Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Emergence of Factual Knowledge in Pretrained Multilingual Language Models

Master's Thesis

Laurin Paech
lpaech@student.ethz.ch

Mrinmaya's Lab
Department of Computer Science
ETH Zürich

Supervisors:
Yifan Hou
Prof. Dr. Mrinmaya Sachan

August 1, 2022

Acknowledgements

I would like to thank my supervisor Yifan for his support and guidance. His feedback and expertise have been very valuable. I would also like to thank Mrinmaya for the support and for the opportunity to work in his lab. Furthermore, I'd like to thank my parents and friends. I would not be here without you.

Abstract

The advent of pretrained language models has led to exceptional progress in natural language processing (NLP) tasks. While most research has focused on monolingual models, an increasing amount has been dedicated to multilingual language models (MLLMs), which exhibit surprisingly good performance on zero-shot cross-lingual tasks (Wu and Dredze, 2019; Hu et al., 2020; Liang et al., 2020). Although these models are not trained on a cross-lingual objective, they appear to create robust language-agnostic representations to share knowledge across languages. Consequently, low-resource languages may benefit from the knowledge of high-resource languages alleviating their lack of data and even surpassing their monolingual performance. While existing work (Libovický et al., 2019; Pires et al., 2019; K et al., 2020) has investigated these language-agnostic representations, the interactions between languages are not fully explored yet. They do not consider how factual knowledge can emerge through language interactions. In particular, factual knowledge could be shared across languages or inferred anew by sharing symbolic rules.

This work investigates the emergence of factual knowledge in pretrained MLLMs. More concretely, we conduct a study to explore factual knowledge sharing and symbolic reasoning in a zero-shot cross-lingual setting. For this, we investigate (i) how much these models depend on a shared representation when probing for factual knowledge and (ii) the ability to use symbolic reasoning across languages to infer factual knowledge not seen explicitly during pretraining. Our experiments demonstrate that while some factual knowledge is shared across languages, sharing heavily depends on key factors. More specifically, we identify lexical overlap, word order, and alignment of language-specific subspaces. The latter can be achieved by using language-agnostic entities functioning as anchor points or through the use of parallel corpora. Furthermore, we find that monolingual BERT is also capable of learning facts from other languages. Cross-lingual symbolic reasoning works well for some rules while others are not learned consistently. Further analysis shows that the model can even combine symbolic reasoning with factual knowledge transfer to infer new knowledge in a target language. Overall, pretrained MLLMs appear to still exhibit systematic deficiencies in their language-agnostic representations, but we find that these can be partially alleviated.

Contents

Chapter 1

Introduction

The rise of pretrained language models such as BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019) and RoBERTa (Liu et al., 2019) has led to significant improvements in NLP, achieving state-of-the-art results on tasks such as named-entity recognition (NER), natural language inference (NLI) and question answering (QA). These transformer models utilize a self-attention mechanism to create contextual word representations, i.e. the representation of a word in the context of the words around it. They are pretrained on large corpora of unlabeled text to acquire a general language understanding and are subsequently fine-tuned with task-specific data on a downstream task.

Their exceptional performance has motivated researchers to explore what language models store in their parameters. These studies find that language models possess a surprising amount of linguistic and factual knowledge (Goldberg, 2019; Petroni et al., 2019). Pretrained language models have often been limited to only one language, mostly English. This restricts accessibility, diversity and can lead to language biases. In contrast, multilingual language models (MLLMs) like multilingual BERT (mBERT) and XLM-R (Conneau et al., 2020) are trained on many languages at once and can thus overcome language barriers and alleviate these issues. In particular, these models seem to exploit similarities in structure between languages to build a combined representation space (Pires et al., 2019; Wu and Dredze, 2019) and enable cross-lingual transfer - the ability to transfer knowledge learned in one language to another. This generalization ability enables low-resource languages to benefit from information in high-resource languages and mitigate their lack of data. Previous work has already shown promising results of cross-lingual transfer for syntactic and semantic tasks such as POS tagging and NLI (Lauscher et al., 2020). A better understanding of the interplay of languages could close the gap in performance between monolingual and multilingual models but language-neutral representations are not fully understood yet. While factual knowledge has been explored in MLLMs to some extent (Jiang et al., 2020a; Kassner et al., 2021), their cross-lingual transfer has not. This could give us an insight into how these models create and share factual knowledge during pretraining. For example a MLLM that is trained on french

news articles could be capable of recalling facts in those articles in other languages. This could drastically reduce the access to information that is currently only available in a few languages.

In this thesis, we investigate the emergence of factual knowledge in pretrained MLLMs. Factual knowledge could be shared across languages through language-agnostic representations. Therefore we investigate how much pretrained MLLMs depend on a shared representation when being probed for factual knowledge. For this we train mBERT on unseen facts in a source language and evaluate the model in a target language. We analyse the various factors that influence the zero-shot cross-lingual transfer. Besides mBERT, we also analyse the ability of monolingual BERT to learn from facts in other languages. Since some facts might have emerged through symbolic reasoning during pretraining, we analyse if these symbolic rules (e.g. symmetry, equivalence, implication) are also shared across languages and used to infer new factual knowledge. For this we extend Kassner et al. (2020) to a zero-shot cross-lingual setting. Finally, we study the combined application of factual knowledge transfer and symbolic reasoning. We pose the following questions: How much factual knowledge can be shared across languages in MLLMs? What are the factors that influence factual knowledge transfer? Can factual knowledge emerge through symbolic reasoning across languages? This - to the best of our knowledge - has not been done before.

Our experiments demonstrate that (i) while limited, mBERT is capable of zero-shot cross-lingual factual knowledge transfer, even if languages are typologically distant. We identify lexical overlap, matching word order, and especially alignment as important factors for successful transfer between language-pairs. We discover that deficiencies in alignment can be partially overcome by training with a small parallel corpus or with multiple source languages. (ii) We find that monolingual BERT is capable of zero-shot cross-lingual factual knowledge transfer as well. This is further evidence that monolingual BERT constructs partial language-neutral representations. (iii) mBERT has the ability to create new factual knowledge not seen during training by reasoning with symbolic rules on memorized facts in other languages. In particular, mBERT can share symbolic rules across languages and even combine symbolic reasoning with factual knowledge transfer to infer new facts. The code for training and data generation is available at: `https://github.com/laurinpaech/emergence-factual-knowledge`.

The thesis is organized as follows: First, we give an overview on related works and how this motivated our work in Chapter 2. In Chapter 3 we discuss our experiments regarding factual knowledge transfer for mBERT and monolingual BERT and analyse the various factors that influence the transfer. In Chapter 4 we study symbolic reasoning and the ability of mBERT to utilize different symbolic rules. Furthermore, we investigate if symbolic reasoning can be combined with factual knowledge transfer by the model. Finally we summarise our results and give an outlook on what future work might entail in Chapter 5.

Chapter 2

Related Work

The introduction of transformer models (Vaswani et al., 2017) have been described as NLPs ImageNet-moment, referring to the ImageNet challenge that showed the viability of deep learning models in computer vision which led to enormous improvements in performance. The success of transformer models and their human-like performance have raised many ethical concerns about misuse and biases. Numerous studies have been devoted to understand their inner-workings and limitations.

In this chapter, we will give an overview on what has been done so far with regards to the understanding of language models, multilingual embeddings, captured relational knowledge, reasoning and how this motivated our work.

Language Models

Since its introduction BERT (Bidirectional Encoder Representation from Transformers) (Devlin et al., 2019) has become the *de facto* standard for language modeling. By pretraining on large amounts of unlabeled text data using the masked language modeling (MLM) and next-sentence prediction objectives, BERT builds a general language understanding and learns to create contextual representations of its input. The model is then extended with additional layers and fine-tuned on task-specific labeled data to create state-of-the-art models on a wide range of tasks. Additionally, Devlin et al. introduce multilingual BERT (mBERT), a variant of BERT that was trained on the Wikipedia corpus of 104 languages, that exhibits remarkable cross-lingual capabilities without using a multilingual objective. Later Conneau et al. (2020) improved on mBERT's design and introduced XLM-RoBERTa (XLM-R). In contrast to mBERT, the goal of XLM-R was to improve cross-lingual language understanding and create a multilingual model that does not sacrifice the performance of individual languages. For this they introduce the so-called *curse of multilinguality* - a tradeoff between the amount of languages, their performance and model capacity. More languages lead to a better cross-lingual performance until performance starts to degrade. By increasing the model capacity XLM-R achieves significant improvements in monolingual and

cross-lingual performance. As a result, XLM-R outperforms mBERT on several tasks and is even competitive with RoBERTa on monolingual benchmarks.

Factual Knowledge

But how do these models achieve such incredible results? Much work has focused on the knowledge captured by pretrained language models (PLMs) during training. These models seem to store syntactic and semantic and even commonsense and factual knowledge encoded in their parameters. In this work we are focusing on factual knowledge.

Petroni et al. (2019) introduce LAMA probe and investigate how much factual knowledge is stored in PLMs. The model is probed with "fill-in-the-blank" cloze-like statements such as "Paris is the capital of `[MASK]`", where the object of the sentence is replaced by a masked token, which has to be inferred by the model. They find that BERT stores large amount of factual knowledge and its ability to recall is even on-par with traditional knowledge bases. Consequently, language models could, so they hope, be used as an alternative to knowledge bases which are costly to maintain. Kassner and Schütze (2020) examine negation and mispriming for LAMA probe and find that PLMs have trouble differentiating between positive and negative sentences. Additionally, PLMs are easily distracted by mispriming. They come to the conclusion that PLMs have trouble with negation because the training corpus consists primarly of positive sentences. Inappropriate prompts are limiting knowledge retrieval as they can always only give a lower-bound on the knowledge retrieved. Jiang et al. (2020b) propose methods to automatically improve prompts to get a higher lower-bound on factual knowledge of LMs.

