, we mean the opposite (the representations for different languages/domains are fairly similar, which means they were possibly abstracted into the common latent representation).

Multilingual abstraction describes how representations in multilingual models evolve from an initial language-specific state at the first layer to a more generalized, language-agnostic state in the intermediate layers. This transformation starts with the model processing distinct linguistic inputs and progresses to a more unified representation of the information.

Conversely, multi-domain specialization shows how models that handle multiple domains maintain and enhance domain-specific representations throughout the network. Unlike the abstraction seen in multilingual contexts, these models preserve domain distinctions from the initial layers and continue to refine them as the representation progresses through the model.

While establishing these phenomena in a particular pre-trained model is significant, showing that these patterns manifest across different models and settings makes findings potentially more scalable. Demonstrating that models repeatedly converge to the same representation phenomena can drive interpretability by studying smaller or different models. This phenomenon, known as ***universality***, indicates that the identified representation patterns are consistent across various models (Olah 2022; Olah et al. 2020). Proving complete universality is out of the scope of this work; instead, we aim to provide empirical evidence for a high degree of universality across Transformer models in general (and not just specific language models, translation models, or training/architectural setups).

Our work studied models of different types, architectural variations, training data, regimes, and scales. Our phenomena occur in all settings studied, allowing us to call them representation patterns instead. However, we acknowledge that absolute universality cannot be guaranteed, but the ample positive evidence and absence of counterevidence allow us to claim a high degree of universality.

Our research aims to demonstrate that multilingual abstraction and multi-domain

specialization occur in Transformer models (Sections 3.2 and 3.4) and are universal phenomena (Sections 3.3 and 3.5). Additionally, we address challenges such as inconsistencies in existing literature (Section 4.1), outlier cases (Section 4.2), and conflicting measurements (Section 4.3), develop a novel methodology for assessing representational similarity (Section 3.1), and introduce practical applications of our findings (Section 4.4).

Our contributions are as follows:

- Characterization of multilingual abstraction and multi-domain specialization in Transformer models.
- Establishing the universality of these phenomena across various Transformer models, tasks, and scales.
- Development of the Averaged Neuronwise Correlation (ANC) methodology for representation comparison.
- Practical application of our findings to enhance domain adaptation in neural machine translation.

The rest of the thesis is organized as follows:

- *Section 2*: Background information on Transformer architecture (Section 2.1), models (Section 2.2), and representational similarity indexes (Section 2.3).
- *Section 3*: Method for effective comparison of multilingual representations (Section 3.1) and exploration of four major claims: multilingual abstraction (Section 3.2), its universality (Section 3.3), multi-domain specialization (Section 3.4), and the universality of multi-domain specialization (Section 3.5).
- *Section 4*: Addresses objections to our multilingual abstraction arguments, covering conflicting evidence from literature (Section 4.1), outlier languages (Section 4.2), conflicting representation comparison measurements (Section 4.3), and presents the practical application of multi-domain specialization (Section 4.4).
- *Section 5*: Summarizes key findings and suggests future research directions.

Our publications related to the thesis are as follows:

- In **Del** et al. (2022), we introduce ANC (Section 3.1), the universality experiments of multilingual abstraction (Section 3.3), and address conflicting representation comparison measurements (Section 4.3).
- In **Del** et al. (2021a), we present experiments on multilingual abstraction (Section 3.2), tackle conflicting literature (Section 4.1), and introduce rebuttals concerning outlier languages (Section 4.2).
- In **Del** et al. (2021b), we explore the concept of multi-domain specialization (Section 3.4), its universality (Section 3.5), and its practical application (Section 4.4).

2. BACKGROUND

2.1. Transformer Architecture

The Transformer is a neural network architecture (Vaswani et al. 2017) that has revolutionized the field of natural language processing and has become the foundation for many state-of-the-art deep learning models. Its innovative design, which relies on self-attention mechanisms rather than traditional recurrent or convolutional layers, allows for significantly improved parallelization and scalability. Consequently, the Transformer has been widely adopted in various domains, including machine translation (Vaswani et al. 2017), text generation (Brown et al. 2020b), and even computer vision (Dosovitskiy et al. 2021) and reinforcement learning tasks (Chen et al. 2021).

At a high level, the Transformer can be implemented in different configurations, such as the encoder-decoder, encoder-only, and decoder-only setups. It consists of multiple components, including encoder and decoder blocks, self-attention mechanisms, multi-head attention, position-wise feed-forward networks, and layer normalization (Figure 1).

These components work hierarchically by processing the input data through a series of interconnected layers. Each layer takes the previous layer's output as its input, refining and enhancing the data representation. The self-attention mechanism captures context and dependencies within the input, which is combined across multiple attention heads in the multi-head attention mechanism. Crucially, self-attention allows the model to capture long-range dependencies in the input sequence and works particularly well for the discrete-input data such as langauge. The resulting output is further processed by the position-wise feed-forward networks, adding non-linearity and complexity to the representations. Layer normalization ensures stability during training by mitigating covariate shift (Ba et al. 2016). By propagating the information through this hierarchical arrangement, Transformer can efficiently capture the underlying structure and dependencies in the input data, producing increasingly abstract and expressive representations at each level.

There are several variations of the original Transformer architecture, with modifications primarily targeting improvements in efficiency and performance. One such aspect explored in this thesis is alternative normalization techniques, which can significantly impact the model's ability to train and generalize. In the following subsections, we will discuss each of the Transformer's components in detail, providing insights into their functionality, purpose, and role within the overall architecture.

2.1.1. Encoder and Decoder Blocks

The Encoder and Decoder blocks are the primary building blocks of the Transformer architecture. They are responsible for processing the input data and gen-

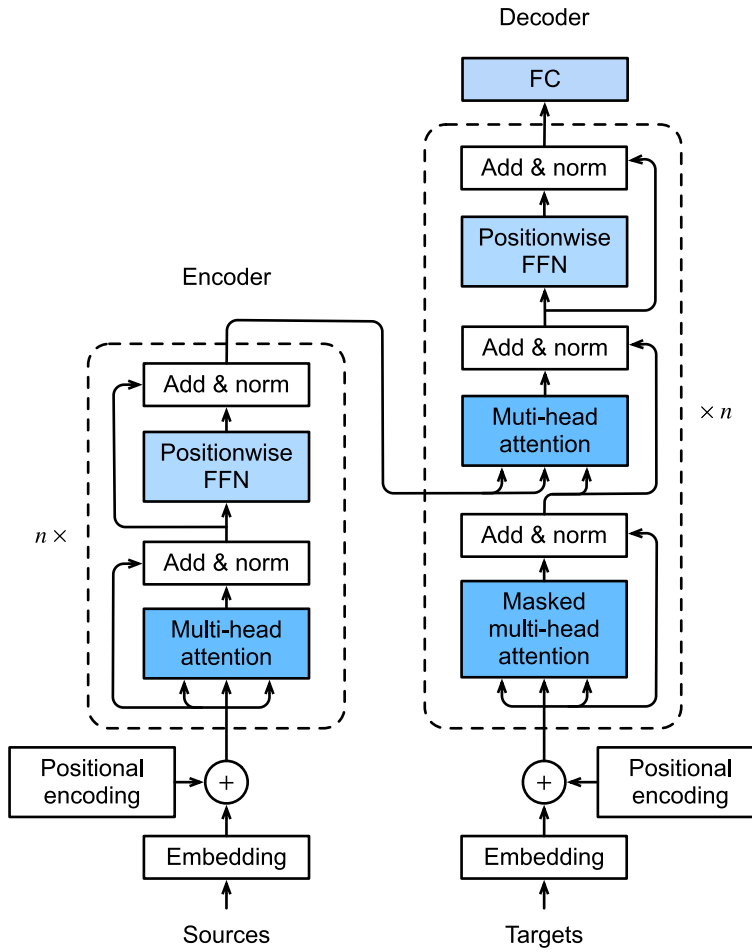

Figure 1: Transformer architecture. Image source: d2l.ai (Zhang et al. 2021).

erating the output representations. Intuitively, the Encoder block captures the contextual information in the input sequence, while the Decoder block generates the output sequence based on the encoded input and its context. These blocks contain multiple layers, each containing self-attention, multi-head attention, position-wise feed-forward mechanisms, and the necessary normalization components. By stacking multiple layers, the Encoder and Decoder blocks can learn increasingly complex and abstract representations of the input data, effectively capturing the underlying structure and dependencies.

Encoder Block. Each encoder block in the transformer architecture consists of two sub-layers: a multi-head self-attention layer and a position-wise feed-forward layer. The input to the encoder block passes through each sub-layer, followed by a residual connection and layer normalization. This process can be described as follows:

$$
\begin{aligned}
x' &= \text{MultiHead}(x) + x \\
o &= \text{LayerNorm}(x') \\
o' &= \text{FFN}(o) + o \\
z &= \text{LayerNorm}(o')
\end{aligned}
$$

The input sequence is processed through n stacked encoder blocks.

Decoder Block. Each decoder block in the transformer architecture consists of three sub-layers: a multi-head self-attention layer, a multi-head cross-attention layer, and a position-wise feed-forward layer. The multi-head self-attention layer processes the output sequence generated so far, the cross-attention layer attends to the encoder output, and the feed-forward layer further processes the combined information. Similar to the encoder block, the input to each sub-layer in the decoder block is followed by a residual connection and layer normalization. Considering that EncOut is an output of the final Encoder layer, this process can be described as follows:

$$
\begin{aligned}
x' &= \text{MultiHead}(x) + x \\
o &= \text{LayerNorm}(x') \\
o' &= \text{MultiHead}(o, \text{EncOut}) + o \\
z &= \text{LayerNorm}(o') \\
z' &= \text{FFN}(z) + z \\
y &= \text{LayerNorm}(z')
\end{aligned}
$$

The decoder output sequence is generated through n stacked decoder blocks.

2.1.2. Self-Attention Mechanism

The self-attention mechanism is at the heart of the Transformer architecture, allowing the model to weigh and combine different parts of the input sequence based on their relevance to the current processing step. This mechanism lets the model focus on the most relevant information while filtering out less important details. The self-attention mechanism helps the model capture local and global dependencies in the input data, leading to more accurate and expressive representations. Self-attention mechanism allows the model to weigh the importance of different tokens in the input sequence when making predictions.

The self-attention mechanism computes an attention score for each token in the input sequence with respect to every other token. Given an input sequence $x = (x_1, x_2, \ldots, x_n)$, the attention scores are computed as follows:

$$
e_{ij} = \frac{\exp(x_i W_Q (x_j W_K)^\top)}{\sum_{k=1}^{n} \exp(x_i W_Q (x_k W_K)^\top)}
$$

where W_Q and W_K are learnable weight matrices representing the query and key, respectively. The output of the self-attention layer is a weighted sum of the input embeddings:

$$z_i = \sum_{j=1}^{n} e_{ij}(x_j W_V)$$

where W_V is another learnable weight matrix representing the value.

2.1.3. Multi-Head Attention

Multi-head attention is an extension of the self-attention mechanism, designed to enable the model to learn multiple attention patterns simultaneously. Intuitively, each "head" in the multi-head attention can be considered an individual attention mechanism that focuses on a different aspect of the input data. By combining the outputs of multiple heads, the model can capture a more diverse and more prosperous set of relationships between the input elements, ultimately leading to more robust and expressive representations.

The multi-head attention layer computes h different self-attention outputs and concatenates them:

$$\text{MultiHead}(x) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O$$

where each head is computed as:

$$\text{head}_i = \text{SelfAttention}(xW_i^Q, xW_i^K, xW_i^V)$$

and W_O is a learnable output weight matrix.

2.1.4. Position-wise Feed-Forward Networks

The position-wise feed-forward networks (FFNs) are responsible for injecting non-linearity into the model and processing the output of the attention mechanisms. The FFNs act independently on each position in the input sequence, learning nonlinear transformations that contribute to the overall expressivity of the model. These networks enable the Transformer to capture certain abstract features in the input.

These networks originally consisted of two linear layers with a ReLU activation function in between:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2$$

where W_1, W_2, b_1, and b_2 are learnable parameters. ReLU is often replaced with other nonlinear functions in subsequent works.

2.1.5. LayerNorm

Layer normalization (LayerNorm) (Ba et al. 2016) is an essential component of the Transformer architecture, responsible for stabilizing the training process and improving generalization. It works by normalizing the activations of the previous layer across the feature dimension. The choice of normalization technique can significantly impact the efficiency and performance of the model, as different approaches may offer varying levels of stability and generalization capabilities. Pre-LayerNorm, Post-LayerNorm, and NormFormer are the three normalization techniques relevant to this thesis.

LayerNorm is defined as follows:

$$\text{LayerNorm}(x) = \frac{x - E[x]}{\sqrt{Var[x] + \varepsilon}} \cdot \gamma + \beta,$$

where γ and β are trainable parameters, and ε is a small constant.

Post-LayerNorm. In the post-LayerNorm variation (the default version discussed before and presented in Figure 1), layer normalization is applied after the multi-head attention and position-wise feed-forward networks:

$$\begin{aligned} x' &= \text{MultiHead}(x) + x \\ o &= \text{LayerNorm}(x') \\ o' &= \text{FFN}(o) + o \\ z &= \text{LayerNorm}(o') \end{aligned}$$

Pre-LayerNorm. In the pre-LayerNorm variation (Xiong et al. 2020a), layer normalization is applied before the multi-head attention and position-wise feed-forward networks:

$$\begin{aligned} x' &= \text{LayerNorm}(x) \\ o &= \text{MultiHead}(x') + x \\ o' &= \text{LayerNorm}(o) \\ z &= \text{FFN}(o') + o \end{aligned}$$

NormFormer. In NormFormer (Shleifer et al. 2021a), the Normalization block is placed before the residuals and FeedForward layer (as in pre-LayerNorm), while residual and Self- Attention layers are also normalized.

Specifically, the output of each attention head is scaled via learned scalar coefficients γ_i:

$$\text{HeadScaleMHA}(Q,K,V) = \text{Concat}(\gamma_1 \text{head}_1, \ldots, \gamma_n \text{head}_n) W^O$$

where γ are learnable parameters initialized to 1.

Then, an additional LayerNorm block is applied to the result of HeadScaleMHA and another one after the fully connected layer.

The Transformer architecture is the basis for our specific models in this thesis. In the next section (Section 2.2), we will provide a background on these models.

2.2. Model Types

This section provides an overview of the various types of models we investigate in this thesis, all of which are built upon the Transformer architecture.

First, we will provide an overview of language modeling, including Masked Language Models (MLMs) and Causal Language Models (CLMs). Then, we will describe the specific pretrained language models used in this study, including mBERT, XLM-RoBERTa, XGLM, and Machine Translation (MT) models.

2.2.1. Language Modeling

In natural language processing, language modeling is a fundamental task to predict the probability distribution of words or tokens in a given context. It is the basis for many downstream applications, such as writing assistants, summarization, and sentiment analysis. It is also the approach behind popular commercial systems such as OpenAI's ChatGPT [1].

Two popular approaches to the task of language modeling are Masked Language Models (MLMs) and Causal Language Models (CLMs). Both approaches involve training the model to predict words or tokens in a given sequence, but they differ in their prediction strategies.

In this thesis, we will analyze pretrained MLM-based and CLM-based multilingual language models (see Section 2.2.2) to investigate the multilingual structure they learn.

Masked Language Models (MLMs). Masked Language Models (MLMs) (Devlin et al. 2019) are language modeling techniques that focus on predicting the masked words in a given context. By randomly masking a subset of tokens in the input and training the model to predict the original words, MLMs can effectively learn contextual representations of words, accounting for both left and right contexts. This bidirectional context learning allows MLMs to capture a more comprehensive understanding of the input text, improving their ability to generate accurate and coherent predictions. MLMs are usually used for natural language understanding tasks such as question answering, sentiment analysis, and text classification.

The objective function for MLMs can be formulated as:

$$\mathscr{L}_{MLM} = -\sum_{i \in \mathscr{M}} \log P_\theta(t_i | t_{\mathscr{U}})$$

[1] `https://openai.com/blog/chatgpt`

where $\mathscr{M}$ denotes the set of masked positions, t_i is the token at position i, $t_{\mathscr{U}}$ represents the unmasked tokens, and P_θ is the model's probability distribution parameterized by θ. MLMs most often consist of a stack of encoder Transformer blocks, omitting decoder parts (apart from the softmax layer needed for making predictions).

Causal Language Models (CLMs). Causal Language Models (CLMs) (Radford et al. 2018) are another type of language modeling technique that predicts the next word in a sequence, given its preceding context. Unlike MLMs, CLMs rely on a unidirectional context, processing the input text from left to right. This approach ensures that the model learns causal relationships between tokens, particularly useful for text generation tasks, where the generated words must follow a coherent and logical order.

The objective function for CLMs can be formulated as:

$$\mathscr{L}_{CLM} = -\sum_{i=1}^{n} \log P_\theta(t_i|t_{<i})$$

where n is the length of the sequence, t_i is the token at position i, $t_{<i}$ represents the tokens preceding t_i, and P_θ is the model's probability distribution parameterized by θ. CLMs most often only consist of a stack of decoder Transformer blocks, omitting encoder parts.

2.2.2. Pretrained Language Models (PLMs)

Pretrained Language Models (PLMs) leverage large-scale unsupervised training on diverse text corpora to learn general linguistic knowledge, which then can be either fine-tuned for specific tasks (Devlin et al. 2019) or prompted without any fineturning (Brown et al. 2020b).

This transfer learning approach enables PLMs to achieve state-of-the-art performance across a wide range of NLP tasks with small-to-none task-specific training data.

This study will focus on three widely-used pretrained multilingual language models: mBERT, XLM-RoBERTa, and XGLM. These models have been pretrained on massive multilingual corpora and have demonstrated impressive performance in various natural language understanding tasks.

mBERT. mBERT (Devlin et al. 2019), or Multilingual BERT, is a pretrained language model based on the BERT architecture, designed to handle multiple languages simultaneously. It was trained on a large-scale multilingual corpus consisting of text from 104 languages, allowing it to learn a shared representation space for different languages. mBERT has been widely used as a robust baseline model for various cross-lingual transfer learning tasks and has served as the foundation for several other multilingual models.

The mBERT is based on the masked language modeling and next-sentence prediction (NSP) objectives. NSP is a binary classification task that aims to predict

whether two sentences are consecutive in the original text. There is a unique CLS token prepended to the input sequence, and the model is trained to predict the probability of the following sentence given the CLS token and the first sentence.

Our contributions in Section 4.1 will be based on the mBERT model.