Kassner et al. (2021) extend LAMA to a multilingual setting and investigate to what extent mBERT can be used as a knowledge base. They come to the conclusion that mBERT is not storing entity knowledge in a language-agnostic way. They support this claim with the performance gap between different languages when probed for knowledge and a systematic bias for the queried language. However, this seems to be an over-generalization. A performance gap between languages reveals only that not all knowledge is uniformly accessible but does not exclude that some is shared through a common language-neutral representation. As they have found, identical knowledge across languages does exist. Contemporaneously Jiang et al. (2020a) create the Crosslingual FACTual Retrieval benchmark (X-FACTR) to asses the amount of factual knowledge stored in multilingual language models. They expand probing from single-token to a multi-token setup and introduce various multi-token decoding algorithms. Further, they design a annotation schema to adjust prompts to the correct morphology. Additionally, they come up with a code-switching objective that replaces entity mentions in one language with those of another to improve performance. Similar to Kass-

ner et al. (2021), they find that overall factual knowledge retrieval is low but greater in high-resource languages compared to low-resource languages. While Jiang et al. (2020a) and Kassner et al. (2021) investigate factual knowledge in MLLMs, both do not consider that MLLMs might be able to share factual knowledge across languages. Thus an analysis of MLLMs and their factual knowledge is incomplete.

Factual knowledge in PLMs can emerge in two forms: either through memorization or reasoning on already memorized knowledge to create new knowledge. While BERT has shown good reasoning abilities on common reasoning tasks such as NLI and commonsense reasoning, one flaw of these tasks is that all relevant facts are available during inference time. Models need to be capable of handling cases of implicit knowledge in the form of knowledge stored in the parameters and explicit knowledge, that is natural language input. Talmor et al. (2020) choose symbolic reasoning tasks to study the reasoning capabilities of various PLMs. The authors find that different language models have different reasoning abilities e.g. RoBERTa can compare numbers while BERT can not. They hypothesize that this is due to different pretraining corpora and training objectives. Limitations also seem to occur due to being trained on co-occurences as their reasoning is context-dependent and not abstract. For example, RoBERTa's ability to compare numbers is restricted to human age ranges. Kassner et al. (2020) train BERT from scratch on synthetic corpora to investigate how factual knowledge emerges through memorization and symbolic reasoning. The authors compare their approach to link prediction in the knowledge base domain and while the model is capable of learning some symbolic rules it seems to overgeneralize when applying them. Our work extends this line of research to a multilingual setting. In particular, we investigate if MLLMs share factual knowledge across languages via zero-shot cross-lingual transfer and if they can apply symbolic reasoning on already memorized facts to infer new facts in other languages.

Cross-lingual representations

Many studies have shown the impressive performance of MLLMs on zero-shot cross-lingual transfer - the ability to fine-tune on task-specific data in a source language and subsequently evaluate on the same task in a target language (Pires et al., 2019; Wu and Dredze, 2019). Recent work has found evidence that these models seem to carry not only language-specific but even language-neutral representations, which might explain their surprisingly good performance (Pires et al., 2019; Libovický et al., 2020; Hu et al., 2020). However, the architectural and linguistic properties that contribute to language-neutral representations and hence facilitate cross-lingual transfer have been the subject of much debate.

At first, it has been speculated that mBERT is only exploiting lexical overlap between typologically similar languages to solve these tasks. Pires et al. (2019)

discard this idea and report that mBERT is even capable of zero-shot cross-lingual transfer when source and target languages are in different script, i.e. have no overlap. Although the impact of lexical overlap is rather negligble, several works find that an overlap of subwords slightly increases performance (Wu and Dredze, 2019; K et al., 2020; Wu et al., 2020). The linguistic properties most important are a comparable, but not necessarily parallel, training corpus (Dufter and Schütze, 2020) and structural similarity (K et al., 2020). In particular, multiple works (Pires et al., 2019; Dufter and Schütze, 2020; K et al., 2020) find that transfer is significantly less effective when both source and target language have a different word order (e.g. SVO and SOV). Further, Pires et al. (2019) hypothesize that mBERT generalizes across languages by sharing similar word pieces in all languages (e.g. URLs, numbers) that are mapped to a common space. As a result, co-occurring words are mapped close to the common space and this aligns the language-specifc subspaces into a language-agnostic space. Similarly, Dufter and Schütze (2020) speculate that a limited capacity forces the model to exploit common structures among languages and shared special tokens might function as anchor points, which Wu et al. (2020) define as "identical strings that appear in both languages in the training corpus". The experiments of Dufter and Schütze (2020) show that multilinguality arises late during training when the model already starts to overfit. This motivated us to fine-tune the model for our experiments instead of training from scratch.

Libovický et al. (2019) explore mBERT's representations directly on semantic tasks such as sentence retrieval, word alignment and MT quality estimation. They find that representations can be divided into a language-specific and a language-neutral components. Both Pires et al. (2019) and Libovický et al. (2019) discover language-identity as constant shift in the embedding space and centering embeddings largely removes language-specific information. An alternative approach is to project the embeddings on a small amount of parallel data (Libovický et al., 2020). K et al. (2020) analyse architectural properties and conclude that depth and total number of parameters are most important, while the number of attention heads is negligible. Both Wu et al. (2020)'s and Wu and Dredze (2019)'s results suggest that especially bottom layers seem to play an important role for cross-lingual transfer and the freezing of bottom layers during fine-tuning improves performance. According to Tanti et al. (2021), fine-tuning causes a reorganization of the representational space of mBERT such that the proportions of the language-specifc and language-neutral components depend on the downstream task. A limitation of cross-lingual transfer is the performance gap between high- and low-resource languages. Lauscher et al. (2020) analyse how to close this gap. They find that, especially on low-resource languages, few-shot learning, even with small numbers of annotated target data, can rapidly improve the performance. They also show that zero-shot transfer performance has high correlation with similarity of task-relevant language features of source and target language. For instance, in a syntactic tasks such as POS tagging, high similar-

ity of syntactic features increases performance. Pretraining corpus size highly correlates with good performance on higher-level tasks such as NLI and QA. Hu et al. (2020) and Liang et al. (2020) introduce novel benchmarks for zero-shot cross-lingual transfer, XTREME and XGLUE respectively. Hu et al. evaluate the cross-lingual generalization of multilinugal language models on 9 syntactic and semantic tasks spanning 40 languages and find that XLM-R performs often better than mBERT. Similarly, Liang et al. introduce 11 tasks covering NLU and language generation.

Wu et al. (2020) find that independent monolingual BERT models exhibit similar structures. They hypothesize that in multilingual language models the monolingual spaces are automatically aligned during pretraining. Hence partial parameter sharing, especially in the bottom layers, is sufficient for the emergence of multilingual representations. Artetxe et al. (2020) provide further evidence that a shared vocabulary and even joint training is not necessary. The authors transfer monolingual models to another language by learning multilingual embeddings while keeping the rest of the models parameters frozen. The transferred monolingual model is highly competitive on zero-shot cross-lingual transfer tasks. They conclude that monolingual models seem to learn general language abstractions. de Souza et al. (2021) take these ideas further by directly fine-tuning monolingual models on a task in a foreign language. These models achieve competitive performance without the need for parallel corpora, suggesting that monolingual models develop a language-neutral representation space, possibly due to the MLM objective. Following this line of work, we also investigate factual knowledge transfer in monolingual models.

Chapter 3

Factual Knowledge Transfer

While existing work has focused on the zero-shot cross-lingual transfer of syntactic and semantic knowledge, we investigate the ability of MLLMs to transfer memorized factual knowledge across language representations. Factual knowledge memorized in one language during training could be available in another it was previously not seen in. We hypothesize that by memorizing a fact in a source language, it also becomes available in other languages since they might share an underlying language-agnostic representation.

In this chapter, we first give an overview of the experimental setup, where we discuss data generation, pre-processing, language selection and model selection. Next, we discuss our experiments on facts with multilingual entities, followed by language-agnostic entities. The following sections analyse the impact of individual relations and the frequency of entities in the training set. Then we study zero-shot cross-lingual factual knowledge transfer on monolingual BERT. Last, we discuss our results and the limitations of our experiments.

3.1 Experimental Setup

Following the LAMA setup, **facts** consist of entities $e, f \in E$ and relations $r \in R$ and are constructed as entity-relation-entity-triples: (e, r, f) that can also be seen as (*subject*, *relation*, *object*) statements. Each fact is converted to a cloze-like statement by masking the object for evaluation. If the model correctly predicts the object, it is said to know the fact. In contrast to prior work, where the model is probed for already existing factual knowledge acquired during pretraining, we fine-tune the model on facts in a **source language** and evaluate the performance of zero-shot cross-lingual transfer on corresponding facts in a **target language**.

Data generation

We want to work with new, unseen facts to investigate how cross-lingual factual knowledge emerges during training. Using facts that were possibly already seen

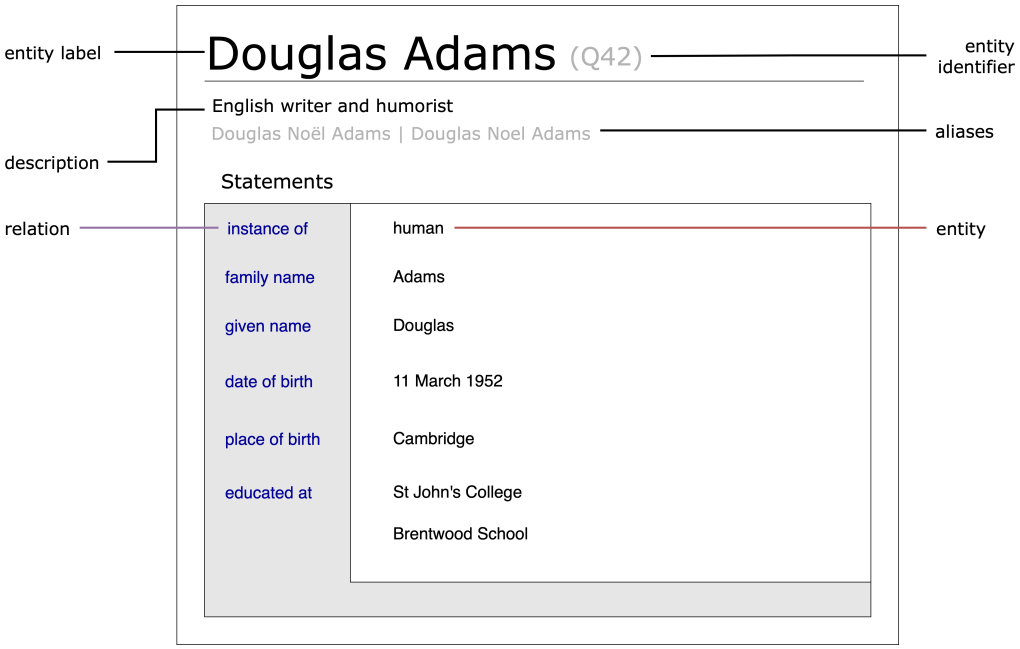

Figure 3.1: Example of Wikidata entity (Douglas Adams). Entity forms statement triples with its relations and other entities.

would bias our results. Therefore facts in both languages are created by sampling relations r and entity-pairs (e, f) from a synthetic dataset. Using a synthetic corpus has the benefit that synthetic facts are unlikely to have appeared in the original training set since PLMs are trained on natural occurring text data such as Wikipedia or Common Crawl. Lastly, synthetic facts easily scale to large amounts of data. The synthetic dataset consists of entities and relations from Wikidata[1] - a free, open and cooperative multilingual knowledge base that is composed of the structured data of Wikipedia (Vrandečić and Krötzsch, 2014). An example of an entity and its corresponding relation-entity pairs is illustrated in Figure 3.1. We were provided with a pre-cleaned in-house version of the Wikidata dataset. It consists of a subset of all entities with their respective IDs and labels and a subset of entity-relation-entity triples.

Entities can be divided into language-agnostic, which have the same label across all languages (e.g. numbers and URLs), and language-specific or multilingual, which vary labels from language to language (e.g. "chair" in English but "Stuhl" in German). However, for our experiments we define language-agnostic entities more broadly as all entities that have the same label across a set of languages (e.g. English and German) in Wikidata. These are mostly names in Latin-script. We distinguish between language-agnostic and multilingual facts depending on which entities are used (Fig. 3.2). Relations are always language-

[1] www.wikidata.org

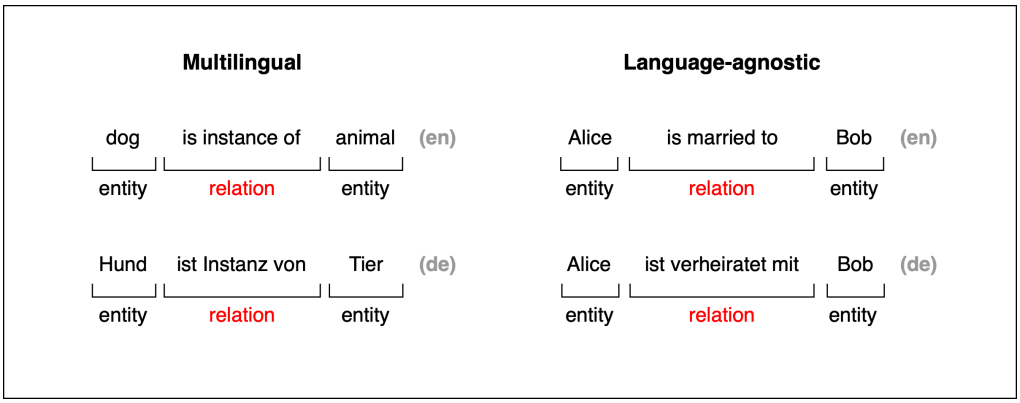

Figure 3.2: Examples of multilingual and language-agnostic facts in English and German with highlighted entities and relations.

specific and link a subject-entity with an object-entity. They range from single words like "occupation" to templates like "is educated at" missing subject and object. Any grammatical errors that may arise when creating facts such as wrong inflection are not corrected for. This might be a limitation since proper inflection gives the model hints about what to predict. However, since Kassner et al. (2021) compared their machine-translated queries with manually corrected ones from Jiang et al. (2020a) and found only marginal performance difference, we assume it will give comparable results. We only consider entities that consist of a single-token since multi-token entities pose issues with selection, decoding and parameter-tuning (Petroni et al., 2019; Jiang et al., 2020a). As additional requirement to reduce the probability of creating already seen facts, entities need to have low-connectivity in the knowledge base, i.e. we select entities that have as few relations with other entities as possible. After collecting entity candidates, they are manually cleaned to remove inappropriate words, non-nouns (e.g. adjective, verbs), single-letter entities (e.g. P, L, K), duplicates, numbers and single Unicode characters. Similarly, candidate relations are manually cleaned to remove time, date, ratings, chemistry or Wikidata-internal relations. Further we remove any duplicates and focus on selecting relations that are as close to natural language as possible. Finally, we have a total of 5913 language-agnostic entities and 712 multilingual entities in English, German, French and Spanish. For other language-pairs the number of entities range from 556 in English-Japanese to 115 for Russian-Chinese with language-pairs in Non-Latin-script having the lowest number of shared entities. In contrast, the number of relations are much closer from 795 for languages in Latin-script to 508 for Russian-Japanese.

Language Selection

Even though Wikidata is multilingual, labels of entities and relations are rarely available in all languages. We select seven languages: English, German, French, Spanish, Russian, traditional Chinese and Japanese. These languages are typologically diverse, belonging to a variety of language-families. Importantly, they also have sufficient entities and relations available to sample from. Furthermore, the languages also need to share enough entities and relations with other languages to construct the synthetic dataset. Traditional Chinese is chosen over simplified Chinese as more entities and relations are available. We are also going to refer to these languages by their ISO code. Further we will refer to Indo-European-languages in Latin script (English, German, French and Spanish) as "Latin-European languages" or simply "Latin languages". Note that low-connectivity filtering for entities only works for Latin-European languages as Japanese, Chinese and Russian have too few entities in Wikidata. Instead we run a LAMA-like probe before each experiment to verify that the facts are unknown by the model. More concretely, we test that the object-entity is not in the top 5 of predicted entities for that fact. As part of their ablation study, Liang et al. (2020) investigate the choice of source language, also referred to as pivot language, in zero-shot cross-lingual transfer and find that, while often used, English is not always the optimal choice. This motivated us to not only create English-centric datasets but pairwise datasets between all Latin-European languages. Since these are all in the same script, we have multilingual and language-agnostic entities available. Russian, Chinese and Japanese have less entities and relations available, especially in pairs with other languages. Therefore we create English-centric datasets, i.e. English-Russian, English-Chinese and English-Japanese, and any language-pair permutation among them (e.g. Japanese-Russian, Chinese-Japanese, Russian-Chinese).

Model Details

Unless stated otherwise, our experiments are all run on multilingual cased BERT[2] provided by HuggingFace[3] (Wolf et al., 2020). For tokenization we use mBERT's WordPiece tokenizer. Using Wikidata entities and relations has the benefit that mBERT has already seen them individually during pretraining and possibly build language-agnostic representations of them. As Dufter and Schütze (2020) find that multilingual representations arise quite late in the training process of multilingual language models, pretraining mBERT from scratch is not only computationally expensive but also unnecessary. Instead we fine-tune the model with a masked language modeling (MLM) objective retaining the MLM head used during pretraining and minimize the cross-entropy loss. Therefore, the fine-tuning

[2] *bert-base-multilingual-cased*
[3] `https://huggingface.co`

can be seen as an extension of the pretraining process, giving an insight into the emergence of factual knowledge during pretraining and the additional benefit of testing already existing language-agnostic representations for cross-lingual factual knowledge transfer. Hyperparameters are selected by searching over ranges on the respective validation sets. Epochs {50, 100, ..., 250, 300}; batch-size {64, 128, 256, 512}; learning-rate {4e−5, 5e−5, 6e−5}. Other parameters are kept as is. The high number of epochs are chosen since we want the model to memorize the facts in the source language, i.e. know them, to investigate if it is able to transfer them to a target language. We also fix the randomness for reproducability (default seed 42). Contrary to the common zero-shot setting, where the validation set is chosen in the source language, we use a validation set in the target language. This is due to the experiments evaluating the transfer of memorized knowledge, not the memorization itself. We take 10% of our test set as validation set. As evaluation metric we use mean Precision@k (mP@k), unless stated otherwise. For a given fact Precision@k is 1 if the object is predicted in the top k ranks of the results. The model is trained on 2 NVIDIA 1080 Ti GPUs with 11GB of memory and if experiments exceed memory capacity, trained on a single NVIDIA RTX A6000 GPU with 48GB memory. Training takes usually 2-3 hours (or $<$ 1 hour respectively), depending on the experiment.