XLM-RoBERTa. XLM-R (Conneau et al. 2020a), or Cross-lingual Language Model RoBERTa, is an extension of the RoBERTa model (Liu et al. 2019). XLM-R improves upon mBERT by incorporating additional training and architectural improvements, resulting in more substantial cross-lingual capabilities. XLM-R uses a token-based batching approach (it groups text by tokens, ignoring sentence boundaries) and removes the NSP objective.

The XLM-Roberta model is the basis for the analysis from Sections 3.2 and 3.3.

XGLM. XGLM, or Cross-lingual Generative Language Model (Lin et al. 2022), is a GPT-style multilingual language model.

Unlike mBERT and XLM-RoBERTa, which are masked language models, XGLM follows an autoregressive training strategy and CLM objective. By learning to predict each token in a sequence conditioned on the preceding tokens, XGLM aims to capture the generative nature of language. This characteristic makes XGLM particularly suitable for various natural language generation tasks across multiple languages.

XGLM in various sizes is studied in Section 3.3.

2.2.3. Machine Translation (MT)

Machine Translation is the task of translating text from one language to another. It is the original task for which the Transformer architecture was proposed, as it demonstrated state-of-the-art performance in neural machine translation.

Definition. Machine translation models are typically trained using a sequence-to-sequence (seq2seq) approach, where the model consists of an encoder and a decoder. The encoder processes the input text in the source language and generates a contextualized representation, while the decoder generates the translation in the target language conditioned on the encoder's output. The objective function for MT models can be formulated as:

$$\mathscr{L}_{MT} = -\sum_{i=1}^{n} \log P_\theta(t_i^{tgt} | t_{<i}^{tgt}, t_{1:m}^{src})$$

where t_i^{tgt} is the token at position i in the target language sequence, $t_{<i}^{tgt}$ represents the target language tokens preceding t_i^{tgt}, $t_{1:m}^{src}$ denotes the source language sequence, and P_θ is the model's probability distribution parameterized by θ. Machine translation models learn to map the representations of the source language to the target language and can be trained across data that spins multiple domains, thus playing a crucial role in understanding the underlying structure and properties of multilingual and multi-domain models.

In this work, we train and analyze multi-domain machine translation models.

Multidomain MT. Multi-domain Machine Translation models are designed to handle translations across multiple domains, such as news articles, technical documents, and conversational texts.

The training process typically involves pretraining on a large, multi-domain corpus.

In Section 3.4, we will investigate how multi-domain machine translation models represented domains in their representation spaces, i.e., we will study the shape of the multi-domain structure that spins the layers of these models.

2.3. Representations Comparison Indexes

In this section, we introduce various representational similarity indexes, which are used to compare the internal representations extracted from multilingual and multi-domain deep learning models. These indexes enable us to quantitatively measure the similarity between representations of the multilingual and multi-domain deep learning models, providing insights into the structure and properties of the representational spaces. The indexes covered in this section include Canonical Correlation Analysis (CCA), Projection Weighted CCA (PWCCA), Singular Vector CCA (SVCCA), and Centered Kernel Alignment (CKA). Representational similarity analysis is the fundamental methodology used to study cross-lingual representational similarity in Section 3.2.

Similarity indexes presented in this section are general-purpose and have been used to compare representations between arbitrary neural networks and between neural networks and biological brains (H.-T. Wang et al. 2020). The first half of this thesis uses these techniques to investigate the emerging cross-lingual structure, and in Section 3.1.2, we introduce a new similarity index that is explicitly aimed at the analysis of representations from multilingual deep learning models.

2.3.1. Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) (Hotelling 1936) is a statistical technique used to analyze the relationship between two sets of multivariate variables.

Given two sets of variables, X and Y, CCA aims to find the linear combinations of X and Y with maximum correlation. Mathematically, CCA seeks to find the weight vectors w_X and w_Y that maximize the following correlation:

$$\rho = \max_{w_X, w_Y} \frac{w_X^T X Y^T w_Y}{\sqrt{w_X^T X X^T w_X} \sqrt{w_Y^T Y Y^T w_Y}}$$

In the context of our study, X and Y correspond to the internal representations of the multilingual models for different languages or domains.

By using CCA-like methods, we can evaluate the similarity of these representations by examining the canonical correlations between them. CCA is a foundation

for enhanced similarity indexes such as PWCCA and SVCCA, which we use to measure representations and will discuss in the following sections.

2.3.2. Projection Weighted CCA (PWCCA)

Projection Weighted CCA (PWCCA) (Morcos et al. 2018) is an extension of CCA that incorporates a weighting factor to account for the variance of different components.

PWCCA's incorporation of a weighting factor based on the variance of different components helps to balance the influence of each component in the similarity calculation, ensuring that the overall measure is more robust to noise and slight variations, thus providing a more reliable assessment of similarity.

This weighting factor is derived from the eigenvalues of the covariance matrices, resulting in a similarity index that is more robust to noise and slight variations. PWCCA is defined as:

$$PWCCA(X,Y) = \frac{\sum_{i=1}^{d} \rho_i \lambda_i}{\sum_{i=1}^{d} \lambda_i}$$

where superscript d is the dimension of the representation space, ρ_i is the i-th canonical correlation, and λ_i is the i-th eigenvalue of the covariance matrix.

2.3.3. Singular Vector CCA (SVCCA)

Singular Vector CCA (SVCCA) (Raghu et al. 2017) is another extension of CCA that combines Singular Value Decomposition (SVD) with CCA.

SVCCA helps by reducing the impact of less significant, noisy dimensions in the representations, which allows for a more explicit focus on the core structure and properties shared, ultimately leading to a more accurate similarity measure.

SVCCA first applies SVD to the internal representations and retains only the top k singular vectors. The reduced representations are then fed into the CCA to calculate the canonical correlations. SVCCA aims to eliminate noisy dimensions and focus on the most significant components of the representations. The process can be summarized as:

1. Perform SVD on X and Y:

$$X = U_X S_X V_X^T, \quad Y = U_Y S_Y V_Y^T$$

2. Retain the top k singular vectors of U_X and U_Y:

$$\tilde{U}_X = U_X(:, 1:k), \quad \tilde{U}_Y = U_Y(:, 1:k)$$

3. Apply CCA to the reduced representations $\tilde{U}_X$ and $\tilde{U}_Y$.

2.3.4. Centered Kernel Alignment (CKA)

Linear Centered Kernel Alignment (we use a linear version unless stated otherwise) quantifies the similarity between two sets of representations X and Y using a simple and computationally efficient approach. The similarity metric is computed as follows:

$$CKA(X,Y) = \frac{\|X^T Y\|_F^2}{\|X^T X\|_F \|Y^T Y\|_F}$$

This formula calculates the squared Frobenius norm of the cross-covariance matrix $X^T Y$ normalized by the product of the Frobenius norms of the covariance matrices $X^T X$ and $Y^T Y$. The result is a normalized score that reflects the degree of similarity between the two sets of representations, with a higher score indicating greater similarity.

SVCCA reduces dimensionality and noise impact by applying Singular Value Decomposition, which emphasizes the most significant features of the data. Similarly, PWCCA counters noise by weighting canonical correlations according to the variance they explain, thus prioritizing the most informative directions. Both methods aim to refine the correlation analysis beyond the linear scope of Linear CKA, enhancing the interpretability of complex data relationships. Linear CKA, similarly to CCA, might suffer from the problems PWCCA and SVCCA aim to solve, but it is an explainable, simple, and efficient way to compare representations linearly.

3. KEY RESULTS

Multilingual language models, such as the Transformer-based mBERT (Devlin et al. 2019)[1] and XLM-R (Conneau et al. 2020a), are designed to handle multiple languages. These models undergo self-supervised training, which differs from machine translation as it involves unsupervised data from multiple languages without direct cross-lingual signals.

Multi-domain Transformers, in contrast, are trained on diverse datasets that span multiple domains or are inherently large and heterogeneous, covering a broad spectrum of topics and styles.

This chapter explores the mechanisms of multilingual and multi-domain Transformers and presents four key findings about their processes. Two findings focus on the existence of these processes, while the other two address their universality:

- **Result I: Multilingual Abstraction** - We show that Transformers start by processing sentences focusing on language-specific features in the initial layers. Gradually, they shift to an abstract representation that transcends linguistic specifics and emphasizes the underlying meaning (Section 3.2).
- **Result II: Universality of Multilingual Abstraction** - This shift from language-specific to abstract representation occurs consistently across various training datasets, architectural choices, and model sizes, indicating universality across multilingual Transformer models (Section 3.3).
- **Result III: Multi-domain Specialization** - Unlike the abstraction in multilingual models, multi-domain Transformers continuously identify and preserve domain-specific features in sentence representations across all layers (Section 3.4).
- **Result IV: Universality of Multi-Domain Specialization** - This finding suggests that domain specialization is also a repeating pattern observed across Transformer models (Section 3.5).

The chapter is structured as follows: Section 3.1 introduces a new technique for comparing representations to support claims about multilingual abstraction, and Sections 3.2, 3.3, 3.4, and 3.5 present the arguments for Claims I-IV respectively.

3.1. Representations Comparison with ANC

This section discusses *representations comparison*, a technique for understanding the internal workings of neural networks (see Section 3.1.1). We focus on applying this method to the analysis of multilingual models. Additionally, we introduce the "Averaged Neuronwise Correlation (ANC)," a new similarity measure specifically created for analyzing multilingual models, which is explained in Section 3.1.2. In the following sections, we will use representations comparison and ANC to explore the concept of 'multilingual abstraction' and its universality.

[1] https://github.com/google-research/bert/blob/master/multilingual.md

3.1.1. Representations Comparison in Multilingual Models

Introduction. Representations Comparison methodically unveils structural characteristics within multilingual language models by examining language representation similarities across various layers. The following steps detail this method's application to multilingual models:

- **Step 1: Assemble Parallel Texts:** Gather a parallel corpus comprising matching sentences in two languages.
- **Step 2: Extract Layer-Wise Token Embeddings:** For each sentence in the corpus, use a multilingual model like mBERT to derive token embeddings from every layer.
- **Step 3: Formulate Sentence Vectors:** For each sentence at every layer, synthesize a singular sentence vector by averaging the token vectors or taking a vector for the CLS token to represent a sentence (CLS-pooling).
- **Step 4: Compute Similarity Across Layers:** Use the similarity index to calculate and record the similarity between sentence vectors in different languages for each layer. This assesses the degree of alignment between languages within each layer.
- **Step 5: Analyze Similarity Scores:** Evaluate a trend in the similarity data across layers to understand the evolution of language representations within the model.

Figure 2 visually illustrates this process.

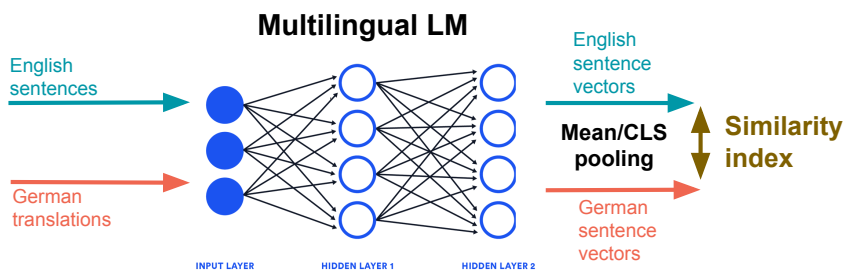

Figure 2: Schematic of the cross-lingual representations comparison process.

Representations Comparison employs similarity indexes, as mentioned in Step 4, which are statistical methods for measuring vector similarities. Widely used options include PWCCA (Morcos et al. 2018), SVCCA (Raghu et al. 2017), and CKA (Kornblith et al. 2019). These methods are rooted in Canonical Correlation Analysis (CCA), designed to identify maximally correlated linear combinations of variables from different sets. Each method uniquely enhances CCA: PWCCA utilizes internal CCA coefficients for refinement, SVCCA applies singular vector decomposition, and CKA uses kernel methods. In multilingual representation analysis, these indexes are crucial for quantifying language alignment within neural networks. Section 2.3 provides detailed mathematical definitions for these indexes.

While CCA-based methods are highly effective for general representation comparison, the unique context of multilingual models, where two vector sets originate from the same model, presents a distinct scenario. The following section introduces the "Averaged Neuronwise Correlation" as a new similarity index designed to leverage this aspect in multilingual cases.

3.1.2. Averaged Neuronwise Correlation

Following the framework outlined in Section 3.1.1 for comparing multilingual representations, we introduce the Averaged Neuronwise Correlation (ANC). This index is designed explicitly for cross-lingual analysis and is grounded in strong assumptions tailored for multilingual language models. Due to its simplicity, the ANC enables stronger claims about these models and ensures straightforward interpretability.

Method: Assumption. We base ANC on the assumption of one-to-one neuron alignment across languages in multilingual representations, proposing that *contents and positions of neurons should be similar across different languages*. We find this assumption reasonable for several reasons:

- **Identical Sentences Scenario:** In a perfectly cross-lingual model, two identical sentences in different languages should yield identical representations. Since representations consist of neurons, identical representations imply identical neurons.
- **Cross-Lingual Transfer Learning:** Our assumption is essentially what enables zero-shot cross-lingual transfer learning in multilingual models. A linear prediction head, typically trained in English, could also effectively work with other languages because key neurons it relies on behave similarly across languages.
- **Enhanced interpretability:** Basing similarity index on neuron-to-neuron similarity allows to break down the resulting score into individual neuron contributions easily. This approach enables the identification of neurons that contribute most or least to similarity, offering insights into their language-specific or universal properties. However, the usefulness of viewing neurons as independent units for interpretability is limited because interpretable features in networks are mainly distributed over neurons, and each neuron plays a role in representing multiple features (Elhage et al. 2022). Nevertheless, we can track the difference in representations to a set of uncorrelated neurons, which can serve as a base for further interpretability investigations.

The assumption of one-to-one neuron correspondence is a strong conservative assumption. It might miss the cases where the representations are similar in non-trivial ways. However, the non-trivial cross-lingual alignment inside the model requires more effort to represent and limits the models' cross-lingual potential, which is otherwise unlocked with a straightforward alignment. ANC, on

the other hand, allows us to make strong statements about representations (such as the presence of one-to-one neuron alignment between languages) if it results in a high degree of similarity. As the following sections will show, despite ANC being conservative, it consistently reveals high degrees of similarity in multilingual models.

Method: Definition. Constructing a similarity index assuming one-to-one neuron correspondance is straightforward. We calculate individual correlations between pairs neurons at same positions in two languages and then compute an average score across all neurons. Therefore, we define the Average Neuron-Wise Correlation (ANC) as follows.

Let L_1 and L_2 be two sets of vector representations of dimension NxD, where N is the number of data points, and D is the number of dimensions (neurons). Let X and Y be the centered (by neurons) representations, defined as:

$$X := L_1 - mean(L_1)$$
$$Y := L_2 - mean(L_2)$$

Thus, the columns of X and Y correspond to neurons across two datasets, and each neuron is defined as a set of activation values it takes over the dataset.

Let $\vec{z_x^i}$ and $\vec{z_y^i}$ be the neurons from position i from X and Y correspondingly. The neuron is thus The Pearson correlation *corr* between the two neurons is then defined as:

$$corr(\vec{z_x}, \vec{z_y}) = \frac{\langle \vec{z_x^i}, \vec{z_y^i} \rangle}{\|\vec{z_x^i}\| \|\vec{z_y^i}\|} \tag{3.1}$$

Consequently, the ANC similarity between two sets of representations L_1 and L_2 is defined as:

$$ANC(X,Y) = \frac{\sum_{i=1}^{N} abs(corr(\vec{z_x^i}, \vec{z_y^i}))}{N} \tag{3.2}$$

In our work, we obtain two sets of representations from the network layers by passing a parallel bilingual corpus through the model.

Limitations and discussion. Firstly, neural networks have permutational symmetries. By reordering neurons, it is possible to get an identical network with different weights. Given a random initialization of weights, the training process cannot lead to identical networks with permuted weights, even for the same training set. For such cases, the ANC score can be quite low. For example, the question might be, why should a score of 0.3 indicate significant discrepancies between sentence representations produced by two different models if the permuted but identical models achieve a score of 0.9?

ANC will not be helpful when comparing representations from two different networks. However, it is not the intended use case. In our work, we are deliberately interested in comparing representations of sentences in other languages *that come from a single multilingual model*. Thus, in our setup, the neurons originate from a single network, and the assumption of alignment between neurons is reasonable. If the neurons were derived from layers of two distinct networks, general-purpose indexes like CKA or CCA-based methods would be more suitable.

Secondly, large networks may contain identical components that are language-specific but capture the same abstract concepts. Consequently, the ACN score can be low even if the layer captures the same concepts.

However, if the network captures the same abstract concepts in two different ways for two different languages, then it means that it implies that it specializes in each language independently (specialization) as opposed to using shared multilingual features (abstraction) and avoiding duplication of concepts in representations. We are interested in capturing exactly this difference.

Third, there is the question of sensitivity to noise. If the model is not minimal, network components might produce irrelevant signals that are ignored by upper layers, while the active and controlled core pathway might not include all the neurons. Consequently, the ACN score can be low even if the useful part of the layer behaves identically.

Part of the reason similarities indexes like PWCCA and SVCCA were introduced is to reduce the sensitivity of CCA to outliers. However, many core pathways can be potentially useful for different tasks, and without knowing a specific task for a multilingual language model, it is hard to know what neurons to value more. Since in our work, we are interested in the base model (not tuned) we expect most neurons to be potentially valuable. Alternatives such as trimmed mean or median are the good modifications that we leave for future work, while, as our experiments show, even the basic version of ANC is enough to make desired claims about multilingual representations coming from a single multilingual model.

3.2. Result I: Multilingual Abstraction

This section examines how mBERT and XLM-R, prominent multilingual language models, exemplify 'multilingual abstraction.' This process begins with models assigning more language-specific representations to the sentences in early layers. As processing advances to deeper layers, the model transitions to a more language-neutral representation, centering on shared meaning across languages.

We present four independent lines of evidence to validate the multilingual abstraction in mBERT:

- **Evidence 1 (Representations Comparison):** Cross-language representation comparisons revealed distinct differences between the initial layers,

signifying language-specific traits. However, deeper layers showed converging representations, indicating a shift towards shared semantic understanding (Section 3.2.1).

- **Evidence 2 (Unsupervised Visualization):** t-SNE visualizations at each layer initially display clear language-based clusters that merge into groups defined by semantic rather than linguistic attributes in the later layers (Section 3.2.2).
- **Evidence 3 (Suggestive Evidence from Related Works):** Insights from related research on probing tasks, zero-shot cross-lingual transfer, and representational alignment support the multilingual abstraction hypothesis. (Section 3.2.3).
- **Evidence 4 (Monolingual Abstraction):** A parallel abstraction pattern in monolingual BERT models, observed with paraphrases, suggests that multilingual models similarly abstract language details in favor of overarching semantic content (Section 3.2.4).

This evidence collectively leads us to conclude that multilingual abstraction aptly characterizes sentence processing in mBERT. This section focuses on mBERT, and Section 3.3 explores the universality of this phenomenon. We further defend the evidence in Section 4.

3.2.1. Evidence 1: Representations Comparison

Applying *representations comparison* with the Averaged Neuronwise Correlation (ANC) offers compelling evidence for the multilingual abstraction process in both mBERT and XLM-R models. ANC scores provide insights into how these models handle language-specific representations across various layers.

Setup. This experiment compared English representations with Estonian, Latvian, Lithuanian, French, and Polish representations. Estonian, Latvian, and Lithuanian represent the Baltic language family, Polish represents Slavic languages, and French is most linguistically similar to English.

We do a setup similar to the one of Singh et al. (2019a) and use the mBERT-cased model and a multi-parallel dataset (en-et, en-lt, en-lv, en-pl, en-fr; 10k examples for each pair). The parallel corpus is composed of Singh et al. (2019a)'s extension of the XNLI (Cross-lingual Natural Language Inference) dataset (Conneau et al. 2018). We chose Estonian, Latvian, and Lithuanian as our case study and used Polish and French to see how results for Baltic languages compare to high-resource Romance and Slavic languages.

We embed the source and target sentences with mBERT, average over token embeddings to obtain sentence representations (mean-pooling) from each layer and repeat for each language pair. Next, we compare two parallel sets of sentence representations using the ANC similarity index.

Results. Figure 3 presents the ANC results. ANC scores were lower in the early layers, indicating that representations were dissimilar between language

pairs. It aligns with the expectation that the initial processing stages will focus more on unique language characteristics. As the models progressed to deeper layers, ANC scores increased, reaching approximately 0.7. This score elevation suggests a significant alignment and correlation of neurons across languages, implying a shift towards a more unified semantic representation. The pattern observed in later layers indicates the models abstracting from language-specific details and converging towards a shared semantic understanding across different languages.

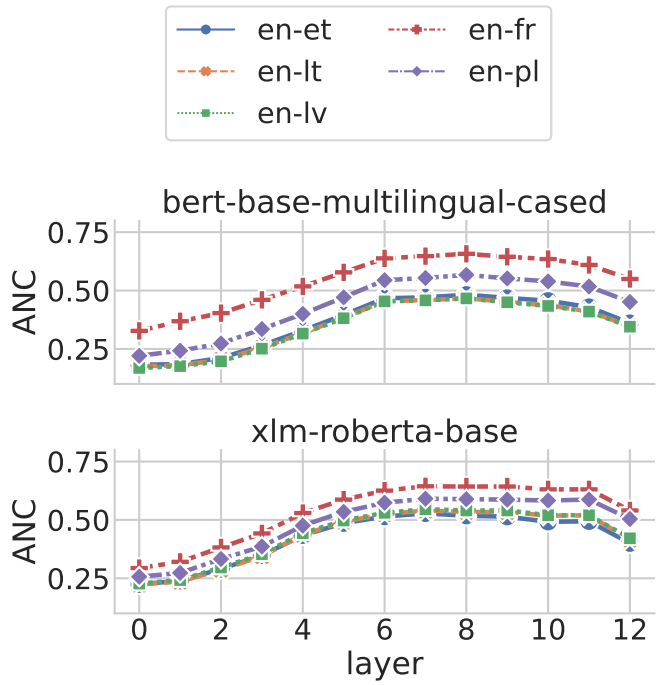

Figure 3: ANC results for the mBERT and XLM-R models.

To confirm whether mBERT treats different languages distinctively in its initial layers, we used additional similarity indexes: PWCCA, SVCCA, and CKA. ANC presumes a one-to-one correspondence of neurons across languages, potentially overlooking cases where the network develops language-specific neurons for similar concepts. In contrast, PWCCA, SVCCA, and CKA are more flexible in capturing similarities, even when neurons are not aligned one-to-one.

Figure 4 presents the results of this analysis.

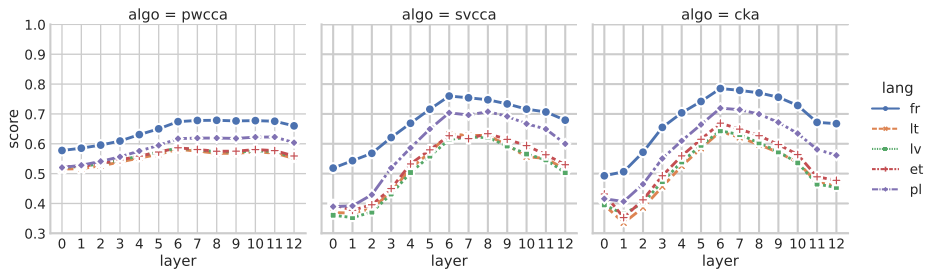

Figure 4: Multilingual abstraction in mBERT revealed by SVCCA, PWCCA, and CCA.

The findings in the figure suggest that, indeed, mBERT's initial layer representations are significantly language-specific. We also estimated a calibration baseline/lower bound (not presented in Figures): if we permute sentences in one of the sets of representations, we get close to zero ANC score at all layers for both CCA-likes and ANC.

Our analysis in this subsection substantiates the hypothesis that mBERT and XLM-R models undergo a process of multilingual abstraction, progressively shifting from language-specific to more abstract semantic representations as information flows through network layers.

3.2.2. Evidence 2: Unsupervised Visualization

Expanding on the representation comparison findings from Section 3.2.1, this subsection explores the multilingual abstraction in mBERT through unsupervised t-SNE visualization.

Setup. In this experiment, we used the same multi-parallel dataset as in section 3.2.1. The multi-parallel nature of the dataset ensures that the sentences are not grouped by the topic/domain of the dataset from which the language is sampled.

We used three high-resource languages (English, German, and French) and two low-resource languages (Swahili and Urdu) to track the evolution of sentence representations across the mBERT layers.

t-SNE algorithm works in an unsupervised way based on the plain sentence representation. Only at the end do we color each data point with the corresponding language to make sense of the resulting clusters that manifest themselves naturally.

Results. Figure 5 provides a visualization of the representational transformation in mBERT. In layers 0-2, the language-specific features dominate the representations that t-SNE. This algorithm does not obtain any language identity with the input groups languages in distinct clusters. However, a critical transformation occurred in the middle layers (5-8). Here, groups of English, German, and French representations merge, indicating that mBERT abstracts away from language-specific features and processes sentences more language-neutrally. This

observation strongly supports the hypothesis of multilingual abstraction, in which language distinctions become less prominent in favor of shared semantic content.

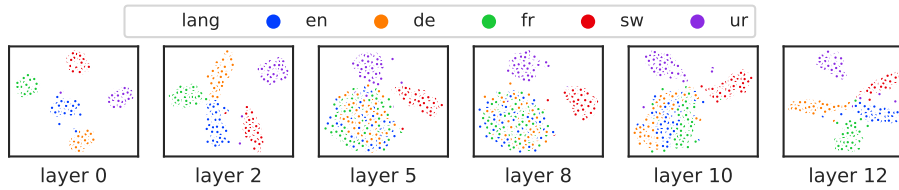

Figure 5: t-SNE visualization of sentence representation transformation in mBERT, highlighting multilingual abstraction.

Interestingly, representations of European languages again show divergence in the final layers, indicating a re-emphasis on language-specific traits. This pattern is likely related to mBERT's task-specific requirements of mBERT for predicting masked tokens in a specific language.

However, the representations of Swahili and Urdu continued to form separate clusters in these layers, suggesting that mBERT retains more language-specific features for these languages. Although this could indicate a limitation in the generality of the multilingual abstraction process, we argue that these languages are outliers and further discuss this in Section 4.2.

t-SNE visualization strongly supports the multilingual abstraction hypothesis. It shows mBERT's capacity to transform sentence representations from language-specific in the early layers to more abstract and language-neutral in the middle layers before reverting to language-specific processing in the final layers. The specialization in the last layers is due to the model needing to generate text in a specific language. This dynamic pattern, combined with the insights from the representation comparison, underscores multilingual abstraction as a critical characteristic of mBERT's processing of multilingual input.

3.2.3. Evidence 3: Suggestive Evidence from Related Works

Zero-shot cross-lingual transfer. The zero-shot cross-lingual transfer ability of mBERT, highlighted in Hu et al. (2020) and Liang et al. (2020), provides indirect yet crucial evidence for multilingual abstraction. This phenomenon, where mBERT fine-tuned on a task in one language, demonstrates proficiency in performing the same task in other languages, invites an analysis of its underlying mechanics.

For such cross-lingual applicability, a key factor is the existence of a language-neutral component within the mBERT representations. When the model fine-tunes for a task in one language, its prediction head adapts to specific neurons that resemble language-neutral behavior. This adaptation allows the same linear combinations of neurons calibrated for one language to be effective for tasks in other languages. The efficacy of this approach suggests that, as mBERT pro-

cesses linguistic input, it abstracts language-specific features into more generalized, language-neutral representations. These representations form the basis upon which the prediction head operates, enabling a cross-lingual functionality.

This indirect evidence from the behavior analysis in related works, particularly the model's capability for zero-shot transfer, supports the hypothesis of multilingual abstraction. It implies that mBERT processes language-specific information and transforms it into more abstract language-neutral forms, crucial for cross-lingual transfer.

Explicit alignment of representations. From an interpretability perspective, aligning vector spaces across languages is not just a technical challenge, but also a method to quantify the degree of similarity between representations. Various approaches to alignment can provide insights into whether multilingual representations are inherently well structured or whether further intervention is needed to obtain effective cross-lingual transfer.

For example, related works aligned representations by learning linear transformations (Cao et al. 2020; Conneau et al. 2020b; Vázquez et al. 2021), suggesting that if good performance is achieved using these methods, the underlying representations are already linearly similar across languages. Thus, if linear transformations suffice, it may indicate a high degree of inherent multilingual abstraction within the representations. Artetxe et al. (2020) train a transformer LM on one language, and just by retraining an embedding matrix, transfer it to a new language. This also suggests a degree of correspondence in cross-lingual structures between languages. Kornblith et al. (2019) provide a deeper discussion of linear alingment methods in the context of representational similarity indexes and Hämmerl et al. (2024) survey allingment methods in relation to language neutrality in LMs.

3.2.4. Evidence 4: Paraphrase Analysis

Following the analyses in Sections 3.2.1 and 3.2.2, this subsection examines the processing of paraphrases in both the monolingual and multilingual BERT models, further substantiating the multilingual abstraction hypothesis. We used a parallel corpus consisting of English sentences, their German translations, and human-generated English paraphrases. This setup allowed us to compare sentences with the same semantic content but different linguistic forms in the BERT models. We opted for the SVCCA similarity index instead of ANC due to the latter's introduction after the analysis.

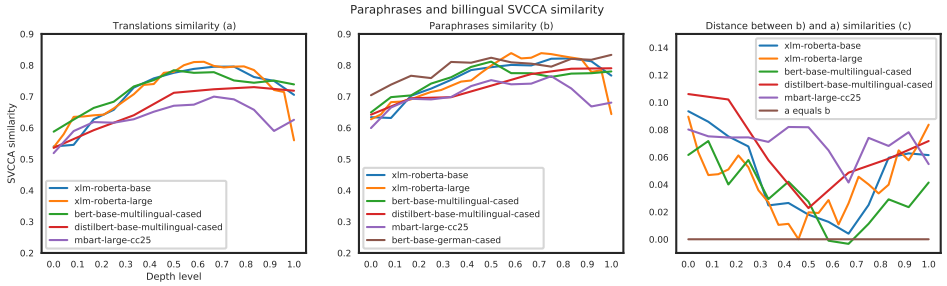

Figure 6: Comparison of sentence representations for translations and paraphrases in BERT models, illustrating the abstraction process.

Figure 6 presents the results of this analysis, showing how both translations and paraphrases underwent a similar abstraction process, particularly in the middle layers of the networks. The critical observation here is that around these middle layers, sentences in different languages (translations) and their paraphrases in English exhibit similar levels of representational similarity. This pattern indicates that, regardless of the language, BERT abstracts sentence representations to a point where the underlying semantic content takes precedence over the linguistic form.

This abstraction pattern is not exclusive to the multilingual models. Monolingual German BERT models also exhibited a similar trend when comparing paraphrases, suggesting that sentence abstraction is an inherent property of BERT-like language models. Therefore, the observed interlinguality in multilingual models can be considered a byproduct of this fundamental sentence abstraction process.

In summary, the parallel abstraction patterns observed in the processing of multilingual sentences and monolingual paraphrases within BERT models provide compelling support for the multilingual abstraction hypothesis. This evidence and the distinct yet complementary insights from representation comparison, unsupervised t-SNE visualization, and behavior analysis in related works collectively reinforce our understanding of mBERT's processing dynamics of mBERT. It underscores that mBERT transcends language-specific representations through its layered architecture, gradually moving towards more abstract, language-neutral processing. This transition is not confined to multilingual contexts and mirrors the model's handling of paraphrases within a single language. As we conclude from this exploration of Claim I, the cumulative evidence paints a coherent picture of multilingual abstraction as a fundamental mechanism in mBERT, influencing how it processes and understands multilingual input.

In this experiment, we expanded our model selection beyond XLM-R and mBERT to include other models from the BERT family. This decision lays the groundwork for our future exploration of multilingual abstraction universality, which we investigate in Section 3.3.

3.3. Result II: Universality of Multilingual Abstraction

In Section 3.2, we explored multilingual abstraction within the mBERT and XLM-R models and identified it as a critical phenomenon in language processing. The occurrence of this phenomenon in both models, despite their architectural differences, suggests a broader universality. In this context, universality refers to the consistent occurrence of a phenomenon across different models and tasks, which is crucial for extending interpretability insights from one model to another.

This chapter delves into how multilingual abstraction is a universal characteristic of transformers, irrespective of their architectural variations, scales, and training objectives. We present three critical pieces of evidence to substantiate this claim.

- **Evidence 1 (Universality Across Architectural Variations):** Despite differences in the training data and slight architectural variations, mBERT and XLM-R exhibited multilingual abstraction. An additional experiment highlights the invariance of this phenomenon to the normalization styles in Transformer architecture (Section 3.3.1).
- **Evidence 2 (Universality Across Scale):** Multilingual abstraction across models of varying sizes suggests its scale invariance (Section 3.3.2).
- **Evidence 3 (Universality Across Training Objectives):** Multilingual abstraction extends beyond masked language models, demonstrating its presence across various training objectives (Section 3.3.3).

This evidence collectively supports the conclusion that multilingual abstraction is a universal feature of multilingual Transformer models. The following subsections elaborate on each piece of evidence and provide detailed justifications for our conclusions regarding the universality of multilingual abstraction.

3.3.1. Evidence 1: Universality Across Architectural Variations

This subsection examines whether architectural variations in Transformer models influence multilingual abstraction. To assess this, we focused on two specific comparisons.

Initially, we considered multilingual abstraction in the mBERT and XLM-R models. Despite the differences in batch composition, objective terms (absence of next token prediction in XLM-R), and pretraining datasets, both models exhibit this phenomenon, as previously shown in Section 3.2.1 and Figure 3. Figure 6 shows that the distilled version of the BERT also follows multilingual abstraction.

Next, we extend our investigation to three XLM-Roberta (Conneau et al. 2020a) language models (base-size versions), each employing different normalization schemas.

Model Details. We trained from scratch the following three XLM-Roberta (Conneau et al. 2020a) language models (`base` size versions).

- ***Post-LN*** (`scale_post`): normalization block is placed *after* the residual connections in the transformer block (part of the original Transformer);
- ***Pre-LN*** (`scale_pre`): normalization block is placed *before* the residuals (this was shown to improve training in Xiong et al. (2020b));
- ***Normformer*** (`scale_normformer`): normalization block is placed *before* the residuals *and* FeedForward, Residual, and Self-Attention layers are also normalized (Shleifer et al. 2021b).

We assessed cross-lingual representations for each model using Averaged Neuronwise Correlation (ANC), as outlined in Section 3.1.

Figure 7 presents the ANC similarity results for these pre-trained models, revealing that for all `scale_post`, `scale_pre`, and `scale_normformer`, ANC shows relatively high cross-lingual performance in all layers.

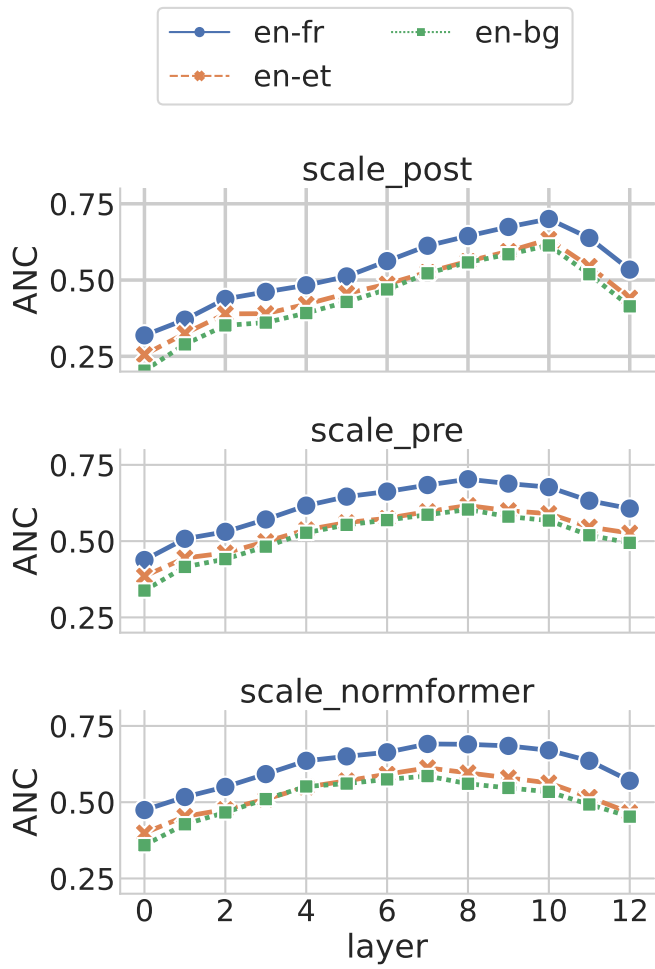

Figure 7: ANC results illustrating multilingual abstraction across different normalization methods in XLM-R.

The consistency in multilingual abstraction across these architectural variations in both the mBERT and XLM-R models, as well as among the different normalization approaches in XLM-R, suggests the resilience of this phenomenon. Although not exhaustive, these findings indicate that multilingual abstraction is a robust feature of Transformer models and is relatively unaffected by specific architectural changes. Next, we explore its universality across scales and training objectives.