3.2 Multilingual Factual Knowledge Transfer

As previously stated we distinguish between multilingual and language-agnostic facts depending on the type of entity used. We first discuss our experiments with multilingual facts, written in a source and target language that possibly do not share entity-labels or even a script. The model has to map the memorized fact in the source language to the target language:

$$(e_{\text{source}}, r_{\text{source}}, f_{\text{source}}) \implies (e_{\text{target}}, r_{\text{target}}, f_{\text{target}})$$

This is especially difficult with multilingual entity labels since not only is the relation different in source and target language, the entities are too. That means that the model needs to be able to map a fact in the source language to an equivalent representation in the target language. In particular, the model needs to be able to recognize that the context in both languages is equivalent and predict the object accordingly. Pires et al. (2019) and Dufter and Schütze (2020) hypothesize that multilingual language models build language-agnostic representations due to sharing word pieces in all languages that are mapped to a common space due to their limited capacity. As a result, co-occurring word pieces are mapped close to the common space and this aligns the language-specific subspaces into a language-agnostic space. We hypothesize that MLLMs can utilize the language-agnostic space for multilingual factual knowledge transfer.

Source	en			de			fr			es		
Target	de	fr	es	en	fr	es	en	de	es	en	de	fr
Transfer	1.2	5.2	1	1.5	0	0	4.2	0	0	1	0	0

Table 3.1: Mean precision at one (mP@1) of zero-shot factual knowledge transfer with multilingual entities from a source to a target language on mBERT.

3.2.1 Initial Experiments

For training and test set, we sample n relations and for each relation m entity-pairs. Since we find that not all relations and entities perform equally well, we choose $n = 10$ and $m = 1000$ which provides a diverse set of relations and entities and is comparable to Kassner et al. (2020)'s setup. This is a total of $10,000$ facts in source and target language and the training can be done in a reasonable time. We fine-tune mBERT on the facts in the source language using the MLM objective and then probe the model in the target language. For the hyperparameter search in a given source language, we combine multiple target languages into a single validation set to mitigate language biases. Note that this is only feasible for Latin-European languages since they share enough entities and relations with each other. However, we usually use the same parameters across all languages.

The results can be found in Table 3.1. Note that we omit any language-pairs with Russian, Japanese and Chinese as all transfers were zero, possibly due to not sharing a common script. As the results show, there is barely any sharing across languages with multilingual entity labels. Even for languages that are typologically close such as English, German, French and Spanish, the model is mostly unable to transfer knowledge, although mBERT is clearly able to memorize the training data with a training accuracy of almost 100%. The best transfer is English-to-French and French-to-English with 5.2% and 4.2% mean Precision@1 respectively. Next, we discuss further improvements and alternative approaches.

Refinements

Since the model was trained on natural language, its contextual representations might be sensitive to improper grammar. Facts that have grammatical errors, might lead to worse contextual representations and therefore to worse transfer performance. As most of our facts do not reflect natural language and relations are usually just single words, we hypothesize that this is partially responsible for the bad performance. To test this, we selected 10 relations that provide templates that are close to natural language and repeat our experiments. Unfortunately this does not lead to performance improvements and strengthens our choice of

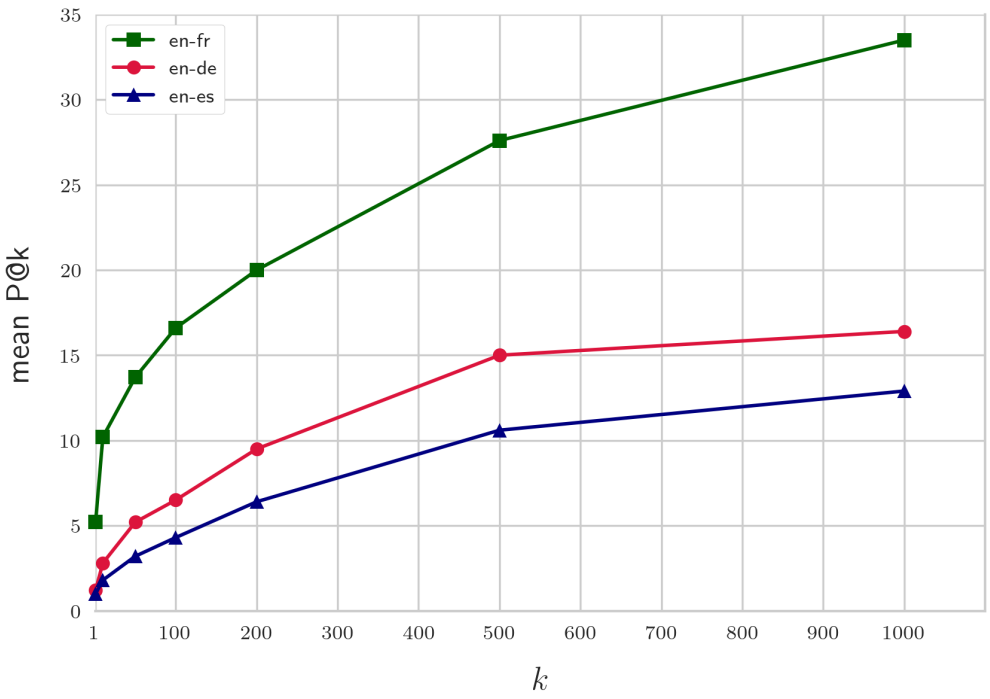

Figure 3.3: Mean P@k curve with varying k $\{10, 50, 100, 200, 500, 1000\}$ for zero-shot transfer from English-to-French (en-fr), English-to-Spanish (en-es) and English-to-German (en-de) using multilingual entities.

using a synthetic dataset. Next, inspired by Jiang et al. (2020a), we tried to improve multilingual performance by code-switching. With a probability of 30% we switch the source-language entities with target-language entities. Since the model could learn the co-occurences of entities, we remove the code-switched facts from the test-set. Code-switching could improve multilingual alignment and therefore improve performance. However, we find that code-switching did not impact performance at all. We also suspected that the poor performance may be due to mBERT's lack of robust multilingual representations and repeat the experiments on XLM-R. Surprisingly, performance was equally poor.

Next, we investigated the effect that the number of facts (or entity-pairs) and relations have on transfer performance. We found that by decreasing the number of relations from 10 to 5, the model can improve its performance for English-to-French from 5.2% to 8.6%. Reducing facts seems to have either no effect or decrease performance. More facts with a relation make it possibly easier to build stronger representations. Similar results were observed for other language-pairs. Although this could be because poor performing relations are removed, we find this to be consistent in other experiments with different relations as well. As multilingual factual knowledge transfer appears to be quite challenging, we

suspect that our metric P@k with $k = 1$ is too ambitious. For this we run selected experiments again with different k $\{10, 50, 100, 200, 500, 1000\}$. The mean P@k curve for English-to-French (en-fr), English-to-Spanish (en-es) and English-to-German (en-de) can be found in Figure 3.3. Most notably, P@10 almost doubles the performance from 5.2% to 10.2% for English-to-Spanish. Further increasing k only leads to marginal percentage gains, leading to the logarithmic shape of the curve. Additionally we tried using *Mean Reciprocal Rank* (MRR) which is defined as the mean of the reciprocal of the rank of the object that is to predicted. We used a cut-off at rank 1000. Using MRR improves English-to-French performance to only 6.4% which is a better reflection of the general performance than a high k.

3.2.2 Multiple source languages

Another hypothesis we had was that seeing a fact in more than one source language could aid cross-lingual transfer. For this we increased the number of source languages to three. Unfortunately, Russian, Japanese and Chinese do not share enough relations and entities with other languages, so our experiments are restricted to Latin-European languages. Indeed we find that cross-lingual knowledge sharing slightly improves by adding more languages (Table 3.2). The best single source language transfer to English increases from 4.2% with only French as source language to 6.5% with three source languages. Transfer to German and Spanish barely increase. Surprisingly, transfer to French slightly decreases from 5.2% with English to 4.7% when German and Spanish are added.

Source Target	de-fr-es en	en-fr-es de	en-de-es fr	en-de-fr es
Baseline	4.2	1.2	5.2	1
Transfer	6.5	1.5	4.7	1.2

Table 3.2: Mean P@1 of zero-shot cross-lingual factual knowledge transfer with multilingual facts using multiple source languages. The baseline is the best single source transfer for the target language.

Few-Shot Learning / Parallel Corpus

Prior work has shown that having a parallel corpus improves language-neutrality and can help align representations (Conneau and Lample, 2019; Libovický et al., 2020). Dufter and Schütze (2020) even suggest that when no parallel corpus is available, the training corpora in different languages should at least be comparable. Inspired by this line of research, we investigate the impact of few-shot learning. For every relation, n target facts are removed from the test set and

added to the training set. By that, few-shot learning refers to having a partial parallel corpus that could potentially help alignment between languages. While we have found no effect on the performance of single source languages, transfer performance for multiple source languages increased considerably. Our results are shown in Table 3.3. The transfer of German-Spanish-French to English almost triples from 6.5% to 17.9% with only 10 parallel facts. These results indicate a lack of alignment that is resolved by adding parallel data to the fine-tuning process. Interestingly, the model does not increase performance linearly when more facts are added. Instead the highest increase comes from only 10 facts in our experiments.