3.3.2. Evidence 2: Universality Across Scale

We assessed whether this phenomenon scales in multilingual language models by exploring multilingual abstraction across various model sizes.

Model Details. Our analysis covers a range of XLM-R models that differ in scale from base to XXL. These models varied in the number of parameters and layer counts, as listed in Table 1.

Name	type	#params	l	n	#lgs
xlm-roberta-base	MLM	270M	12	758	100
xlm-roberta-large	MLM	550M	24	1024	100
xlm-roberta-xl	MLM	3.5B	36	2560	100
xlm-roberta-xxl	MLM	10.7B	48	4096	100

Table 1: Model details for XLM-R models we study. *type*: training objective of the model, *#params*: number of parameters, *l*: number of layers, *n*: number of hidden units (neurons at each layer), *#lgs*: number of languages used in pertaining.

Results. Figure 8 shows each model's ANC cross-lingual similarities averaged across English, French, and Bulgarian.

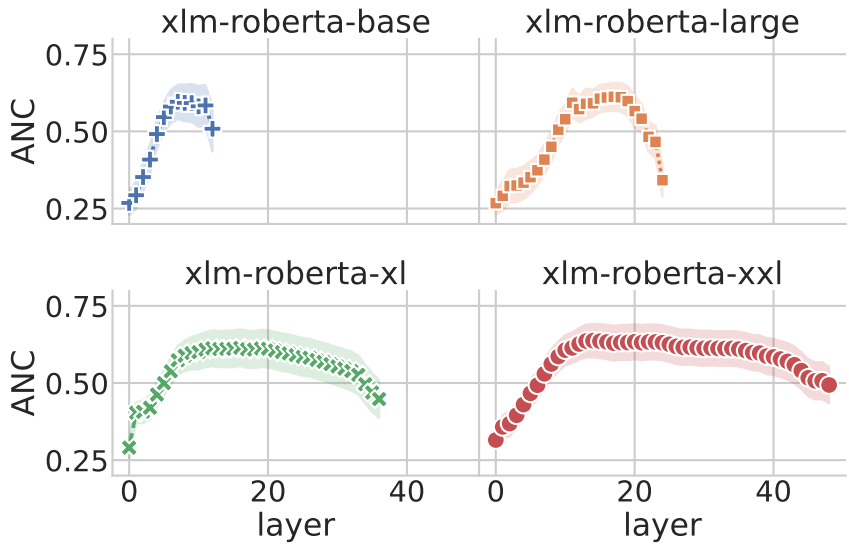

Figure 8: ANC analysis across XLM-R models of varying sizes, indicating scale-independent multilingual abstraction.

The consistent pattern of multilingual abstraction for all sizes highlights its scale independence. These findings confirm that multilingual abstraction is a fundamental feature of multilingual Transformer models unaffected by scale.

With multilingual abstraction established across architectural variations and model scales, we next focus on its universality across different training objectives and explore how varying tasks may influence multilingual abstraction.

3.3.3. Evidence 3: Universality Across Training Objectives

While we showed that Multilingual Abstraction occurs in models of different scales, these were solely masked language models. This evidence indicates that causal GPT-style models (trained to predict the next token) and machine translation models also converge reliably to the Multilingual Abstraction structure.

Kudugunta et al. (2019) made the case for machine translation models, where they employed SVCCA to reveal converging representation patterns in multilingual machine translation models. The following presents an experiment on GPT-style casual language models.

Model Details. We describe the models in Table 2. The Table shows the causal (decoder-only) models.

Name	type	#params	l	n	#lgs
xglm-564M	CLM	564M	24	1024	30
xglm-1.7B	CLM	1.7B	24	2048	30
xglm-2.9B	CLM	2.9B	48	2048	30
xglm-4.5B	CLM	4.5B	48	4096	134
xglm-7.5B	CLM	7.5B	32	4096	30

Table 2: Model details for XGLM models we study. *type*: training objective of the model, *#params*: number of parameters, *l*: number of layers, *n*: number of hidden units (neurons at each layer), *#lgs*: number of languages used in pertaining.

Results. Figure 9 presents the ANC results, demonstrating that *multilingual abstraction* persists across various training objectives, including masked language modeling, machine translation, and causal language modeling.

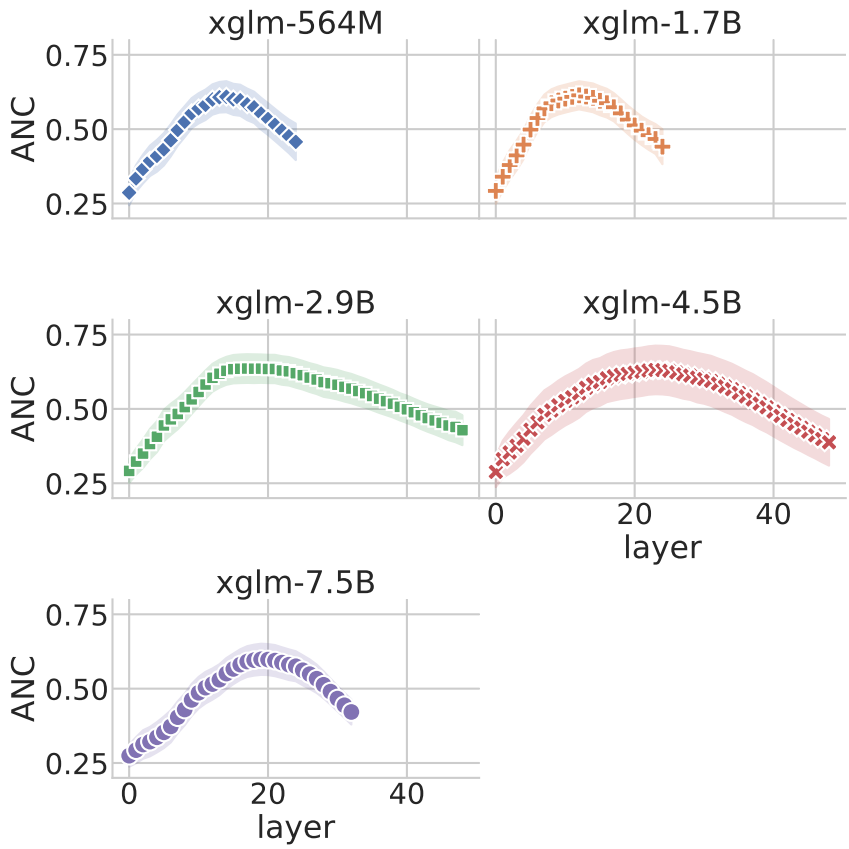

Figure 9: ANC cross-lingual representational similarity for the XGLM CLM-style models of different sizes. All models follow a similar Multilingual Abstraction pattern. We aggregate among en-fr, en-de, en-ru, and en-et pairs and show similarity average and spread.

Multilingual abstraction has been consistently observed across different ar-

chitectural designs, model scales, and training objectives, strongly suggesting its universality in Transformer models. Such universality suggests that multilingual abstraction is an inherent feature of multilingual Transformers integral to their language processing capabilities.

3.4. Result III: Multi-Domain Specialization

Moving from the concept of *multilingual abstraction*, this section delves into *multi-domain specialization* within multi-domain transformers. In contrast to the language-neutral processing typical of multilingual contexts, this phenomenon states that Transformers consistently recognize and preserve domain-specific details in sentence representations throughout all processing layers.

Our analysis uses *Transformer-base* NMT models (Vaswani et al. 2017), covering diverse domains such as parliamentary speeches, medical texts, subtitles, and legal documents, with English-Estonian as the language pair. Each domain's dataset comprises approximately 500k sentences.

In this thesis, by domain, we simply mean text grouped together by distinct interpretable shared characteristics, either stylistic or thematic. For example, we consider a corpus of a medical text a separate domain, while another example might include a dataset of very long sentences (style-based domain). Thus, there may be "domain leakage" in a classical sense where the network identifies the medical domain by the sentence length. Still, in this case, we just say that that is the domain of long sentences, which is an equivalently exciting claim about the inner workings of the converged network.

To demonstrate *multidomain specialization* , we employ three types of evidence:

- **Evidence 1** *(t-SNE Visualization):* Unsupervised visualization with t-SNE reveals multi-domain Specialization (Section 3.4.1).
- **Evidence 2** *(PCA Visualization)*: Unsupervised visualization with PCA reveals multi-domain specialization (Section 3.4.2).
- **Evidence 3** *(k-means Clustering)*: Unsupervised clustering with k-means revealed multi-domain specialization (Section 3.4.3).

In the following subsections, we present the evidence above.

3.4.1. Evidence 1: t-SNE Visualization

In this study, we visualized how sentence representations maintain domain-specific features within a Transformer model. We extracted representations for the development set sentences from all domains and averaged token embeddings to represent each sentence.

By applying a cosine-based t-SNE for dimensionality reduction, we generated a two-dimensional (2D) visualization of these representations. Each sentence is a color-coded post-factum based on its domain, as shown in Figure 10.

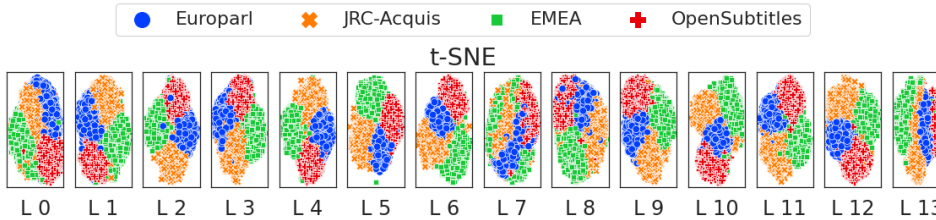

Figure 10: t-SNE visualization of sentence representations across Transformer layers, color-coded by domain. Layer 0 represents fixed encoder embeddings, Layers 1–6 are encoder layers, Layer 7 represents fixed decoder embeddings, and Layers 8–13 are decoder layers. The model learns to distinguish between domains despite not being explicitly provided with any domain information.

Figure 10 reveals distinct domain-based partitioning in the representations across all encoder and deep decoder layers. The initial encoder layer (Layer 0) shows a nascent stage of this separation, which becomes more pronounced in deeper layers. This trend suggests that, as sentences progress through the Transformer, domain-specific features are identified and prominently maintained. While the encoder learns to partition the hidden space based on domains from scratch, the decoder accesses the encoder's hidden states via encoder-decoder attention, which may simplify domain discovery.

The persistence of domain-specific features in sentence representations, particularly in deeper layers, implies that the domain may be integral to the core meaning of the sentence, which the model preserves throughout its processing. This visualization supports *multi-domain specialization* characterization in contrast to multilingual abstraction.

3.4.2. Evidence 2: PCA Visualization

Following the t-SNE analysis, we used PCA to examine multi-domain specialization further. PCA helps to identify global linear relationships in the data. As Figure 11 shows, PCA separates sentences by domain, similar to the t-SNE results.

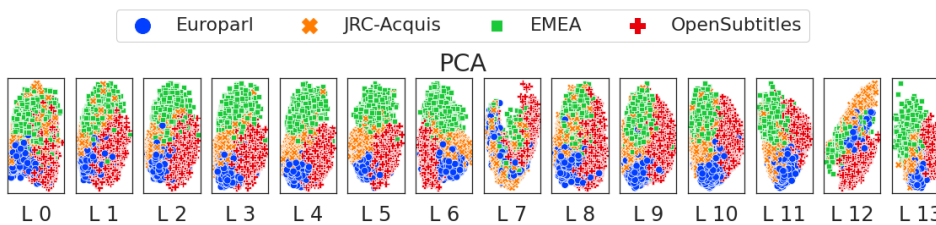

Figure 11: PCA visualization showing sentence representations by domain across Transformer layers.

We see that domains at layer 7 (L7) are separated a bit less clearly, but we assume this is due to L7 being the fixed decoder embedding layer (non-contextualized).

This makes processing at L7 quite shallow, focusing on the target sequence. This mirrors slightly less clear separation at L0 (non-contextual encoder layer). Starting from the L8, second decoder layer (as well as L1, second encoder layer), the separation becomes cleaner again.

Thus, results indicate that domain-specific features in sentence representations are not only locally differentiated (as demonstrated by the cosine t-SNE) but also globally distinct and linearly separable. This consistency across both analyses confirms multi-domain specialization in the Transformer model, highlighting the domain as a critical aspect of sentence meaning maintained throughout the model's layers.

3.4.3. Evidence 3: k-means Clustering

As a final step in confirming multi-domain specialization, we applied k-means clustering with four clusters to sentence representations from the fourth layer of the NMT model using the same multi-domain data as in the t-SNE and PCA analyses. Unlike previous methods, clustering does not rely on dimensionality reduction and offers quantitative insights.

Figure 12 presents the clustering results as a confusion table, showing how the sentence representations from different domains correspond to the clusters.

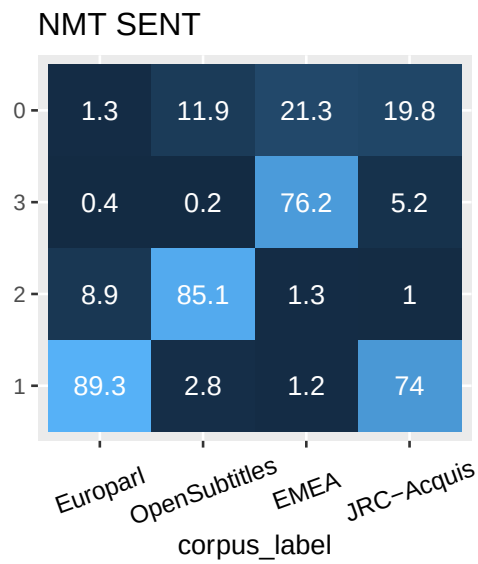

Figure 12: Cluster distribution of sentence representations in the EN-ET NMT model, indicating predominant domain-driven clustering.

The results predominantly align with the original data domain. A notable exception is the legal (JRC-Acquis) domain blending with parliamentary speeches (Europarl) in one cluster. This outcome contrasts with the t-SNE and PCA visualizations, in which the two domains were entirely distinct. It suggests that while

the JRC and Europarl differentiate within the same cluster, the distinction needs to be stronger for k-means to divide them into separate clusters.

These findings collectively demonstrate *multi-domain specialization* in the multi-domain NMT model. Unlike *multilingual abstraction*, where transformers abstract language-specific features, *multi-domain specialization* indicates the model's consistent identification and retention of domain-specific details in sentence representations. This differentiation is evident across various analytical techniques, including t-SNE, PCA, and k-means clustering, each revealing aspects of how the domain context is integral to sentence processing in the transformer model.

3.5. Result IV: Universality of Multi-Domain Specialization

We now examine the universality of multi-domain specialization after introducing it for multi-domain NMT models in Section 3.5. This mirrors our analysis of the universality of multilingual abstraction in Section 3.3.

Universality refers to the consistent presence of a phenomenon across diverse models and tasks. We present three pieces of evidence to support this hypothesis:

- **Evidence 1 (XLM-R):** Beyond NMT models, we found multi-domain specialization in XLM-R, indicating its presence across Transformer architectures with different training objectives.
- **Evidence 2 (Document-level Representations):** By examining document-level representations, we show that multi-domain specialization is even more distinct, confirming its persistence across textual scales.
- **Evidence 3 (Another Language Pair):** Analysis with another language pair also reveals multi-domain specialization, suggesting that it is universal across languages.

The following subsections elaborate on each piece of evidence, reinforcing multi-domain specialization as a universal trait in multi-domain Transformer models.

3.5.1. Evidence 1: Universality Across Training Objectives

To explore the universality of multi-domain specialization, we extend our investigation to the XLM-R model and examine its behavior across different training objectives. This analysis helps determine whether multi-domain specialization is consistent in diverse Transformer models. We replicated the clustering setup used in the NMT model in layer seven of the XLM-R model 3.2.2. The results are presented in Figure 13.

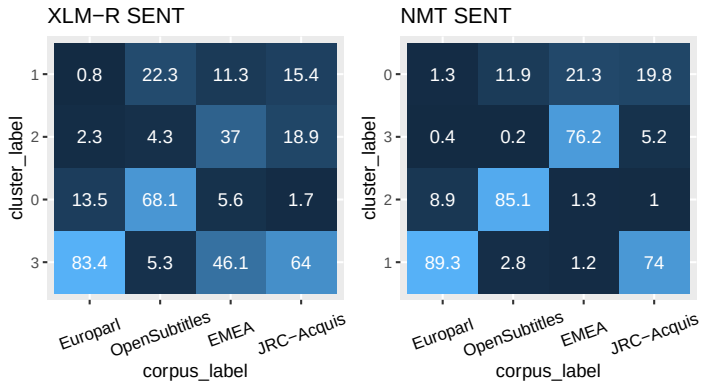

Figure 13: Clustering of domains in the XLM-R model, illustrating multi-domain specialization.

The clustering shows an interesting pattern: while subtitles data (OpenSubtitles) get its separate cluster, the other three domains group together. Despite this grouping, domains remain distinctly identifiable within the cluster, as highlighted by citetgoldber-clusters. We explicitly reveal this phenomenon by clustering in the following subsection.

3.5.2. Evidence 2: Document-level Representations

Building on our previous sentence-level analysis, we now investigate multi-domain specialization using document-level representations.

We averaged sentence representations across all documents within the multi-domain dataset for this experiment and then applied clustering.

The clustering results are shown in Figure 14.

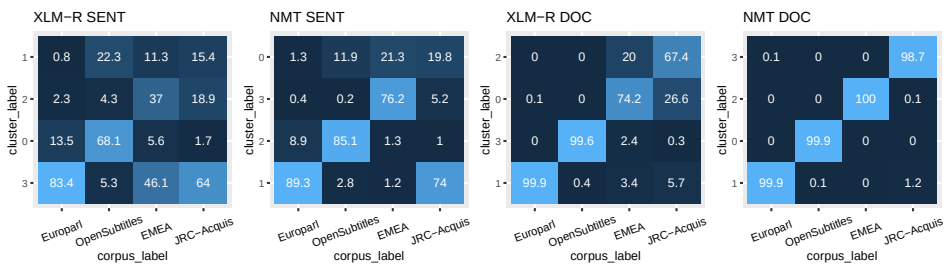

Figure 14: Document clustering in the EN-ET NMT model, closely aligning with the original data domains.

The results show that document-level clustering reflects the original data domains more accurately than sentence-level clustering. This improved alignment likely stems from individual sentences sometimes lacking strong domain-specific markers. In contrast, these markers are more prevalent at the document level as the representation encompasses multiple sentences. This content aggregation allows the model to apply the domain specialization process more successfully.

This evidence shows that multi-domain specialization becomes more apparent at the document level at both NMT and XLM-R, reinforcing the multi-domain specialization and universality hypotheses.