Source Target	de-fr-es en	en-fr-es de	en-de-es fr	en-de-fr es
0-shot	6.5	1.5	4.7	1.2
10-shot	17.9	6.6	11.3	5.9
25-shot	21.5	**9.3**	**14.6**	5.6
50-shot	21	7.5	14.2	7.7
100-shot	17.6	5.7	12.5	6.6
250-shot	**24.4**	7.7	16.2	**10.2**
Highest Increase	+17.9	+7.8	+9.9	+9

Table 3.3: Mean P@1 of n-shot learning for zero-shot factual knowledge transfer with multiple source languages. n is defined as the number of facts in target language added to training set.

Data reduction & Alternative Metrics

Similar to the single source language experiments, we find that reducing the number of relations increases the performance slightly. When evaluating the choice of metric we find again that the highest percentage gain in performance is from P@1 to P@10, specifically for transfer to English from 6.5% to 16.8%. We combined few-shot learning and the change in metric for multiple source languages to get an additive performance boost. In particular, we reduce the number of relations from 10 to 5, add a small parallel corpus of 50 facts and use P@10 as evaluation metric. We report our results in Table 3.4. For the transfer from German-French-Spanish to English, we see an increase of 11% on few-shot learning to a total of 32.3%. For the transfer from English-French-Spanish to German the model achieves a total transfer of 14.2%.

Source Target	de-fr-es en	en-fr-es de	en-de-es fr	en-de-fr es
Transfer	6.5	1.5	4.7	1.2
Transfer + Refinements	32.3	14.2	25.7	15.4

Table 3.4: Comparing zero-shot factual knowledge transfer on multiple source languages. Baseline transfer uses mean P@1. Refinements combine mean P@10, 5 relations and 50 parallel facts for additive increase.

3.2.3 Combine Multilingual and Language-agnostic Entities

In our experiments we found that facts that have language-agnostic entities have especially good performance. In this section, we discuss experiments on facts that have language-agnostic entities for either subject or object, while the other is multilingual, e.g. "Emmanuel is the president of France" with "Emmanuel" as language-agnostic entity and "France" as multilingual entity:

$$(e, r_{\text{source}}, f_{\text{source}}) \implies (e, r_{\text{target}}, f_{\text{target}})$$

$$(e_{\text{source}}, r_{\text{source}}, f) \implies (e_{\text{source}}, r_{\text{target}}, f)$$

We hypothesize that language-agnostic entities could function as anchor points, which Wu et al. (2020) define as "identical strings that appear in both languages in the training corpus". As a result, these anchor points are likely mapped to a language-neutral representation that is shared by all languages and make cross-lingual transfer more accessible.

Single Source Languages

Source	en			de			fr			es		
Target	de	fr	es	en	fr	es	en	de	es	en	de	fr
multilingual	1.2	5.2	1	1.5	0	0	4.2	0	0	1	0	0
aSubject + mObject	5.7	15.8	6.1	**6.5**	**4.0**	2.2	**14.0**	**3.6**	**3.3**	7.3	2.2	3.6
mSubject + aObject	**6.0**	**17.6**	**8.3**	4.9	3.0	**2.3**	10.6	2.6	3.1	**8.2**	**2.7**	**4.8**

Table 3.5: Mean P@1 of zero-shot cross-lingual factual knowledge transfer with language-agnostic and multilingual entities for single source languages. Prefixes "a" and "m" refer to agnostic and multilingual respectively.

Table 3.5 shows our results combining language-agnostic with multilingual entities for single source languages. Note that we restrict the analysis to Latin-languages as there are no clear language-agnostic entities between languages in different scripts. We will discuss this further in Section 3.3. For a language-agnostic subject and multilingual object (aSubject + mObject), we found that the

Source Target	de-fr-es en	en-fr-es de	en-de-es fr	en-de-fr es
multilingual	6.5	1.5	4.7	1.2
aSubject + mObject	15.5	6.4	9.0	5.3
mSubject + aObject	**31.6**	**10.0**	**23.8**	**12.8**

Table 3.6: Mean P@1 of zero-shot cross-lingual factual knowledge transfer with language-agnostic and multilingual entities for multiple source languages. Prefixes "a" and "m" refer to agnostic and multilingual respectively.

English-to-French transfer increases from 5.2% to almost 16%. Other language-pairs such as German-French and Spanish-French that did not have any transfer before, display it now. For multilingual subjects and language-agnostic objects, we find similar improvements (mSubject + aObject). Further, while transfer from English seems to have increased even more, transfer from German seems to perform better with a language-agnostic subject.

Multiple Source Languages

We repeat our experiments for multiple source languages. As shown in Table 3.6, improvements with language-agnostic subjects are not as large as for some single source languages. In contrast, using a language-agnostic object increases the performance substantially. The transfer performance with English as target language is considerably higher than for single source languages increasing from 6.5% to 31.6% by 25.1% mean P@1. We speculate that this is not only due to using a language-agnostic entity. Instead the language-agnostic object possibly enables the model to fully make use of the additional training data. We hypothesize that predicting a language-agnostic object is easier than a multilingual object because the model needs to only recognize the context without the need for an additional language-specific signal for the object-entity.

3.3 Language-Agnostic Factual Knowledge Transfer

Our previous results motivated us to experiment with only using language-agnostic entities. As previously mentioned, we hypothesize that these likely function as anchor points for mBERT benefiting the zero-shot cross-lingual knowledge transfer. Language-agnostic entities are much easier to predict since subject and object are the same in both source and target language:

$$(e, r_{\text{source}}, f) \implies (e, r_{\text{target}}, f)$$

Note that to discourage learning co-occurrences (i.e. entity pattern-matching), the subject-entities are re-used for every relation. If the model predicts the

Source	en			de			fr			es		
Target	de	fr	es	en	fr	es	en	de	es	en	de	fr
multilingual	1.2	5.2	1	1.5	0	0	4.2	0	0	1	0	0
language-agnostic	28.6	58.6	53.4	26.4	26.4	25.4	60	41.4	54.4	52.6	25.4	54.5

Table 3.7: Mean P@1 for zero-shot cross-lingual factual knowledge transfer with language-agnostic and multilingual facts on Latin-languages.

object only based on a given subject, then its probability to guess the object correctly is $1/n$, where n is the number of relations. Note that the validation set reflects this change as well by splitting along the same subject for every relation. The remaining setup is unchanged. Again we choose $n = 10$ and $m = 1000$ (number of entity-pairs) and use mean Precision@1 as metric. We repeat the transfer experiments with language-agnostic entities for single source languages and report the results in Table 3.7.

In contrast to multilingual entities, the performance for zero-shot cross-lingual transfer with language-agnostic entities is vastly improved. Again, we report a performance gap for using different source languages. Selecting French as source language, mBERT achieves a mean P@1 of 51.2% averaged over all Latin-European languages. English is second with 46.9% above Spanish with 44.2%. German is last with an average transfer performance of 26.1%. Note, that all of our results are greatly above chance, i.e. the model guessing objects based on co-occurrences (10%). While some transfer performances are symmetric such as French-Spanish and English-German, most are not. Transfers from and to German are by far the lowest.

Russian, Chinese, Japanese

Since Russian, Chinese and Japanese are in different script, there are no apparent language-agnostic entities in a common script among them. Instead, we use language-agnostic entities in Latin-script as Latin-words, especially names, can sometimes be found in texts in these languages. The results for English-centric language-pairs and Non-Latin language-pairs are shown in Table 3.8. mBERT achieves the best transfer efficiency from Japanese to Chinese (77%) and vice-versa (72.2%). This might be due to Japanese Kanji using Chinese characters, so they are typologically similar languages. The lowest performance is Japanese-to-Russian with 17.1%. All transfer from and to Russian is surprisingly low. Interestingly, despite English and Russian belonging to the same language family (Indo-European), English-to-Russian transfer performance is lower than with more distant languages like Chinese.

So far we have restricted our analysis to using Latin-entities. We repeat our experiments and investigate the impact of choosing entities in the script of the

Source	en			ru		zh		ja	
Target	ru	zh	ja	zh	ja	ja	ru	zh	ru
Latin Entities	19.7	46.78	24.85	24.7	26.4	72.2	26.7	77	17.1
Source Entities	-	-	-	25.8	19.4	51.18	25.3	58	22.9

Table 3.8: Mean P@1 for zero-shot cross-lingual factual knowledge transfer with language-agnostic entities for Non-Latin languages. Since the languages are in different scripts, we use either Latin entities or source language entities as language-agnostic entities.

source language for both source and target language as quasi-language-agnostic-entities. While transfer performance is still high, most language-pairs decrease in performance except for Japanese-Russian transfer, which slightly increased. Most notably, Japanese-Chinese transfer deteriorates in both directions by around 20% despite the lexical overlap between the languages. Our results also show that the high performance of language-agnostic entities is not restricted to only Latin-entities but can also be achieved with Non-Latin-entities, albeit not as well. We hypothesize that the deterioration in performance might be due to the fact that Latin-entities are more common in the pretraining corpus and therefore have a better representation across languages.

Multiple Source Languages

We extend our experiments to multiple source languages again. The results of mBERT being fine-tuned on two source languages are shown in Table 3.9 and on three source languages in Table 3.10. We see that the source-language-triple English-German-Spanish transferring to French has the highest performance with 84.5%. Surprisingly, the results with three source languages have only small improvements on the best performing transfer of two source languages (cf. transfer to French). The only exception is the transfer to German, which even decreases when English is added to French and Spanish. This is counter-intuitive since English and German are both Germanic languages compared to French and Spanish, which are Romance languages. In general, our results are consistent with our hypothesis that facts are easier to transfer when they are seen in multiple languages. During fine-tuning the source languages likely act as a parallel corpus that help align the representational spaces using the entities as anchor points, similar to the research of Libovický et al. (2020).

3.3.1 Aliases

Since mBERT needs to map the contextual representation of r_{source} onto r_{target}, we consider that our setup might suffer from a selection bias. While for a mul-

Source	en-de		en-fr		en-es		de-es		de-fr		fr-es	
Target	fr	es	de	es	fr	de	en	fr	en	es	de	en
Transfer	82.7	71	43.7	68.2	77.2	44.2	81.7	81.1	57.8	62.3	52.9	76

Table 3.9: Mean P@1 for zero-shot cross-lingual factual knowledge transfer using language-agnostic entities with two source languages.

Source	de-fr-es	en-fr-es	en-de-es	en-de-fr
Target	en	de	fr	es
Transfer	82.9	46.7	84.5	73.5

Table 3.10: Mean P@1 for zero-shot cross-lingual factual knowledge transfer using language-agnostic entities with three source languages.

tilingual setup the translations of entities in Wikidata are one-to-one mappings, relations are usually not. As example, the relation "instance of"[4] is also known under its aliases "is a", "is an example of" and many other. However, its German label is "ist ein(e)" which is a possible translation but its contextual meaning is not always equivalent. Fortunately Wikidata also provides aliases for relations in other languages. We use their respective aliases to construct alias-facts. Similar to Jiang et al. (2020b) using automatic prompt generation, using aliases provides a higher lower-bound on the factual knowledge shared between languages. Further, in cases where Wikidata does not provide an alias for a relation in a specific language, we compensate for this by translating all English alias-labels with GoogleTranslate[5].

We add alias-facts to our training set to reduce our selection bias and investigate if the added facts can improve the factual knowledge transfer. Some aliases might facilitate transfer more than others. However, our preliminary results showed that adding alias-facts to the training set even decreased performance for single source languages. Results for multiple source languages show that only the transfer to French benefited from training with alias-facts increasing mP@1 from 84.5% to 97.2%. All other Latin-European languages had a substantial decrease in performance instead. Next, we add alias-facts in the target language to the test set but not to the training set. When the object is inferred correctly for either the relation or one of its aliases, mBERT is considered to know the fact in the target language. This gives us a higher lower-bound on its transferability. The results are shown in Table 3.11. The performance increase is substantial, revealing that we have under-reported performance so far. The transfer with German as source or target language even doubles. Repeating the experiments for multiple source languages shows also a considerable improvement (Table 3.12). Transfer to Ger-

[4] www.wikidata.org/wiki/Property:P31
[5] translate.google.com

Source	en			de			fr			es		
Target	de	fr	es	en	fr	es	en	de	es	en	de	fr
Transfer	28.6	58.6	53.4	26.4	26.4	25.4	60	41.4	54.4	52.6	25.4	54.5
Transfer + test-aliases	68.2	72.5	73.4	54.3	58.6	50.2	79.6	69.6	79.4	75.2	56.4	67.2

Table 3.11: Mean P@1 for zero-shot factual knowledge transfer with language-agnostic entities with single source languages. Testing on relations and their aliases to reduce selection bias and increase knowledge lower-bound.

Source	de-fr-es	en-fr-es	en-de-es	en-de-fr
Target	en	de	fr	es
Transfer	82.9	46.7	84.5	73.5
Transfer + train-aliases	63	40	94.6	62
Transfer + test-aliases	**94.2**	**90.5**	**97.2**	**87.1**

Table 3.12: Mean P@1 for zero-shot factual knowledge transfer with multiple source languages. Trained with relations and their aliases, or instead tested on relations and on their aliases together.

man even increases from 46.7% to 90.5%. This could indicate that relation-labels for German are not optimally chosen by using the pre-selected ones from Wikidata. Our results demonstrate the importance of alternative prompt generation, in line with research of Jiang et al. (2020b). We also repeat our experiments in the multilingual setup with aliases in either training- or test-set but, unfortunately, we found no improvement in performance. Indicating that relations are likely not the transfer-bottleneck but instead multilingual entities.

3.3.2 Word Order

Although Japanese has a Subject-Object-Verb (SOV) word order, we have constructed our facts only in Subject-Verb-Object (SVO) word order so far. In this section, we investigate the compatibility of different word orders between source and target languages. For this we train our model with English facts in SVO word order and then evaluate on Japanese in SOV. Our experiments show that English-to-Japanese performance completely collapses to 0%. To validate this finding and to rule out possible typological differences impacting this result, we repeat the experiment with English as source language in SVO and German as target language in SOV word order. Again, transfer collapses to 0%, indicating that the model is incapable of transferring between languages in different word orders. Our results coincide with previous results on word order in zero-shot cross-lingual transfer (Pires et al., 2019; Dufter and Schütze, 2020; K et al., 2020).

Investigating this further, we conduct an experiment where source and tar-

Source	en			de			fr			es		
Target	de	fr	es	en	fr	es	en	de	es	en	de	fr
Transfer	28.6	58.6	53.4	26.4	26.4	25.4	60	41.4	54.4	52.6	25.4	54.5
Transfer + Dot	35.9	60.8	57	31.6	39.4	28.5	58.7	42.8	54.6	71.6	31.7	65

Table 3.13: Mean P@1 of zero-shot factual knowledge transfer on language-pairs with and without a dot added to the end of the facts.

get language are both in SOV word order. Again using English as source and Japanese and German as target languages. Not only is factual knowledge transfer possible but the performance from English to Japanese slightly improved from 24.8% in SVO to 26.5% in SOV. This likely comes from the pretraining where Japanese was trained with SOV. Further, the transfer from English to German, now both in SOV, increases from 28.6% to 51%. German not using either SVO or SOV word order but a combination of both, gives no clear indication on why the performance improvements are so drastic compared to Japanese.

3.3.3 Dot Test

So far we have constructed facts as entity-relation-entity triple statements. As some papers such as Petroni et al. (2019) choose to end their statements with a ".", we tested if it has an effect on our results. Interestingly, we found that adding a dot actually impacts performance quite considerably. We report the language-pair performance for Latin-European languages in Table 3.13. Our results show that a simple change such as adding a dot can have a positive impact on knowledge transfer. All language-pairs increase their mP@1 by at least multiple percentage points with the only exception of French-to-English and French-to-Spanish, where performance even slightly decreases or does not change. We also added a dot to the facts in multiple source languages and repeated the experiments. While transfer to German and Spanish increased by 5%, the transfer to English decrease by 4% and the transfer to French remained the same. We hypothesize that the dot can function as an additional anchor point that helps to transfer the contextual representation since the model was trained on complete sentences. Further, intuitively it seems reasonable that mBERT is more confident about its predictions when the sentence has ended. Single source languages could benefit more since the model is likely less confident about its prediction compared to multiple source languages.

3.3.4 Few-Shot Learning / Parallel Corpus

As discussed in the prior chapter, we found that adding a parallel corpus of facts to the training set increases results. We repeat these experiments again with single and multiple source languages for language-agnostic facts. A selection of

Source	en	en	en	ja
Target	de	fr	zh	ru
0-shot	28.6	58.6	46.8	17.1
50-shot	80	66.4	73.3	36.2
100-shot	88.7	87.3	77.7	42.3
200-shot	**95.4**	**98**	**87.8**	**82.2**
Highest Increase	+61.4	+39.4	+41	+65.1

Table 3.14: Mean P@1 of n-shot learning for language-agnostic facts. n-facts in the target language are added to the training set. Japanese-Russian experiments are done with Latin-entities.

results is illustrated in Table 3.14. By adding only 50 parallel facts to English, the English-to-German transfer increases from 28.6% to 80%. Similarly English-to-Chinese increases by 27% and Japanese-to-Russian more than doubles its transfer from 17.1% to 36.2%. As we can see not only languages in different scripts benefit from the parallel corpus. Transfer to German seems to have the largest deficiency that can be overcome by training with the parallel corpus. When using multiple source languages, mean P@1 increases to 100% by adding only 50 facts to the training language. This is in accordance with our previous results and shows the overwhelming impact parallel corpora can have on factual knowledge transfer.

3.4 Relations

In this section we investigate the individual performance of relations to better understand their impact on cross-lingual transfer. As previously mentioned, we find that some relations perform vastly better than others. Their performance is also not dependent on the language-pair and can change depending on the direction of the transfer. The same holds for aliases and even subwords of relations. During early analysis with language-agnostic entities, we found that mBERT sometimes transfers a fact only partially, i.e. instead of being able to predict the object for a given subject and relation in the target language, the model might only be able to predict the object in the context of a subword of the multi-token relation. For example, for the fact "Dresdner funding scheme Damm", Wikidata states the German label of the relation "funding scheme" as "Fördertopf". However, the model is not capable of predicting "Damm" for "Dresdner Fördertopf `[MASK]`". A reason might be that "Fördertopf" is not a commonly used word in German. Yet, the object "Damm" ranks high for German translations of the subword "scheme" such as "Plan", "Schema" and "planen". We see this as a partial knowledge transfer, where the model is capable of transferring the object in the context of a more common subword of a relation.

When analysing the individual performance of relations, we find that lexical overlap between the source and target relations is often indicative of their performance. Note, we measure the similarity between relations by tokenizing them into sets of tokens and computing their Jaccard index and overlap coefficient. Shared tokens can serve as a cross-lingual signal for the model to choose the right object for a given subject. Nevertheless, high transfer is not only achieved for relations with high similarity between source and target language. We identify many relations that do not share tokens but have high transfer performance. Further, we find instances of relations that have a high similarity but still a rather poor performance. For example in the multilingual setting, the model is only able to transfer 11.3% of facts for the source relation "programmer" (en) and target relation "programmeur" (fr). Part of the multilingual setting is that the entities are dissimilar and therefore the model has to not only transfer the context from a source to a target relation but also to align the context with the target entities. This leads us to believe that entities are more important for the transferability. This would explain why the performance increases when using language-agnostic entities. As a result the performance of the language-agnostic setting could be much more dependent on the choice of relation. To summarize, we find that high lexical overlap between relations often benefits the transferability of the facts but lexical overlap is not a necessary condition as we find examples of relations without any that perform very well.