3.5.3. Evidence 3: Another Language Pair

Our final evidence for universality comes from analyzing the German-English language pair.

The clustering results are depicted in Figure 15.

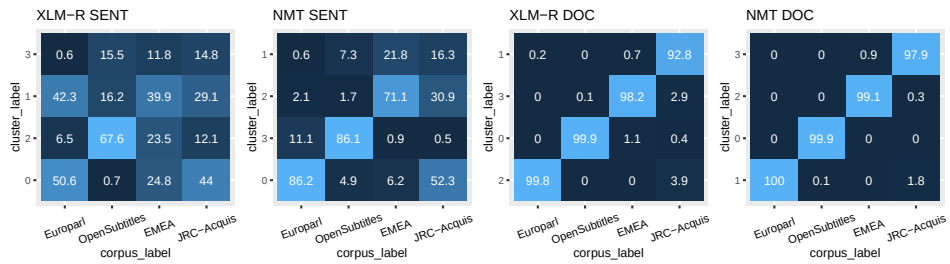

Figure 15: Domain clustering in the DE-EN NMT model, strongly aligned with original data domains.

The German-English results display a more precise clustering of document representations into their respective domains than the Estonian-English pair, further confirming the universality of the multi-domain specialization process.

A consistent pattern emerges across the three pieces of evidence presented in this chapter, demonstrating the universality of multi-domain specialization in Transformer models. Regardless of the language pair, whether Estonian-English or German-English, XLM-R and NMT models tend to organize representations according to domains. This pattern holds at both sentence and document levels, with the latter showing even stronger domain alignment. The implications of these findings extend to various applications, particularly in enhancing Domain Adaptation in NMT models. The universality of multi-domain specialization across different models, language pairs, and levels of textual granularity establishes it as a fundamental characteristic of multi-domain Transformers.

We also note that there is an interplay between multilingual abstraction and multi-domain specialization. Perfect abstraction and perfect specialization are two endpoints of the same spectrum. The model can choose to unify some latent concepts (abstraction) or to differentiate between them (specialization). In the case of multilingual models, even though the model has to retain some language specialization (at least in order to generate text in the correct language with correct grammar specific to that language), the amount of abstraction is large enough in the middle layers for us to describe multilingual models as following multilingual abstraction process.

This chapter has methodically explored and established four significant claims regarding Transformer models: multilingual abstraction, its universality, multi-

domain Specialization, and the universality of multi-domain specialization. Our evidence-based analyses demonstrate that these phenomena are fundamental and universal in Transformer models across various architectures, scales, training objectives, and languages. Recognizing these phenomena as inherent algorithms in Transformers demystifies these models and paves the way for more tailored and effective language models across low-resource languages and specific domains.

However, we made several statements requiring further investigation, and objections and conflicting evidence related to multilingual abstraction exist. In the next chapter, we address these issues and showcase a practical application of multi-domain specialization.

4. ADDRESSING OBJECTIONS AND PRACTICAL APPLICATION

4.1. Multilingual Abstraction: Addressing Conflicting Literature

Problem. Researchers have found mixed results in comparing multilingual representations in models like mBERT. Some studies suggest sentence representations in different languages become more distant in higher layers, while others see them becoming more alike. This difference is crucial for understanding multilingual abstraction.

Specifically, Singh et al. (2019a) used CLS-pooling and found languages diverged in upper layers. In contrast, Muller et al. (2021) used mean-pooling and observed the opposite. These conflicting findings are shown in Figure 16.

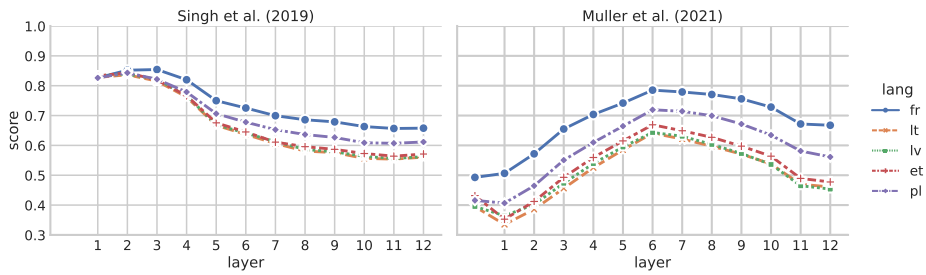

Figure 16: Different views on language representation in mBERT. The representations go from similar to dissimilar and thus diverge (left) vs from dissimilar to similar and thus converge (right). Adapted from Singh et al. (2019a) and Muller et al. (2021). The graph illustrates that two related works arrive at two opposing answers to the same question (convergence/divergence of language representations). As we debunk in this section, authors use different similarity indexes and sentence representation types which results in different patterns.

Methodological Discrepancy. When we extract representation for a specific sentence, we first get a list of token representations for each token. The second step is to get a single-vectored sentence representation, and there are two approaches to this: CLS-pooling and mean-pooling. CLS-pooing refers to simply taking the first token and treating it as a representation of the whole sentence, while mean-pooling corresponds to averaging all token representations overall.

We found that the main difference between these studies is their approach to pooling or summarizing a sentence's information: mean-pooling reveals multilingual abstraction while CLS-pooling does not. Therefore, establishing which approach is appropriate is crucial for our claims about multilingual abstraction.

To show that ***mean-pooling*** is a more reasonable pooling strategy to use, we considered three types of evidence:

- **Evidence 1 (Cosine Similarity):** CLS vectors in early layers are similar to random vectors, while mean-pooled vectors do not have this problem.
- **Evidence 2 (Probing Task):** Probing shows that mean-pooled vectors are more expressive then CLS-pooled vectors.
- **Evidence 3 (Argument from Self-Attention):** We present the analytical argument about using mean-pooling representations being more adherent to the properties of Transformer architecture.

This evidence leads us to conclude that mean-pooling is the right choice for analyzing language representation in mBERT, defending the multilingual abstraction. The following subsections explain each evidence in detail.

4.1.1. Evidence 1: Cosine Similarity

We employed cosine similarity as an alternative metric to the correlation-based indexes to assess the effectiveness of different pooling strategies for sentence representations in mBERT. Cosine similarity, a widely used measure in NLP, evaluates the angular proximity between vectors, indicating their semantic similarity.

Our methodology involved computing the average cosine similarity between English sentences and their corresponding translations in different languages. We also calculated the cosine similarity between English and randomly permuted target sentences. We performed these computations using three distinct pooling strategies: CLS pooling, mean pooling, and first token pooling. We present the results in Figure 17.

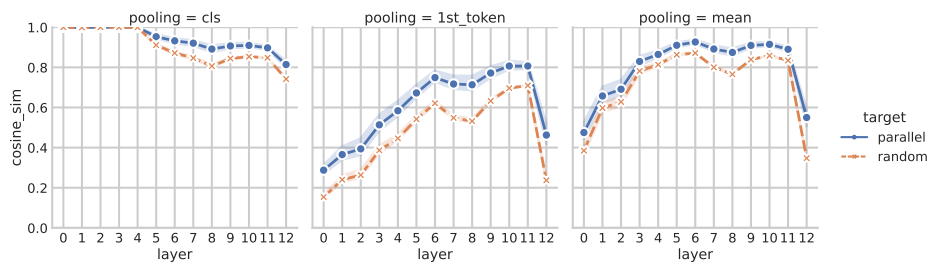

Figure 17: Cosine similarities of sentence representations under three different pooling strategies. The legend indicates whether similarity is measured between parallel sentences ("parallel") or arbitrary pairs of sentences ("random").

The most striking observation is that CLS-pooled representations exhibit high cosine similarities for the initial five layers, almost equal to one, regardless of whether sentences are parallel or randomly paired. This suggests that CLS-pooled vectors in the early layers cannot distinguish between translations and non-translations, pointing towards a limitation in their semantic representational power.

In contrast, the mean-pooled vectors demonstrate a more nuanced pattern, indicating a better capability for semantic distinction. This finding challenges the

reliability of CLS pooling in early layers for cross-lingual representational analysis, leading us to question the method's overall effectiveness in capturing semantically meaningful sentence representations.

4.1.2. Evidence 2: Probing Task Analysis

The probing task evaluates the effectiveness of pooling strategies in mBERT for sentence representation, focusing on cross-lingual closeness.

Methodology. Our approach involved using a multi-parallel corpus of 10,000 multi-parallel examples from XNLI dataset as in section 3.2.1. For each English sentence, we identified the most similar sentence in the target language based on vector representations. The accuracy was calculated on the basis of whether the closest sentence was the actual translation. We compared CLS pooling, mean pooling, and first-token pooling.

Figure 18 shows the averaged results across different languages.

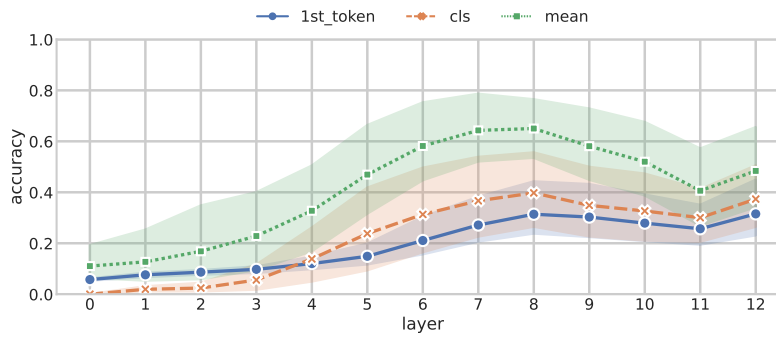

Figure 18: Accuracy of matching the closest sentence vector to the source sentence across pooling strategies averaged over four languages.

The results indicate that CLS pooling could be more effective in the first four layers, with accuracy near zero. This suggests its limited ability to capture cross-lingual similarities early on. Meanwhile, first-token pooling performs better than CLS pooling in these layers. Although CLS pooling improves in deeper layers, mean pooling consistently shows higher accuracy.

The probing task reveals that mean pooling is more effective than CLS pooling for sentence representation in mBERT, especially in tasks involving multiple languages. Mean-pooling demonstrates a more substantial capacity to capture semantically meaningful and consistent representations across languages, making it preferable.

4.1.3. Evidence 3: Argument from Self-Attention

The final piece of evidence examines the learning signal distribution in mBERT, mainly focusing on the CLS token's role across layers.

In mBERT, the CLS token is integral for the next sentence prediction task, but this is specifically relevant at the final layer. During pretraining, the model

uses the CLS token representation from the last layer to predict the subsequent sentence. However, this explicit learning signal for sentence-level representation is absent in the CLS token at layers other then last.

In the Transformer architecture, attention layers disperse information about each token across positions in each layer. As a result, the CLS token at intermediate layers is not obligated to accumulate information about the entire sentence until the very last layer: sentence information can be aggregated across many tokens, or be concentrated at some other token, and then moved by attention heads to CLS only at the very last layers. This can be observed in the earlier layers' low accuracy in the probing task (Figure 18).

Mean pooling, in contrast, aggregates information from all tokens at every layer, ensuring a comprehensive capture of sentence-level meaning. This approach is accessible from the limitation of relying on a single token that lacks a direct learning signal in all but the final layer.

Given these considerations, mean pooling emerges as a more reliable and conceptually sound choice for sentence representation in cross-lingual analysis. It provides a more holistic and evenly distributed representation of sentence meaning instead of the CLS token's more limited and layer-specific utility.

The evidence we presented addresses the conflicting interpretations of cross-lingual models' structural phenomena and is further supported by indirect evidence from related works in Section 3.2.3. By examining the impact of pooling strategies on representational analysis, we have reconciled the differing views on language representation in models like mBERT. Our findings support using mean-pooling as a more accurate method for extracting sentence representations, revealing a consistent pattern of multilingual abstraction. This supports the notion that multilingual models initially process language-specific features and gradually transition to more abstract, language-neutral representations.

4.2. Multilingual Abstraction: Addressing Outliers

This section addresses the presence of outlier languages in multilingual abstraction, particularly following up on evidence from Claim I concerning unsupervised visualization (Figure 5 from Section 3.2.2). We aim to demonstrate that specific languages not aligning with the multilingual abstraction pattern, such as Swahili and Urdu, are exceptions rather than indicative of a broader trend.

We base our analysis on an extensive set of 406 pairwise language comparisons across 29 languages in mBERT and XLM-R. Figure 19 presents a boxplot of CKA distances between language pairs for mBERT and XLM-R. Mean-pooled sentence vectors, representing the average of constituent word vectors, are used in this analysis.

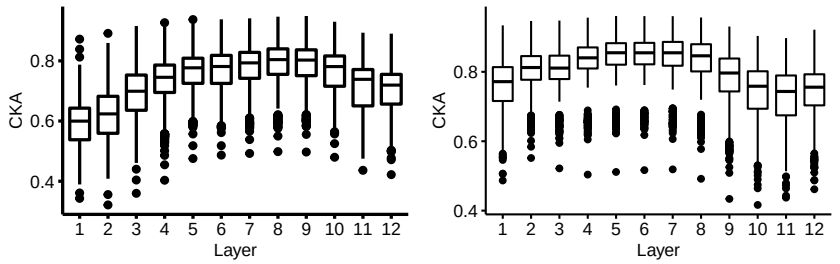

Figure 19: Multilingual abstraction across language pairs in mBERT and XLM-R, with outliers highlighted.

The box plot primarily confirms the multilingual abstraction pattern in most languages. However, some outliers deviate significantly, notably languages such as Urdu and Hindi.

To further investigate this, we focus on the most cross-lingual layers within mBERT and XLM-R and visualize the pairwise language similarities in these layers. Figure 20 provides a detailed heatmap of these relationships.

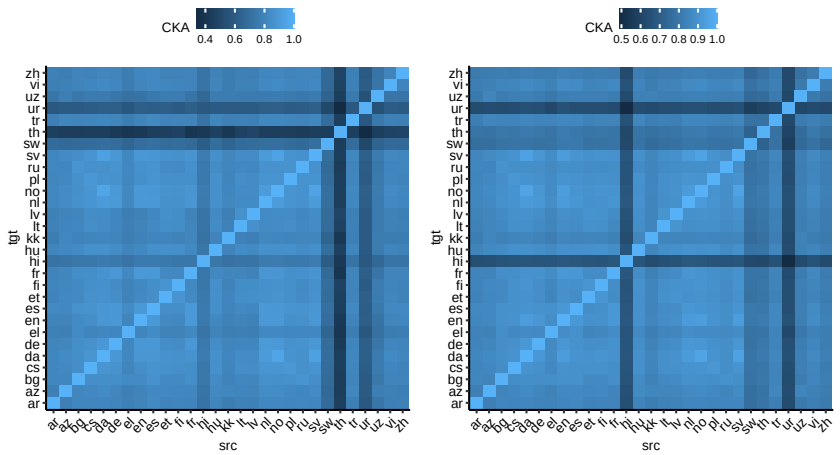

Figure 20: Layer-specific language similarities, identifying Urdu, Hindi, Swahili, and Thai as outliers.

This heatmap clarifies that languages like Urdu, Hindi, Swahili, and Thai are indeed outliers, diverging from the primary trend of multilingual abstraction.

To comprehend how languages group within their representational space, we executed agglomerative clustering based on CKA distances. We present results in Figure 21.

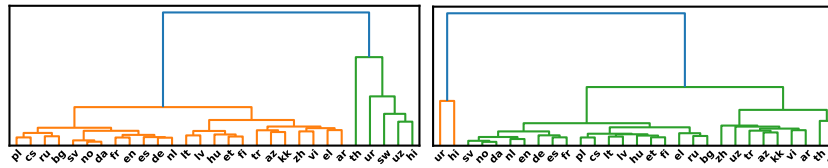

Figure 21: Linguistic clustering based on CKA distances, demonstrating language groupings and isolating outliers.

The linguistic tree shows that similarities in language-specific elements in representations follow cross-lingual linguistic similarities.

Swahili and Urdu are among the most low-resource languages in the datasets used by the models, while Hindi and Thai are high-resource languages that use distinct writing scripts. These factors could explain their divergence from the other languages, but identifying the exact reasons for Urdu, Hindi, Swahili, and Thai being the outliers would require a lot of experimental ablation studies and falls outside the scope of the thesis. We leave the checking of this hypothesis for future work.

4.3. Multilingual Abstraction: Addressing Conflicting Measurements

In Section 3.3.1, we argue that multilingual abstraction is universal in all normalization approaches. We used ANC, which is simple, interpretable, and reliable, for this analysis. However, as it turns out, the same study with CKA shows a drastically different result. This issue has to be addressed for our claim about the universality of multilingual abstraction to stay strong, so we address the issue in this section.

Problem. We train the following three XLM-Roberta (Conneau et al. 2020a) language models (`base` size versions) from scratch (each with a different normalization schema):

- Post-LN (`scale_post`): normalization block is placed *after* the residual connections in the transformer block (part of the original Transformer);
- Pre-LN (`scale_pre`): normalization block is placed *before* the residuals (this was shown to improve training by Xiong et al. (2020b));
- Normformer (`scale_normformer`): normalization block is placed *before* the residuals *and* FeedForward, Residual, and Self-Attention layers are also normalized (Shleifer et al. 2021b).

Next, we compare the representations with CKA, as described in Section 3.2.1.

We present the results in Figure. 22. This reveals that while `scale_post` and `scale_pre` generally follow multilingual abstraction evidence, the Normformer results are drastically different. While the similarity for the first half of the layers increases (layers 0-5), the CKA score drops dramatically in the middle layer of

the network and continues to hang around zero for all remaining layers (layers 6-12).

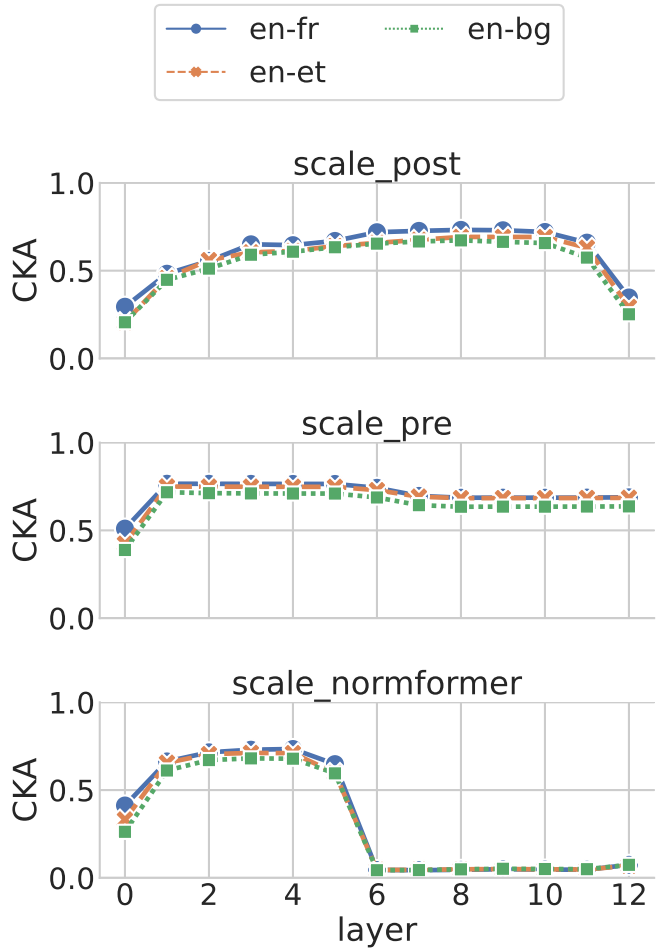

Figure 22: Counter-intuitive CKA (dis)similarity of XLM-Normformer layers. CKA index shows drastic dissimilarity for layers 6-12.