Further analysis shows that relations do not have the same performance in all languages. This is even the case when the relations have a high similarity. For instance, in the language-agnostic setting the transfer of facts using the relation "symbolizes" (en) and "symbolisiert" (de) has a mean P@1 of 90% from English to German but only 77.9% from German to English. Note that we keep the entity-pairs consistent. This performance gap can also be reported for transfer to different target languages while keeping the source language constant, as we have already seen in prior sections. We hypothesize that the performance of relations depends on the representations and alignment within the model. This might depend on the frequency and the context of the relation in the pretraining corpus. By seeing the relation in more contexts, the model can build more robust representations that have a higher transfer performance. To test this we compute the correlation between the performance of the relations and the Wikidata counts[6], which represent the frequency of the relation in the Wikipedia corpus. We compute both the Pearson Correlation and Spearman's Rank Correlation Coefficient. For this we use language-agnostic facts and pair-datasets with single source languages. We find that there is no correlation with either metric between the performance of a relation and its frequency. This can also be seen in Figure 3.4.

[6] www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all

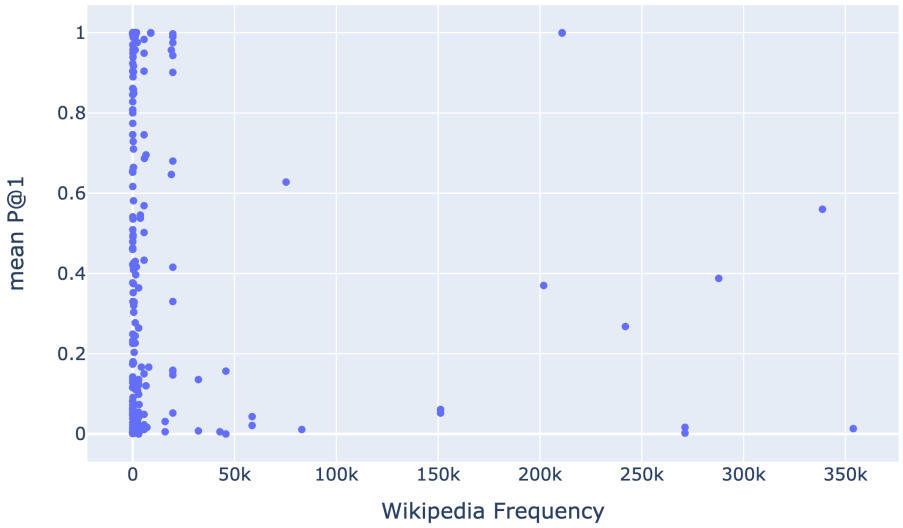

Figure 3.4: Mean P@1 of the transfer of relations compared to their Wikipedia Frequency (Wikidata count). Factual knowledge transfer was done with English as source language and performance is averaged across Latin target languages.

3.5 Entity-Frequency

In this section we investigate the effect of entity-frequency in the training set on the transfer performance. In particular, we first analyse the effect of re-using a subject-entity for a specific relation. Illustrated below is an example where subject-entities are re-used three times for each relation:

$$\left.\begin{matrix}(e1, r1, f1)\\(e1, r1, f2)\\(e1, r1, f3)\end{matrix}\right\} 3$$
$$\left.\begin{matrix}(e2, r1, f4)\\(e2, r1, f5)\\(e2, r1, f6)\end{matrix}\right\} 3$$
$$\vdots$$

For this, we control the number of instances of a subject. In our default experimental setup, every subject-entity is unique for a relation. Note that by increasing the number of instances of a particular subject, we also need to increase k of Precision@k. We hypothesize that the model could benefit from seeing a similar context (subject-entity and relation) many times, creating a better representation of the context itself and therefore increasing transfer performance.

Source	en	
Target	de-fr-es	
Entities	multilingual	agnostic
Re-Use 1 - 500 Facts	**18.9**	95.3
Re-Use 1 - 5 Relations	18.5	**97.6**
Re-Use 2	14.1	76.9
Re-Use 5	11.2	57.3
Re-Use 10	10.6	59.1

Table 3.15: Mean P@1 of re-using subject-entities n times across relations for multilingual and language-agnostic factual knowledge transfer. Note that multilingual refers to using a multilingual object with an agnostic subject. English as source language and performance is averaged across Latin-languages as target.

The experiments are conducted with English as source language and the remaining Latin-languages as target languages, evaluating their average transfer performance. Our results with language-agnostic entities show that re-using the subject decreases the precision substantially. Re-using the subject 10 times decreases mP@1 from 59.13% to 39.8%. Interestingly, increasing the re-use factor to 100 decreases the performance to only 44.29%. We repeat the experiments with multilingual entities and find that for these the transfer increases. From 2.5% for unique subject-entities up to 15.9% mean P@1 for using the same entity for all facts in a relation. We hypothesize that re-used subject-entities make it easier since they remove the additional complexity of transferring multiple subjects.

Next we analyse the influence of re-using subject-entities across relations. Before, we used the same subject-entities across relations to control for co-occurrences. Below we illustrate an example of re-using subject-entities three times across relations:

$$\left.\begin{array}{l}(e1, r1, f1)\\(e1, r2, f2)\\(e1, r3, f3)\end{array}\right\} 3$$
$$\left.\begin{array}{l}(e2, r1, f4)\\(e2, r2, f5)\\(e2, r3, f6)\end{array}\right\} 3$$
$$\vdots$$

We hypothesize that seeing entities with different relations might aid alignment during fine-tuning and therefore encourage cross-lingual transfer. We select English as source language and the remaining Latin-languages as target languages, taking the average over the performance, and use language-agnostic entities. The results of our experiments are shown in Table 3.15. Surprisingly, repeating the subject in only 5 instead of 10 relations decreases the performance slightly. However, further reducing subject-entities to 2 relations increases the performance

by 17.8% above our default setup. Since we only have around 6000 entities, we were not able to repeat the experiment with only unique subject-entities in the default setup. Instead we first reduced the total number of relations from 10 to 5. In a second experiment we reduced the number of facts from 1000 to 500. In both, the model achieves almost perfect transfer. However, this is likely due to the model simply learning co-occurrences between subject and object. Therefore we conduct additional experiments with facts using a language-agnostic subject but multilingual object to remove the chance of the model learning from co-occurrences. In our default setup English achieves an average mP@1 of 10.6%. Reducing the frequency of the subject-entities increases the performance steadily. With unique subject entities the model can achieve 18.5% average mP@1. This seems to confirm that, at least for multilingual facts, the model can benefit from a variety of subject-entities.

3.6 Multilingual Training for Monolingual Models

Motivated by the results of de Souza et al. (2021) and Artetxe et al. (2020), we investigate zero-shot cross-lingual factual knowledge transfer on pretrained monolingual English-BERT[7] with language-agnostic entities. We follow our default setup of using 10 relations and 1000 entity-pairs per relation. We take the average results over three seeds (42, 69, 10) to reduce selection bias. Again, if monolingual BERT learns only entity co-occurrences, its test accuracy would be at best 10% for a given relation. The results for each source language are presented in Table 3.16. Our results indicate that English BERT has the ability to learn factual information from fine-tuning on facts in other languages. The performance of Latin-languages is clearly above chance with French at over 30% and Spanish and German with 26% and 18.8% respectively. Non-Latin-languages like Russian, Chinese and Japanese have only a transfer performance of around 5%.

To better understand these results, we investigate the individual performance of relations. We find that some relations contain Latin-subwords. For instance, in Japanese "以下のMOD" with corresponding target "mod of" in English. Some relations in source and target language are similar enough for learning language shortcuts, i.e. learning shared subword co-occurences as cross-lingual signal. However, further investigation shows that these relations are rare in Non-Latin languages. Even though average performance is poor, we find relations that have a high individual transfer performance without lexical overlap. For example in Russian we identify "запрещает" (ru) and "forbids" (en) with a mP@1 of 29.8%, clearly above chance. The same holds true for other languages. In Chinese we find "批准方" (zh) and "approval" (en) with 37.7%. With Latin-languages a lexical overlap between relations in source and target language is much more likely. We

[7] bert-base-cased in HuggingFace Library

Source	en	de	fr	es	ru	zh	ja
Untrained	39.2	0	0	0	0	3.5	3.4
Transfer	100	18.8	30.8	26.0	5.7	4.7	4.2
no overlap	-	13.1	14.8	17.3	8.7	6.9	3.8
no overlap + test-aliases	-	28.9	29.4	31.2	21.5	15.1	8.6

Table 3.16: Mean P@1 of factual knowledge transfer with language-agnostic entities on English-BERT. Transfer is done on trained and untrained BERT, relations explicitly without overlap between source and target language and with additional testing on aliases to reduce selection bias. Performance is averaged over three seeds.

measure the subword overlap again by token-set-similarity with the Jaccard index and the overlap coefficient. In French we find "symbolise" and "symbolizes" (fr-source and en-target respectively). These have a large subword overlap and also a high transfer of 95.7%. In contrast, "parentèle" and "grandparent" (fr-source and en-target) have no subword overlap but a mP@1 of 50.5% and similarly "défluent" and "distributary" have a performance of 71.6%. In Latin-languages average performance is high but might be artificially increased by relations with very high overlap. To verify this claim, we repeat the experiments with only relations that have no token-overlap between the source and target language (*no overlap*). While average performance has substantially decreased, e.g. French decreased from 30.8% to only 14.8%, the performances without token-overlap are still above chance. Next, we added testing with aliases to reduce possible selection bias (*no overlap+testing aliases*). This increased the performance again to 29.4% for French and to 28.9% and 31.2% for German and Spanish respectively. Non-Latin languages also increase quite drastically in performance indicating that especially Russian and Chinese were underreported. Additionally, we train English-BERT from scratch on our data to verify that it does not simply exploit hidden structures in training and test set (*untrained*). For English, the model learns to memorize 39.2% of the facts. Interestingly, only for Chinese and Japanese does the model learn 3.5%, likely through co-occurrences. The results indicate that BERT does not just exploit structures in the data and strengthens our hypothesis that BERT is capable of zero-shot cross-lingual factual knowledge transfer by utilizing language-agnostic representations.