To double-check the result, we retrain the `scale_normformer` the second time with a different random restart and get the same results.

Evidence for Unreliability of CKA. In what follows, we argue about the unreliability of CKA by showing it gave a misleading answer in our experiment. Specifically, we present three pieces of evidence that suggest that language representations in later layers of XLM-Normformer do not drastically diverge despite CKA suggesting the opposite. The evidence is as follows:

- **Evidence 1** *(Zero-shot Cross-lingual Transfer)*: Zero-shot cross-lingual transfer normally occurs in the XLM-Normformer, which would be highly unlikely given the CKA result.
- **Evidence 2** *(Per-Layer Matching Accuracy)*: The probing task of cross-

lingual sentence matching hints at the representational similarity of the later layer in the XLM-Normformer.

- **Evidence 3** *(Results from SVCCA and PWCCA):* Both other representational similarity indexes produce the opposite results on the same dataset.

In the following subsections, we describe experiments that produce each piece of evidence.

4.3.1. Evidence 1: Zero-shot Cross-lingual Transfer

We conducted a zero-shot cross-lingual transfer test to compare to the CKA results for the Normformer model. This test involved training the models on an English language task and then assessing their performance on the same task in different languages. The aim was to see if the models' performance contradicted the CKA's implication of diverging language representations in later layers.

We used the XNLI dataset (Conneau et al. 2018) for a Natural Language Inference task. The models were first fine-tuned on the English data and then evaluated for transfer performance in French and Bulgarian.

The results in Table 3 display the models' zero-shot transfer accuracy.

Normalization	en	fr	bg
`scale_post`	0.79	0.72	0.70
`scale_pre`	0.81	0.72	0.72
`scale_normformer`	0.79	0.72	0.71

Table 3: Accuracy of XLM-Roberta Base Transformers pre-trained with different normalization schemes and fine-tuned on the English portion of the XNLI sentence classification task. The models show similar zero-shot cross-lingual transfer performance.

The `scale_normformer` shows effective zero-shot transfer, with minimal performance drop across languages. This contradicts the CKA's suggestion of significant language representation divergence in later layers.

These findings imply that despite CKA's results, the `scale_normformer` model maintains effective cross-lingual capabilities. This evidence questions the divergence suggested by CKA and leads us to explore more proofs to validate our claim.

4.3.2. Evidence 2: Per-Layer Matching Accuracy

Following our examination of zero-shot transfer, we further scrutinized the cross-lingual representational capabilities at each layer using a sentence-matching probing task.

This task involved identifying the closest sentence representation in one language to a given sentence in another language. We used mean-pooled sentence representations for this purpose. We located the closest target sentence in another

language for each English sentence from a set of 10,000 sentences based on cosine similarity. We marked an instance correct if the closest sentence was the actual translation. We calculated the matching accuracy as the ratio of correctly identified translations.

Figure 23 displays the layer-wise matching accuracy for the XLM-Normformer.

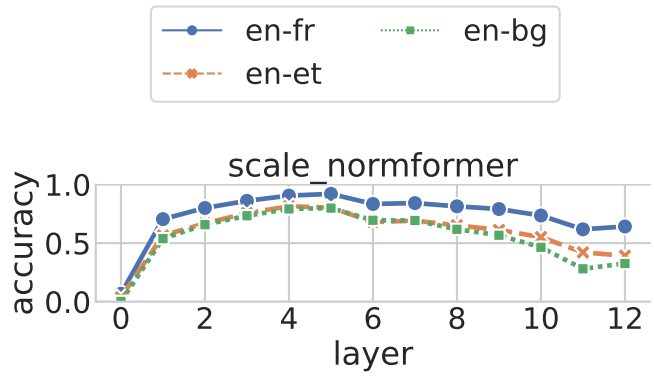

Figure 23: Layer-wise sentence matching accuracy in XLM-Normformer.

Notably, the deeper layers (6-12) exhibit significant cross-lingual matching scores, with more than 50% accuracy for French, contradicting the CKA findings of near-zero similarity.

This evidence suggests that these layers retain substantial cross-lingual representational similarity despite CKA's claim of drastic divergence. The slight decline in accuracy in deeper layers is far from the negligible similarity suggested by CKA.

This task reinforces our position on the CKA's misleading results in the context of cross-lingual representational similarity. In the following subsection, we examine evidence from two other widespread similarity indexes to strengthen our argument further.

4.3.3. Evidence 3: Results from SVCCA and PWCCA

After examining the zero-shot transfer and matching accuracy, we now assess the reliability of CKA by comparing it with results from other similarity indexes, namely SVCCA and PWCCA.

SVCCA offers a different perspective on representational similarity. In Figure 24, SVCCA results for XLM-Normformer display a clear multilingual abstraction pattern. This pattern is markedly different from the CKA findings.

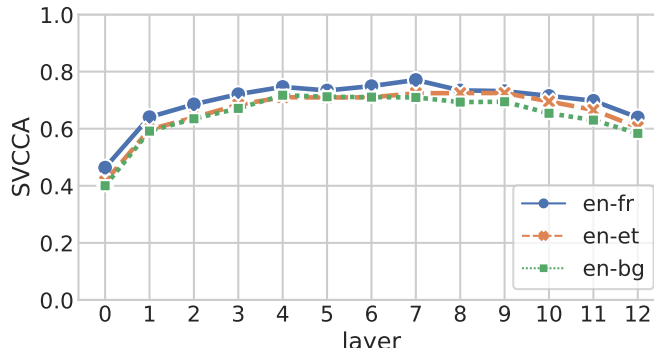

Figure 24: SVCCA similarity analysis for XLM-Normformer layers, showing a multilingual abstraction pattern.

Similarly, we applied PWCCA to analyze the same layers. As shown in Figure 25, similarity gradually increases, although not as pronounced as in SVCCA. However, it significantly differs from the CKA results, particularly in later layers with high similarity.

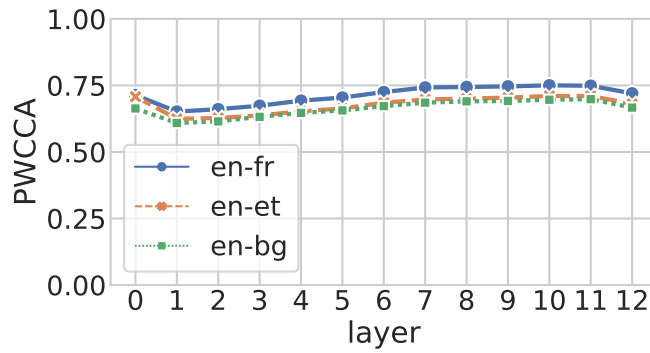

Figure 25: PWCCA similarity analysis for XLM-Normformer layers, demonstrating a discrepancy from CKA findings.

The contrasting results from SVCCA and PWCCA, compared to CKA, reinforce our argument regarding the unreliability of CKA in this specific context. These findings challenge the notion that representations in later layers of XLM-Normformer diverge to the extent suggested by CKA. Consequently, the cumulative evidence from various metrics and tasks supports the conclusion that CKA provided misleading insights in our Normformer experiment.

While the primary purpose of the NormFormer CKA experiment is to provide suggestive evidence motivating the use of more robust similarity indices, the exact reason why CKA fails to capture similarity at layers 6–12 is beyond the scope of this thesis. However, it is intriguing to hypothesize why this drastic drop occurs in CKA but not in other metrics like SVCCA.

As demonstrated by Kornblith et al. (2019), both CKA and SVCCA can be

expressed as normalized sums of dot products between the singular values obtained from the singular value decompositions (SVDs) of the representations X and Y. The key difference lies in how they handle these singular values. SVCCA considers only the top K singular values, truncating the rest, whereas CKA computes a weighted sum of all singular values, with weights corresponding to the eigenvalues.

At layers 6–12, it's possible that the top singular values of the representations are dissimilar, while the singular values accounting for less variance are more similar. In CKA, these similar but less dominant singular values receive minimal weights due to the weighting scheme, resulting in an overall similarity score that approaches zero. In contrast, SVCCA treats all top K singular values equally, allowing the similarity in the less dominant components to contribute more significantly to the overall score.

This suggests that one of the big discrepancies between CKA and other indexes arises from the CKA weighting strategy. Whether to prefer a weighted measure like CKA or an unweighted one like SVCCA or ANC is an empirical question. The experiments presented in this chapter aim to address this question and suggest that weighting by singular values can be not the best option for comparing multilingual models in some cases.

4.4. Multi-Domain Specialization: Practical Application

Exploring Language Alignment in Sections 3.2 and 3.3 presented challenges like conflicting literature and unreliable similarity indexes. However, our investigation into multi-domain specialization revealed no such complexities. At the same time, interpretability research often focuses on understanding machine learning models without necessarily translating these insights into practical applications. While this theoretical exploration is valuable, demonstrating how interpretability findings can be leveraged in practice enhances the impact of the field. Motivated by this perspective, this section presents a practical application that utilizes interpretability to improve domain adaptation in Neural Machine Translation (NMT) models.

Our contribution leverages the universal nature of domain specialization, meaning that the specialization is observed not only in LMs but also in NMT models. By using the NMT model itself as a source of embeddings to extract unsupervised clusters, we simplify the existing automatic domain adaptation framework. This approach eliminates the need for an external LM and achieves improved domain adaptation performance, with gains of up to 3 BLEU points for document-level clusters. Moreover, NMT-based clusters are conceptually more suited for fine-tuning because they align with the model's inherent specialization. This section, therefore, aims to demonstrate how insights from our interpretability work can enhance neural network systems in practical applications.

Our focus here is enhancing domain adaptation in neural machine translation

(NMT) by leveraging our multi-domain specialization finding. We observed that NMT models produce more distinct domain clusters than XLM-R models. This observation and the universality of multi-domain specialization suggest a promising approach for advancing automated domain adaptation frameworks in NMT.

We propose modifying an existing framework for automatic domain generation in NMT by utilizing the NMT model to generate domain-specific clusters, which we describe in the following subsections.

Our contribution in this section leverages the fact that domain specialization is universal. In particular, since specialization appears not only in LMs but also in NMT models, we can use an NMT model as a source of embeddings to extract unsupervised clusters as opposed to the external LM. Based on this observation, we simplify the existing automatic domain framework (Tars et al. 2018), eliminating the need for external LM without loss of performance.

4.4.1. Existing Framework

The existing framework, as described by Tars et al. (2018), uses automatic domain clusters for NMT domain adaptation. The process involves several steps, as illustrated in Figure 26 (left).

The first step (1.1) uses a heterogeneous dataset to train a baseline NMT model. Simultaneously (1.2), the same dataset is passed through an external pre-trained **XLM-R** model to extract sentence or document representations. The second step (2) involves training a k-means clustering model using these extracted representations. In the third step (3), the original dataset is divided into sub-datasets corresponding to the identified clusters. The final step involves fine-tuning the baseline NMT model on each cluster-specific dataset, resulting in specialized models for each domain.

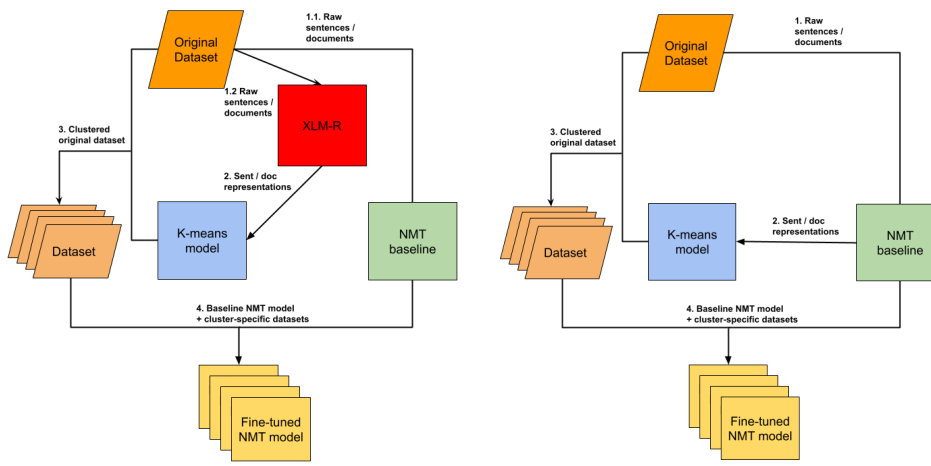

Figure 26: Comparison of the existing automatic domains framework (left) and our proposed revision (right).

4.4.2. Revising the Existing Framework

Our proposed revision of this framework (Figure 26, right) suggests using the NMT baseline for generating sentence/document representations, replacing the external XLM-R model used in step 1.2. This change is motivated by our finding that NMT models produce better-suited clusters for translation tasks. The rest of the process remains unchanged.

The coauthor of the corresponding publication conducted extensive experiments comparing both approaches. These experiments spanned two data scenarios, three language pairs, and sentence-level and document-level clustering. The results showed equal or superior performance (the gain is up to 3 BLUE points for document-level clusters) in terms of BLEU scores when compared to approaches using external language models, validating our proposed method.

Throughout this chapter, we have rigorously addressed objections and conflicting measurements, defended and extended the key claims regarding multilingual abstraction, and introduced a practical application of multi-domain specialization.

5. CONCLUSION

This thesis has explored the interpretability of multilingual and multi-domain Transformer-based models, providing insights into how these models internally represent and manipulate information across different languages and domains. The primary focus was on understanding two key representational phenomena: multilingual abstraction and multi-domain specialization. Through detailed analysis and the development of a novel methodology, we have demonstrated the consistent emergence of these phenomena across various models, training datasets, and architectural variations, suggesting their universal nature.

The concept of multilingual abstraction reveals how sentence representations in multilingual models transition from language-specific to language-agnostic states as they process information. This transformation highlights the model's ability to generalize across languages, focusing more on the underlying meaning rather than linguistic specifics. In contrast, multi-domain specialization shows that multi-domain models maintain and enhance domain-specific features throughout their layers, ensuring that specialized knowledge is preserved and utilized effectively.

As a means to our end, we introduced the Averaged Neuronwise Correlation (ANC) methodology to support our analysis, providing a robust tool for comparing representations in multilingual contexts. This methodology allowed us to make stronger claims about the multilingual abstraction, helped to resolve conflicting evidence and clashing literature, and drove out further analysis. Although not central to our thesis, by introducing a method for improving domain adaptation in neural machine translation, we also demonstrated the practical applicability of our findings.

Future research can continue to explore the inner workings of multilingual and multi-domain models on a more fine-grained level, employing methods such as dictionary learning and circuit discovery (Templeton et al. 2024) while integrating these insights into the design and training of new models can lead to advancements in AI technology that are both innovative and ethically sound.

In summary, this thesis has comprehensively analyzed high-level mechanistic decision-making in multilingual and multi-domain Transformer-based models, establishing key representation patterns and their universality. This thesis's insights can help improve the safety, fairness, and accessibility of AI technology, particularly for underrepresented languages and domains. By enhancing our understanding of how multilingual and multi-domain models operate, we can work towards a more democratized artificial intelligence.

BIBLIOGRAPHY

Anthropic (2023). *Introducing Claude*. URL: https://www.anthropic.com/index/introducing-claude.

Artetxe, Mikel, Labaka, Gorka, and Agirre, Eneko (Oct. 2018). "Unsupervised Statistical Machine Translation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3632–3642. DOI: 10.18653/v1/D18-1399. URL: https://aclanthology.org/D18-1399.

Artetxe, Mikel, Ruder, Sebastian, and Yogatama, Dani (July 2020). "On the Cross-lingual Transferability of Monolingual Representations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4623–4637. DOI: 10.18653/v1/2020.acl-main.421. URL: https://aclanthology.org/2020.acl-main.421.

Ba, Jimmy Lei, Kiros, Jamie Ryan, and Hinton, Geoffrey E. (2016). *Layer Normalization*. cite arxiv:1607.06450. URL: http://arxiv.org/abs/1607.06450.

Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *CoRR* abs/1409.0473.

Bergsma, Wicher (Sept. 2013). "A bias-correction for Cramér's V and Tschuprow's T". In: *Journal of the Korean Statistical Society* 42. DOI: 10.1016/j.jkss.2012.10.002.

Bricken, Trenton, Templeton, Adly, Batson, Joshua, Chen, Brian, Jermyn, Adam, Conerly, Tom, Turner, Nick, Anil, Cem, Denison, Carson, Askell, Amanda, Lasenby, Robert, Wu, Yifan, Kravec, Shauna, Schiefer, Nicholas, Maxwell, Tim, Joseph, Nicholas, Hatfield-Dodds, Zac, Tamkin, Alex, Nguyen, Karina, McLean, Brayden, Burke, Josiah E, Hume, Tristan, Carter, Shan, Henighan, Tom, and Olah, Christopher (2023). "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning". In: *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Brown, Tom, Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared D, Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, Agarwal, Sandhini, Herbert-Voss, Ariel, Krueger, Gretchen, Henighan, Tom, Child, Rewon, Ramesh, Aditya, Ziegler, Daniel, Wu, Jeffrey, Winter, Clemens, Hesse, Chris, Chen, Mark, Sigler, Eric, Litwin, Mateusz, Gray, Scott, Chess, Benjamin, Clark, Jack, Berner, Christopher, McCandlish, Sam, Radford, Alec, Sutskever, Ilya, and Amodei, Dario (2020a). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 1877–1901.

URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Brown, Tom, Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared D, Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, Agarwal, Sandhini, Herbert-Voss, Ariel, Krueger, Gretchen, Henighan, Tom, Child, Rewon, Ramesh, Aditya, Ziegler, Daniel, Wu, Jeffrey, Winter, Clemens, Hesse, Chris, Chen, Mark, Sigler, Eric, Litwin, Mateusz, Gray, Scott, Chess, Benjamin, Clark, Jack, Berner, Christopher, McCandlish, Sam, Radford, Alec, Sutskever, Ilya, and Amodei, Dario (2020b). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Cao, Steven, Kitaev, Nikita, and Klein, Dan (2020). "Multilingual Alignment of Contextual Word Representations". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=r1xCMyBtPS.

Chen, Lili, Lu, Kevin, Rajeswaran, Aravind, Lee, Kimin, Grover, Aditya, Laskin, Misha, Abbeel, Pieter, Srinivas, Aravind, and Mordatch, Igor (2021). "Decision Transformer: Reinforcement Learning via Sequence Modeling". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., pp. 15084–15097. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/7f489f642a0ddb10272b5c31057f0663-Paper.pdf.

Conneau, Alexis, Khandelwal, Kartikay, Goyal, Naman, Chaudhary, Vishrav, Wenzek, Guillaume, Guzmán, Francisco, Grave, Edouard, Ott, Myle, Zettlemoyer, Luke, and Stoyanov, Veselin (July 2020a). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747.

Conneau, Alexis, Rinott, Ruty, Lample, Guillaume, Williams, Adina, Bowman, Samuel R., Schwenk, Holger, and Stoyanov, Veselin (2018). "XNLI: Evaluating Cross-lingual Sentence Representations". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics.

Conneau, Alexis, Wu, Shijie, Li, Haoran, Zettlemoyer, Luke, and Stoyanov, Veselin (July 2020b). "Emerging Cross-lingual Structure in Pretrained Language Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6022–6034. DOI: 10.18653/v1/2020.acl-main.536. URL: https://aclanthology.org/2020.acl-main.536.

Costa-jussà, Marta Ruiz, Cross, James, cCelebi, Onur, Elbayad, Maha, Heafield, Kenneth, Heffernan, Kevin, Kalbassi, Elahe, Lam, Janice, Licht, Daniel, Mail-

lard, Jean, Sun, Anna, Wang, Skyler, Wenzek, Guillaume, Youngblood, Alison, Akula, Bapi, Barrault, Loïc, Gonzalez, Gabriel Mejia, Hansanti, Prangthip, Hoffman, John, Jarrett, Semarley, Sadagopan, Kaushik Ram, Rowe, Dirk, Spruit, Shannon L., Tran, C., Andrews, Pierre Yves, Ayan, Necip Fazil, Bhosale, Shruti, Edunov, Sergey, Fan, Angela, Gao, Cynthia, Goswami, Vedanuj, Guzm'an, Francisco, Koehn, Philipp, Mourachko, Alexandre, Ropers, Christophe, Saleem, Safiyyah, Schwenk, Holger, and Wang, Jeff (2022). "No Language Left Behind: Scaling Human-Centered Machine Translation". In: *ArXiv* abs/2207.04672. URL: https://api.semanticscholar.org/CorpusID:250425961.

Del, **Maksym** and Fishel, Mark (2021a). "Similarity of Sentence Representations in Multilingual LMs: Resolving Conflicting Literature and a Case Study of Baltic Languages". In: *Baltic Journal of Modern Computing* 10. URL: https://api.semanticscholar.org/CorpusID:249921326.

Del, **Maksym** and Fishel, Mark (Nov. 2022). "Cross-lingual Similarity of Multilingual Representations Revisited". In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online only: Association for Computational Linguistics, pp. 185–195. URL: https://aclanthology.org/2022.aacl-main.15.

Del, **Maksym** and Fishel, Mark (July 2023). "True Detective: A Deep Abductive Reasoning Benchmark Undoable for GPT-3 and Challenging for GPT-4". In: *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*. Toronto, Canada: Association for Computational Linguistics, pp. 314–322. URL: https://aclanthology.org/2023.starsem-1.28.

Del, **Maksym**, Korotkova, Elizaveta, and Fishel, Mark (Nov. 2021b). "Translation Transformers Rediscover Inherent Data Domains". In: *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics, pp. 599–613. URL: https://aclanthology.org/2021.wmt-1.65.

Del, **Maksym**, Tättar, Andre, and Fishel, Mark (Oct. 2018). "Phrase-based Unsupervised Machine Translation with Compositional Phrase Embeddings". In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, pp. 361–367. URL: https://aclanthology.org/W18-6407.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics,

pp. 4171–4186. DOI: 10.18653/v1/n19-1423. URL: https://doi.org/10.18653/v1/n19-1423.

Ding, Frances, Denain, Jean-Stanislas, and Steinhardt, Jacob (2021). "Grounding Representation Similarity with Statistical Testing". In: *ArXiv* abs/2108.01661.

Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, Uszkoreit, Jakob, and Houlsby, Neil (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. URL: https://openreview.net/forum?id=YicbFdNTTy.

Elhage, Nelson, Hume, Tristan, Olsson, Catherine, Schiefer, Nicholas, Henighan, Tom, Kravec, Shauna, Hatfield-Dodds, Zac, Lasenby, Robert, Drain, Dawn, Chen, Carol, Grosse, Roger, McCandlish, Sam, Kaplan, Jared, Amodei, Dario, Wattenberg, Martin, and Olah, Christopher (2022). "Toy Models of Superposition". In: *Transformer Circuits Thread*. URL: https://transformer-circuits.pub/2022/toy_model/index.html.

Esplà, Miquel, Forcada, Mikel, Ramírez-Sánchez, Gema, and Hoang, Hieu (Aug. 2019). "ParaCrawl: Web-scale parallel corpora for the languages of the EU". In: *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*. Dublin, Ireland: European Association for Machine Translation, pp. 118–119. URL: https://www.aclweb.org/anthology/W19-6721.

F.R.S., Karl Pearson (1901). "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11, pp. 559–572. DOI: 10.1080/14786440109462720.

Gamma, Erich, Helm, Richard, Johnson, Ralph, and Vlissides, John (1995). *Design patterns: elements of reusable object-oriented software*. Pearson Deutschland GmbH.

Google (2023). *An important next step on our AI journey*. URL: https://blog.google/technology/ai/bard-google-ai-search-updates/.

Gupta, Abhijeet, Boleda, Gemma, Baroni, Marco, and Padó, Sebastian (Sept. 2015). "Distributional vectors encode referential attributes". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 12–21. DOI: 10.18653/v1/D15-1002. URL: https://aclanthology.org/D15-1002.

Hämmerl, Katharina, Libovický, Jindřich, and Fraser, Alexander (Aug. 2024). "Understanding Cross-Lingual Alignment—A Survey". In: *Findings of the Association for Computational Linguistics ACL 2024*. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, pp. 10922–10943. URL: https://aclanthology.org/2024.findings-acl.649.

Hotelling, Harold (Dec. 1936). “Relations Between Two Sets Of Variates*”. In: *Biometrika* 28.3-4, pp. 321–377. ISSN: 0006-3444. DOI: 10.1093/biomet/28.3-4.321. eprint: https://academic.oup.com/biomet/article-pdf/28/3-4/321/586830/28-3-4-321.pdf. URL: https://doi.org/10.1093/biomet/28.3-4.321.

Hu, Junjie, Ruder, Sebastian, Siddhant, Aditya, Neubig, Graham, Firat, Orhan, and Johnson, Melvin (July 2020). “XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation”. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 4411–4421. URL: https://proceedings.mlr.press/v119/hu20b.html.

Koehn, Philipp (Jan. 2004). “Statistical Significance Tests for Machine Translation Evaluation.” In: pp. 388–395.

Koehn, Philipp (2005). “Europarl : A Parallel Corpus for Statistical Machine Translation”. In: *MT Summit* 11.

Köhn, Arne (Sept. 2015). “What’s in an Embedding? Analyzing Word Embeddings through Multilingual Evaluation”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 2067–2073. DOI: 10.18653/v1/D15-1246. URL: https://aclanthology.org/D15-1246.

Kornblith, Simon, Norouzi, Mohammad, Lee, Honglak, and Hinton, Geoffrey (June 2019). “Similarity of Neural Network Representations Revisited”. In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 3519–3529. URL: https://proceedings.mlr.press/v97/kornblith19a.html.

Korotkova, Elizaveta, Luhtaru, Agnes, **Del**, **Maksym**, Liin, Krista, Deksne, Daiga, and Fishel, Mark (2019). *Grammatical Error Correction and Style Transfer via Zero-shot Monolingual Translation*. arXiv: 1903.11283 [cs.CL]. URL: https://arxiv.org/abs/1903.11283.

Kudugunta, Sneha, Bapna, Ankur, Caswell, Isaac, and Firat, Orhan (2019). “Investigating Multilingual NMT Representations at Scale”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. DOI: 10.18653/v1/d19-1167. URL: http://dx.doi.org/10.18653/v1/D19-1167.

Lample, Guillaume, Ott, Myle, Conneau, Alexis, Denoyer, Ludovic, and Ranzato, Marc’Aurelio (Oct. 2018). “Phrase-Based & Neural Unsupervised Machine Translation”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 5039–5049. DOI: 10.18653/v1/D18-1549. URL: https://aclanthology.org/D18-1549.

Li, Yixuan, Yosinski, Jason, Clune, Jeff, Lipson, Hod, and Hopcroft, John E. (2015). "Convergent Learning: Do different neural networks learn the same representations?" In: *FE@NIPS*.

Liang, Yaobo, Duan, Nan, Gong, Yeyun, Wu, Ning, Guo, Fenfei, Qi, Weizhen, Gong, Ming, Shou, Linjun, Jiang, Daxin, Cao, Guihong, Fan, Xiaodong, Zhang, Ruofei, Agrawal, Rahul, Cui, Edward, Wei, Sining, Bharti, Taroon, Qiao, Ying, Chen, Jiun-Hung, Wu, Winnie, Liu, Shuguang, Yang, Fan, Campos, Daniel, Majumder, Rangan, and Zhou, Ming (Nov. 2020). "XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 6008–6018. DOI: 10.18653/v1/2020.emnlp-main.484. URL: https://aclanthology.org/2020.emnlp-main.484.

Lin, Xi Victoria, Mihaylov, Todor, Artetxe, Mikel, Wang, Tianlu, Chen, Shuohui, Simig, Daniel, Ott, Myle, Goyal, Naman, Bhosale, Shruti, Du, Jingfei, Pasunuru, Ramakanth, Shleifer, Sam, Koura, Punit Singh, Chaudhary, Vishrav, O'Horo, Brian, Wang, Jeff, Zettlemoyer, Luke, Kozareva, Zornitsa, Diab, Mona, Stoyanov, Veselin, and Li, Xian (Dec. 2022). "Few-shot Learning with Multilingual Generative Language Models". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 9019–9052. URL: https://aclanthology.org/2022.emnlp-main.616.

Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, and Stoyanov, Veselin (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692. arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692.

Lloyd, S. (1982). "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

Luhtaru, Agnes, Purason, Taido, Vainikko, Martin, **Del**, **Maksym**, and Fishel, Mark (2024). *To Err Is Human, but Llamas Can Learn It Too*. arXiv: 2403.05493 [cs.CL]. URL: https://arxiv.org/abs/2403.05493.

Maaten, Laurens van der and Hinton, Geoffrey (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Mikolov, Tomas, Chen, Kai, Corrado, G., and Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *ICLR*.

Morcos, Ari, Raghu, Maithra, and Bengio, Samy (2018). "Insights on representational similarity in neural networks with canonical correlation". In: *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., pp. 5732–5741. URL: http://papers.nips.cc/paper/7815-insights-on-representational-similarity-in-neural-networks-with-canonical-correlation.pdf.

Muller, Benjamin, Elazar, Yanai, Sagot, Benoıt̂, and Seddah, Djamé (Apr. 2021). "First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 2214–2231. URL: https://www.aclweb.org/anthology/2021.eacl-main.189.

Murtagh, Fionn and Contreras, Pedro (Jan. 2012). "Algorithms for hierarchical clustering: An overview". In: *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery* 2, pp. 86–97. DOI: 10.1002/widm.53.

Nanda, Neel (2022). *A Comprehensive Mechanistic Interpretability Explainer & Glossary*. URL: https://www.neelnanda.io/mechanistic-interpretability/glossary.

Olah, Chris (2022). *Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases*. URL: https://transformer-circuits.pub/2022/mech-interp-essay/index.html.

Olah, Chris, Cammarata, Nick, Schubert, Ludwig, Goh, Gabriel, Petrov, Michael, and Carter, Shan (2020). "Zoom In: An Introduction to Circuits". In: *Distill*. https://distill.pub/2020/circuits/zoom-in. DOI: 10.23915/distill.00024.001.

OpenAI (2023). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL].

Radford, Alec, Narasimhan, Karthik, Salimans, Tim, and Sutskever, Ilya (2018). "Improving language understanding by generative pre-training". In.

Raghu, Maithra, Gilmer, Justin, Yosinski, Jason, and Sohl-Dickstein, Jascha (2017). "SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability". In: *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pp. 6076–6085. URL: http://papers.nips.cc/paper/7188-svcca-singular-vector-canonical-correlation-analysis-for-deep-learning-dynamics-and-interpretability.pdf.

Räuker, Tilman, Ho, Anson, Casper, Stephen, and Hadfield-Menell, Dylan (2023). *Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks*. arXiv: 2207.13243 [cs.LG].

Rikters, Matīss, Amrhein, Chantal, **Del**, **Maksym**, and Fishel, Mark (Sept. 2017). "C-3MA: Tartu-Riga-Zurich Translation Systems for WMT17". In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 382–388. URL: https://aclanthology.org/W17-4738.

Shleifer, Sam, Weston, Jason, and Ott, Myle (2021a). *NormFormer: Improved Transformer Pretraining with Extra Normalization*. arXiv: 2110.09456 [cs.CL].

Shleifer, Sam, Weston, Jason, and Ott, Myle (2021b). "NormFormer: Improved Transformer Pretraining with Extra Normalization". In: arXiv. DOI: 10.48550/ARXIV.2110.09456. URL: https://arxiv.org/abs/2110.09456.

Singh, Jasdeep, McCann, Bryan, Socher, Richard, and Xiong, Caiming (Nov. 2019a). "BERT is Not an Interlingua and the Bias of Tokenization". In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, pp. 47–55. DOI: 10.18653/v1/D19-6106. URL: https://www.aclweb.org/anthology/D19-6106.

Singh, Jasdeep, McCann, Bryan, Socher, Richard, and Xiong, Caiming (2019b). "BERT is Not an Interlingua and the Bias of Tokenization". In: *EMNLP*.

Steinberger, Ralf, Pouliquen, Bruno, Widiger, Anna, Ignat, Camelia, Erjavec, Tomaž, Tufiş, Dan, and Varga, Dániel (2006). "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages". In: *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*.

Tars, Sander and Fishel, Mark (May 2018). "Multi-Domain Neural Machine Translation". In: *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*. Alicante, Spain, pp. 279–288. URL: https://aclanthology.org/2018.eamt-main.26.

Templeton, Adly, Conerly, Tom, Marcus, Jonathan, Lindsey, Jack, Bricken, Trenton, Chen, Brian, Pearce, Adam, Citro, Craig, Ameisen, Emmanuel, Jones, Andy, Cunningham, Hoagy, Turner, Nicholas L, McDougall, Callum, MacDiarmid, Monte, Freeman, C. Daniel, Sumers, Theodore R., Rees, Edward, Batson, Joshua, Jermyn, Adam, Carter, Shan, Olah, Chris, and Henighan, Tom (2024). "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet". In: *Transformer Circuits Thread*. URL: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Touvron, Hugo, Martin, Louis, Stone, Kevin, Albert, Peter, Almahairi, Amjad, Babaei, Yasmine, Bashlykov, Nikolay, Batra, Soumya, Bhargava, Prajjwal, Bhosale, Shruti, Bikel, Dan, Blecher, Lukas, Ferrer, Cristian Canton, Chen, Moya, Cucurull, Guillem, Esiobu, David, Fernandes, Jude, Fu, Jeremy, Fu, Wenyin, Fuller, Brian, Gao, Cynthia, Goswami, Vedanuj, Goyal, Naman, Hartshorn, Anthony, Hosseini, Saghar, Hou, Rui, Inan, Hakan, Kardas, Marcin, Kerkez, Viktor, Khabsa, Madian, Kloumann, Isabel, Korenev, Artem, Koura, Punit Singh, Lachaux, Marie-Anne, Lavril, Thibaut, Lee, Jenya, Liskovich, Diana, Lu, Yinghai, Mao, Yuning, Martinet, Xavier, Mihaylov, Todor, Mishra, Pushkar, Molybog, Igor, Nie, Yixin, Poulton, Andrew, Reizenstein, Jeremy, Rungta, Rashi, Saladi, Kalyan, Schelten, Alan, Silva, Ruan, Smith, Eric Michael, Subramanian, Ranjan, Tan, Xiaoqing Ellen, Tang, Binh, Taylor, Ross, Williams, Adina, Kuan, Jian Xiang, Xu, Puxin, Yan, Zheng, Zarov, Iliyan, Zhang, Yuchen, Fan, Angela, Kambadur, Melanie, Narang, Sharan, Rodriguez, Aurelien, Stojnic, Robert, Edunov, Sergey, and Scialom, Thomas (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv: 2307.09288 [cs.CL].

University of Tartu (2018). *UT Rocket*. DOI: 10.23673/PH6N-0144.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, and Polosukhin, Illia (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vázquez, Raúl, Celikkanat, Hande, Creutz, Mathias, and Tiedemann, Jörg (Aug. 2021). "On the differences between BERT and MT encoder spaces and how to address them in translation tasks". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*. Online: Association for Computational Linguistics, pp. 337–347. DOI: 10.18653/v1/2021.acl-srw.35. URL: https://aclanthology.org/2021.acl-srw.35.

Voss, Chelsea, Goh, Gabriel, Cammarata, Nick, Petrov, Michael, Schubert, Ludwig, and Olah, Chris (2021). "Branch Specialization". In: *Distill*. https://distill.pub/2020/circuits/branch-specialization. DOI: 10.23915/distill.00024.008.

Wang, Hao-Ting, Smallwood, Jonathan, Mourao-Miranda, Janaina, Xia, Cedric Huchuan, Satterthwaite, Theodore D., Bassett, Danielle S., and Bzdok, Danilo (2020). "Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists". In: *NeuroImage* 216, p. 116745. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2020.116745. URL: https://www.sciencedirect.com/science/article/pii/S1053811920302329.

Wang, Liwei, Hu, Lunjia, Gu, Jia-Yuan, Wu, Yue Kris, Hu, Zhiqiang, He, Kun, and Hopcroft, John E. (2018). "Towards Understanding Learning Representations: To What Extent Do Different Neural Networks Learn the Same Representation". In: *NeurIPS*.

Xiong, Ruibin, Yang, Yunchang, He, Di, Zheng, Kai, Zheng, Shuxin, Xing, Chen, Zhang, Huishuai, Lan, Yanyan, Wang, Liwei, and Liu, Tie-Yan (2020a). "On Layer Normalization in the Transformer Architecture". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org.

Xiong, Ruibin, Yang, Yunchang, He, Di, Zheng, Kai, Zheng, Shuxin, Xing, Chen, Zhang, Huishuai, Lan, Yanyan, Wang, Liwei, and Liu, Tie-Yan (2020b). "On Layer Normalization in the Transformer Architecture". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org.

Zhang, Aston, Lipton, Zachary C., Li, Mu, and Smola, Alexander J. (2021). "Dive into Deep Learning". In: *arXiv preprint arXiv:2106.11342*.

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Mark Fišel, for his unwavering support, insightful guidance, and constant encouragement throughout this journey. His mentorship has been essential in shaping this research and my personal and professional growth.

Special thanks to my colleagues and friends at the Institute of Computer Science of the University of Tartu, whose camaraderie and collaboration made this process more enriching.

I also express my gratitude to the university's administration and all the staff for fostering an environment that supports research and personal growth and for continuously striving to make life better for the university community.

I am deeply grateful to my family in Bamberg for graciously hosting me during a summer month, where the initial draft of this thesis was written. Their warmth and hospitality created the ideal environment for focus and creativity, for which I am truly thankful.

Finally, this work is dedicated to all those who have inspired my curiosity and passion for research and to the broader community of scholars working to push the boundaries of knowledge and make the world a better place.

SISUKOKKUVÕTE

Mitmekeelsed ja mitut tekstivaldkonda hõlmavad esituste mustrid transformeripõhistes mudelites

Käesolevas töös uuritakse mitmekeelsete ja mitut valdkonda hõlmavate transformeritel põhinevate mudelite sisemist toimimist, keskendudes sellele, kuidas nad esitavad ja manipuleerivad teavet eri keeltes ja tekstivaldkondades. Transformeritel põhinevad mudelid nagu mBERT ja XLM-R on nende võime tõttu käsitleda suurt hulka mitmekeelseid ja mitut valdkonda hõlmavaid andmeid muutunud loomuliku keele töötluse (NLP) oluliseks osaks. Hoolimata nende tõhususest on endiselt väljakutseks mõista, kuidas need mudelid abstraheerivad ja spetsialiseerivad oma sisemisi esitusi. Käesolev väitekiri keskendub sellele väljakutsele, tutvustades kahte põhinähtust - *mitmekeelne abstraktsioon* ja *mitut valdkonda hõlmav spetsialseerumine* - ja käsitledes neid nii teoreetilisest aspektist kui ka pakkudes mudelite tõlgendatavusele välja praktilisi rakendusi. Kaks põhilist sisupeatükki on peatükid 3 and 4.

Peatükis 3 esitasime neli peamist järeldust mitmekeelsete ja mitut tekstivaldkonda hõlmavate transformerite kohta.

- **ANC metoodika**: Varjatud olekute esituste võrdlemise hõlbustamiseks tutvustas doktoritöö *Averaged Neuronwise Correlation (ANC)* metoodit. See uus meetod oli tugev ja tõlgendatav vahend mitmekeelsete esituste võrdlemiseks, võimaldades selgemaid järeldusi mitmekeelse abstraktsiooni kohta ja aidates lahendada varasemate uuringute vastuolulisi järeldusi.
- **Multilingual Abstraction**: Mitme erineva keele peal treenitud transformeritel põhinevad mudelid nagu mBERT ja XLM-R töötlevad oma algkihis keelespetsiifilisi tunnuseid. Kui kujutised liiguvad sügavamale tehisnärvivõrku, muutuvad nad abstraktsemaks ja keeleliselt neutraalsemaks, keskendudes lausete ühisele tähendusele eri keeltes. Seda *mitmekeelset abstraktsiooni* toetasid mitmed tõendid, sealhulgas visualiseerimine meetoditega t-SNE ja PCA, mis näitasid keelekohaste klastrite konvergentsi semantiliselt ühtsemateks esitusteks sügavamates kihtides.
- **Universality of Multilingual Abstraction**: Üleminekut keelespetsiifilistelt esitustelt abstraktsetele esitustele täheldati mudelite erinevate arhitektuuride, treeningeesmärkide ja parameetrite ning treeningandmete mahu puhul, mis viitab sellele, et mitmekeelne abstraktsioon on transformeritel põhinevate mudelite *universaalne* omadus. See järeldus kehtis nii mBERT-i kui ka XLM-R-i puhul, vaatamata nende mudelite arhitektuurilistele erinevustele, mis kinnitab mitmekeelse abstraktsiooni üldist rakendatavust erinevatele transformeritel põhinevatele mudelitele.
- **Multi-Domain Specialization**: Erinevalt mitmekeelsetest mudelitest säilitavad *multi-domain* mudelid, mis on treenitud eri tekstivaldkondade (nt me-

ditsiinilised, juriidilised ja parlamentaarsed tekstid) andmekogumite peal, kõigis oma kihtides valdkonnapõhiseid representatsioone. Selle asemel, et erinevused ära abstraheerida, säilitavad ja täpsustavad need mudelid valdkondlikke erinevusi igal tasandil, tagades, et valdkonnateadmised jäävad alles. Seda spetsialiseerumist kontrolliti juhendamata klasterdamismeetoditega t-SNE, PCA ja k-means, mis näitasid valdkonnapõhiste esituste selget eristumist kõigis kihtides.

- **Universality of Multi-Domain Specialization**: Sarnaselt mitmekeelsele abstraktsioonile leiti, et *mitut tekstivaldkonda hõlmav spetsialiseerumine* on universaalne muster erinevate mudelite ja treeningeesmärkide puhul. Spetsialiseerumine mudelites säilis erinevate keelepaaride ja treeningandmestike suuruste puhul, tugevdades selle rakendatavust mitme tekstivaldkonna kontekstis.

Peatükis 4 käsitlesime vastuväiteid peamiste järelduste kohta ja esitasime näite tõlgendatavuse kohta käivate järelduste praktilisest rakendamisest:

- **Conflicting Literature on Multilingual Abstraction**: Varasemates uuringutes esitati erinevaid seisukohti selle kohta, kuidas mitmekeelsed mudelid töötlevad lausete esitusi, eelkõige seoses ahendusstrateegiatega. Käesolevas töös käsitleti neid vastuolusid, analüüsides ahendusmeetodeid ja näidati, et *mean-pooling* annab täpsema esituse, kinnitades mitmekeelse abstraktsiooni mustrit ja lepitades vastuolulist kirjandust.
- **Outliers in Multilingual Abstraction**: Mõned keeled nagu *Swahili* ja *Urdu* ei vastanud ootuspärastele mitmekeelsete abstraktsioonimustritele. Analüüsides 406 keelepaari, selgus, et piiratud treeningressursid ja kirjaviiside erinevused aitavad tõenäoliselt kaasa nende kõrvalekallete tekkimisele. Kuigi nende keelte näol on tegemist eranditega, ei muuda see vähemtähtsamaks enamiku keelte puhul täheldatud üldisi suundumusi.
- **Conflicting Measurements in Multilingual Abstraction**: Vastuolud sarnasusindeksite vahel, nagu *ANC* ja *CKA*, tekitasid probleeme. See väitekiri näitas, et *kaalutamata mõõdikud*, nagu ANC, on usaldusväärsemad esituste võrdlemiseks mitmekeelsetes mudelites, lahendades vastuolud ja toetades mitmekeelse abstraktsiooni universaalsust.
- **Praktiline rakendus: Domain Adaptation in NMT**: *Mitut tekstivaldkonda hõlmava spetsialiseerumise* kohta tehtud järeldused viisid praktiliste täiustusteni tehisnärvivõrkudel põhinevate masintõlke (NMT) mudelite *tekstivaldkonnale kohandamise* näol. Tavaliselt vajavad valdkonnale kohandamise meetodid valdkondlike klastrite genereerimiseks välist keelemudelit. Käesolevas väitekirjas tehti aga ettepanek kasutada nende klastrite genereerimiseks NMT-mudelit ennast, mis lihtsustab protsessi ja parandab tulemusi kuni *3 BLEU-punkti* võrra dokumenditasemel kohandamisel. See näitab, kuidas tõlgendatavuse tulemusi saab rakendada reaalsete tehisintellekti süsteemide täiustamiseks.

Kokkuvõtteks võib öelda, et käesolev väitekiri täiendab meie arusaamist sellest, kuidas mitmekeelsed ja mitut tekstivaldkonda hõlmavad transformeritel põhinevad mudelid toimivad, paljastades *mitmekeelse abstraktsiooni* ja *mitut tekstivaldkonda hõlmava spetsialiseerumise* järjepidevad mustrid. Väitekiri näitas, et need nähtused on universaalsed erinevate mudelite, skaalade ja ülesannete puhul. Lisaks annab *ANC* metoodika kasutuselevõtt väärtusliku vahendi mitmekeelsete mudelite esituste sarnasuste analüüsimiseks. Käsitledes peamisi vastuväiteid ja esitades praktilise rakenduse NMTs, toob see töö esile tõlgendatavuse uurimise laiema mõju mitmekeelsete ja mitut valdkonda hõlmavate ülesannete jaoks mõeldud tehisintellekti süsteemide täiustamisele.

Selle uurimistöö tulemusena saadud teadmistel on potentiaali muuta tehisintellekt õiglasemaks, turvalisemaks ja kättesaadavamaks, eriti väheste ressurssidega keelte ja spetsiifiliste valdkondade puhul. Mõistes paremini, kuidas need mudelid sisemiselt teavet töötlevad ja esitavad, saame tulevikus luua läbipaistvamaid, usaldusväärsemaid ja paremini kohandatavaid tehisintellekti süsteeme.

6. PUBLICATIONS

CURRICULUM VITAE

Personal data

Full name:	Maksym Del
Date of birth:	02.02.1995
E-mail:	maksym.del@gmail.com

Education

2018 – 2024	Ph.D. in Computer Science, University of Tartu
2016 – 2018	MSc in Computer Science, University of Tartu
2012 – 2016	BSc in Software Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Employment

2023 – ...	Scientific Programmer
2019 – 2023	Junior Research Fellow in Natural Language Processing

Teaching

Spring 2024	Seminar in Mechanistic Interpretability
Spring 2019	Natural Language Processing (teaching assistant)
Fall 2017, 2018	Neural Machine Translation (teaching assistant)

Scientific work

Main fields of research and interest:

- Mechanistic Interpretability
- Large Language Models
- AI Safety
- Multilingual and Multi-Domain Modeling
- Machine Translation

Research grants and scholarships

- Grant to work on artificial intelligence safety research, AI Safety Support Ltd (ABN: 72 653 039 984)

Other administrative and professional activities:

- Reviewed for ICLR, ACL, EMNLP, AACL, EACL, WMT
- ACL membership: 2018, 2019, 2021

ELULOOKIRJELDUS

Isikuandmed

Täisnimi:	Maksym Del
Sünniaeg:	02.02.1995
E-mail:	maksym.del@gmail.com

Haridus

2018 – 2024	Ph.D. Arvutiteaduses, Tartu Ülikool
2016 – 2018	MSc Arvutiteaduses, Tartu Ülikool
2012 – 2016	BSc Tarkvarainseneriõppes, Ukraina Riiklik Tehnikaülikool "Igor Sikorsky Kiievi Polütehniline Instituut"

Teenistuskäik

2023 – ...	Teaduslik programmeerija
2019 – 2023	Loomuliku keele töötlemise nooremteadur

Õppetöö

Kevad 2024	Seminar mehanistlikus interpretatsioonis
Kevad 2019	Loomuliku keele töötlemine (õppeassistent)
Sügis 2017, 2018	Neuromasintõlge (õppeassistent)

Teadustöö

Peamised uurimisvaldkonnad ja huvid:

- Mehhanistlik interpretatsioon
- Suured keelemudelid
- Tehisintellekti turvalisus
- Mitmekeelne ja multidomeenne modelleerimine
- Masintõlge

Uurimisgrantid ja stipendiumid

- Grant tehisintellekti turvalisuse uurimiseks, AI Safety Support Ltd (ABN: 72 653 039 984)

Muud administratiivsed ja professionaalsed tegevused:

- ICLR, ACL, EMNLP, AACL, EACL, WMT retsensent
- ACL liikmesus: 2018, 2019, 2021

DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.
12. **Sulev Reisberg.** Developing Computational Solutions for Personalized Medicine. Tartu 2019, 126 p.
13. **Huishi Yin.** Using a Kano-like Model to Facilitate Open Innovation in Requirements Engineering. Tartu 2019, 129 p.
14. **Faiz Ali Shah.** Extracting Information from App Reviews to Facilitate Software Development Activities. Tartu 2020, 149 p.
15. **Adriano Augusto**. Accurate and Efficient Discovery of Process Models from Event Logs. Tartu 2020, 194 p.
16. **Karim Baghery.** Reducing Trust and Improving Security in zk-SNARKs and Commitments. Tartu 2020, 245 p.
17. **Behzad Abdolmaleki.** On Succinct Non-Interactive Zero-Knowledge Protocols Under Weaker Trust Assumptions. Tartu 2020, 209 p.
18. **Janno Siim.** Non-Interactive Shuffle Arguments. Tartu 2020, 154 p.
19. **Ilya Kuzovkin.** Understanding Information Processing in Human Brain by Interpreting Machine Learning Models. Tartu 2020, 149 p.
20. **Orlenys López Pintado.** Collaborative Business Process Execution on the Blockchain: The Caterpillar System. Tartu 2020, 170 p.
21. **Ardi Tampuu.** Neural Networks for Analyzing Biological Data. Tartu 2020, 152 p.

22. **Madis Vasser.** Testing a Computational Theory of Brain Functioning with Virtual Reality. Tartu 2020, 106 p.
23. **Ljubov Jaanuska.** Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification. Tartu 2021, 192 p.
24. **Arnis Parsovs.** Estonian Electronic Identity Card and its Security Challenges. Tartu 2021, 214 p.
25. **Kaido Lepik.** Inferring causality between transcriptome and complex traits. Tartu 2021, 224 p.
26. **Tauno Palts.** A Model for Assessing Computational Thinking Skills. Tartu 2021, 134 p.
27. **Liis Kolberg.** Developing and applying bioinformatics tools for gene expression data interpretation. Tartu 2021, 195 p.
28. **Dmytro Fishman.** Developing a data analysis pipeline for automated protein profiling in immunology. Tartu 2021, 155 p.
29. **Ivo Kubjas.** Algebraic Approaches to Problems Arising in Decentralized Systems. Tartu 2021, 120 p.
30. **Hina Anwar.** Towards Greener Software Engineering Using Software Analytics. Tartu 2021, 186 p.
31. **Veronika Plotnikova.** FIN-DM: A Data Mining Process for the Financial Services. Tartu 2021, 197 p.
32. **Manuel Camargo.** Automated Discovery of Business Process Simulation Models From Event Logs: A Hybrid Process Mining and Deep Learning Approach. Tartu 2021, 130 p.
33. **Volodymyr Leno.** Robotic Process Mining: Accelerating the Adoption of Robotic Process Automation. Tartu 2021, 119 p.
34. **Kristjan Krips.** Privacy and Coercion-Resistance in Voting. Tartu 2022, 173 p.
35. **Elizaveta Yankovskaya.** Quality Estimation through Attention. Tartu 2022, 115 p.
36. **Mubashar Iqbal.** Reference Framework for Managing Security Risks Using Blockchain. Tartu 2022, 203 p.
37. **Jakob Mass.** Process Management for Internet of Mobile Things. Tartu 2022, 151 p.
38. **Gamal Elkoumy.** Privacy-Enhancing Technologies for Business Process Mining. Tartu 2022, 135 p.
39. **Lidia Feklistova.** Learners of an Introductory Programming MOOC: Background Variables, Engagement Patterns and Performance. Tartu 2022, 151 p.
40. **Mohamed Ragab.** Bench-Ranking: A Prescriptive Analysis Approach for Large Knowledge Graphs Query Workloads. Tartu 2022, 158 p.
41. **Mohammad Anagreh.** Privacy-Preserving Parallel Computations for Graph Problems. Tartu 2023, 181 p.
42. **Rahul Goel.** Mining Social Well-being Using Mobile Data. Tartu 2023, 104 p.

43. **Anti Ingel.** Algorithms using information theory: classification in brain-computer interfaces and characterising reinforcement-learning agents. Tartu 2023, 142 p.
44. **Shakshi Sharma.** Fighting Misinformation in the Digital Age: A Comprehensive Strategy for Characterizing, Identifying, and Mitigating Misinformation on Online Social Media Platforms. Tartu 2023, 158 p.
45. **Kristiina Rahkema.** Quality Analysis of iOS Applications with Focus on Maintainability and Security Aspects. Tartu 2023, 182 p.
46. **Ivan Slobozhan.** Studying Online Social Media Engagement in CIS Countries during Protests, Mass Demonstrations and War. Tartu 2023, 81 p.
47. **Nurlan Kerimov.** Building a catalogue of molecular quantitative trait loci to interpret complex trait associations. Tartu 2023, 248 p.
48. **Pavlo Tertychnyi.** Machine Learning Methods for Anti-Money Laundering Monitoring. Tartu 2023, 117 p.
49. **Abasi-amefon Obot Affia.** A Framework and Teaching Approach for IoT Security Risk Management. Tartu 2023, 180 p.
50. **Raimond-Hendrik Tunnel.** Video Game Design and Development Bachelor's Curriculum for Estonia. Tartu 2024, 137 p.
51. **Ahto Salumets.** Bioinformatics analysis of various aspects in immunology. Tartu 2024, 198 p.
52. **Mohammed Abdulhameed Shaif Ali.** Deep Learning Methods for Cell Microscopy Image Analysis. Tartu 2024, 143 p.
53. **Pille Pullonen-Raudvere**. Foundations of Efficient and Secure Algorithm Development for Secure Multiparty Computation. Tartu 2024, 265 p.
54. **Marili Rõõm.** Multiple approaches to learners' success and factors affecting it in computer programming MOOCs. Tartu 2024, 170 p.
55. **Shivananda Rangappa Poojara.** Design and Orchestration of Scalable, Event-Driven Serverless Data Pipelines for Internet of Things (IoT) Applications. Tartu 2024, 172 p.
56. **Hassan Abdulgaleel Hassan Salim Eldeeb.** Empowering Machine Learning Pipelines with Automated Feature Engineering. Tartu 2024, 121 p.
57. **Muhammad Uzair.** Soft decision making for agri-food 4.0. Tartu 2024, 158 p.
58. **Kirill Milintsevich.** Estimation of Depression Level from Text: Symptom-Based Approach, External Knowledge, Dataset Validity. Tartu 2024, 130 p.