Prior works hypothesize that monolingual models are capable of zero-shot cross-lingual transfer because they develop a language-neutral representation space, possibly as the by-product of the MLM training objective. In addition, the pretraining corpus of monolingual BERT such as BookCorpus and Wikipedia could contain text in other languages. For instance, English Wikipedia contains several articles with phrases and words from other language and the same possibly holds for books in the BookCorpus. Additionally, Wiktionary as a subset

of the Wikipedia corpus contains many translations of words. These could aid monolingual BERT in constructing a language-neutral representation. Our results seem to support this. However, the performance with no token-overlap could also be explained by BERT representing similar subwords close together as these were probably seen in similar contexts during pretraining. Therefore achieving a cross-lingual signal and evading our similarity measures, i.e. having no token overlap.

3.7 Discussion

We studied zero-shot cross-lingual transfer for factual knowledge by fine-tuning multilingual BERT on synthetic facts in a source language and evaluating the zero-shot transfer performance in a target language. While our experiments reveal that mBERT is capable of zero-shot cross-lingual transfer of factual knowledge, the model exhibits only minor abilities to transfer multilingual facts from a source to a target language without any additional refinements. We suspect this to be possibly due to deficiencies in the alignment between the languages. This might also restrict the utilization of factual knowledge transfer during pretraining. Based on our results, these deficiencies can be partially overcome by using multiple source languages. We speculate that this is due to them functioning as parallel dataset facilitating the alignment of their embedding spaces. This holds similarly for few-shot learning, which adds a small amount of facts from the test set to the training set and creates a parallel corpus of source and target languages. We demonstrated that even by adding a small amount of parallel data, the performance increases substantially. Possibly due to aligning language-specific subspaces of the source languages with the subspace of the target language. Further, we find that using language-agnostic entities improves the transfer performance substantially. These might act as anchor points during fine-tuning.

Experiments on Russian, Chinese and Japanese show that cross-lingual factual knowledge transfer is possible even without a common script albeit using language-agnostic entities. Note that especially Chinese-Japanese factual knowledge transfer performs very well. We believe this is due to a partial lexical overlap. So, while not necessary, the model can benefit from a partial overlap. In contrast, the distance between languages seems to be negligible as English and Russian are both Indo-European languages but the transfer from English to Chinese or Japanese is much higher. For Latin-languages we find that German performs surprisingly bad as either source or target language. We hypothesize that this is due to relation-labels for German not optimally chosen by using the pre-selected relation labels from Wikidata. The results of our experiments with a combined evaluation with aliases support this and indicate that German is simply underreported.

Our results showed that a matching word order between source and target language is necessary for cross-lingual transfer. Moreover, we found that the true word order of source and target language is not relevant. When English and German are changed to SOV, transfer performance greatly increases. When word order diverges between source and target language, performance of cross-lingual transfer collapses. This is in agreement with the analyses of prior work (Pires et al., 2019; Dufter and Schütze, 2020; K et al., 2020). We used aliases of relations to create additional facts to reduce the selection bias of the pre-selected relation-labels. Instead of testing the transfer of a fact by using the relation's target label provided by Wikidata, we use aliases of the relations as well and examine if these aid the model in its prediction. The results reveal that we have underreported the transfer so far and that testing with aliases increases the lower-bound on the transfer performance for language-agnostic facts. We also investigated the impact of adding a dot to the end of our facts and found that this has a surprisingly large effect on language-agnostic facts with a single source language. A possible reason could be that by ending the sentence, the model has a higher confidence in its prediction. Alternatively the dot could function as an additional anchor point. For multiple source languages, the transfer is already quite high, so the model does not benefit from it as much. Last, we find that monolingual BERT has the ability to learn factual knowledge by being fine-tuned on a different language than it was trained on, albeit using language-agnostic entities. Put differently, pretrained monolingual BERT seems to develop a language-neutral representation space that can use of factual knowledge even from other languages, providing further evidence for monolingual BERT's zero-shot cross-lingual transfer capabilities.

Following Pires et al. (2019) and Dufter and Schütze (2020), we hypothesize that zero-shot cross-lingual transfer of factual knowledge works because language-specific representation spaces are aligned during pre-training as a result of common words being mapped into a shared representation, likely because of capacity constraints on the model. This forces it to use its parameters as efficient as possible which leads to abstractions among languages. As a result, co-occurring words are mapped close to the common words, which leads to alignment. This also explains why parallel corpora lead to substantial improvements in transfer performance and why language-agnostic entities perform so well compared to multilingual entities in our experiments. Language-agnostic entities are likely already mapped in a language-neutral space and contextual representation can be transferred more easily.

3.7.1 Limitations

Our experimental setup has some limitations. First, we only use single-token entities which leads to a selection bias. For some languages that use ideograms like Japanese and Chinese, this is very restrictive. Further, since the Word-

Piece tokenization-algorithm creates subwords based on their frequency in the training corpus, single-token entities have by definition a high frequency and therefore likely exhibit very different representations compared to multi-token entities. Additionally, we not only select single-token entities but pairs of single-token entities, e.g. for the language-pair English-Russian an entity needs to be a single-token in both of them. This is easier for languages in the same family or that have a significant lexical overlap. Second, the synthetic corpus is not very representative of natural language. mBERT could have a vastly different performance on natural language queries and transfer much better on natural text since it is closer to the pretraining data, although some of our results indicate that performance is comparable. Further, facts could accidentally introduce morphological errors such as mismatch in grammatical gender. This could make the performance worse than it actually is. While this might be the case, we believe that in practice the performance difference is negligible or at least comparable as we have discussed earlier. Last, although fine-tuning mBERT, especially with our setup, is making use of existing multilingual representations and more akin to a continuation of pretraining, training mBERT from scratch and controlling the pretraining process can lead to a more intricate analysis.

Chapter 4

Symbolic Reasoning Transfer

In the previous chapter, we found that factual knowledge is shared across languages through language-agnostic representations. However, Kassner et al. (2020) have found that new factual knowledge can also emerge through the application of symbolic reasoning. In this chapter, we extend their investigation to a zero-shot cross-lingual setting and examine the ability of MLLMs to use symbolic reasoning across languages to infer previously unseen factual knowledge. In particular, we fine-tune mBERT on symbolic rules in a source language and evaluate if the model can apply them in a target language. Further, we investigate if the model has the ability to combine symbolic reasoning with factual knowledge transfer. Following Kassner et al., we study equivalence, symmetry, inversion, negation, implication and composition (Table 4.1). Note that in contrast to other reasoning tasks, we do not provide all information at inference time. Instead, the model has to reason with symbolic rules on memorized factual knowledge to infer previously unseen facts. For instance, if the model has learned that the relation "is married to" is symmetric, then for a fact like "Alice is married to Bob" the statement "Bob is married to Alice" should be inferred. A summary of selected results for zero-shot cross-lingual rule-transfer can be found in Table 4.2.

The chapter is structured as follows: First, we discuss the experimental setup and the data acquisition process for rule-specific relations. Next, we analyse the cross-lingual transfer for each rule for general and rule-specific relations. Last, we discuss the results of the experiments in a broader context.

Rule		**Definition**	**Example**
EQUI	Equivalence	$(e, r, f) \iff (e, s, f)$	(bird, can, fly) $\iff$ (bird, is able to, fly)
SYM	Symmetry	$(e, r, f) \iff (f, r, e)$	(bob, married, alice) $\iff$ (alice, married, bob)
INV	Inversion	$(e, r, f) \iff (f, s, e)$	(john, loves, soccer) $\iff$ (soccer, thrills, john)
NEG	Negation	$(e, r, a) \iff (e, not\ r, b)$	(jupiter, is, big) $\iff$ (jupiter, is not, small)
IMP	Implication	$(e, r, f) \Rightarrow (e, s, a), (e, s, b), \ldots$	(dog, is, mammal) $\Rightarrow$ (dog, has, hair)
COMP	Composition	$(e, r, f) \wedge (f, s, g) \Rightarrow (e, t, g)$	(tiger, faster than, sheep) $\wedge$ (sheep, faster than, snail) $\Rightarrow$ (leopard, faster than, snail) with $r = s = t$

Table 4.1: Overview of the symbolic rules with entities $e, f, g, a, b \in E$ and relations $r, s, t \in R$ with example in natural language (Kassner et al., 2020).

Source	en			de			fr			es		
Target	de	fr	es	en	fr	es	en	de	es	en	de	fr
EQUI	8.1	20.8	21.5	14	11.7	8.7	17.4	8.3	18	23.3	7.8	27.4
SYM	0	2.6	7.4	0	1	0	0.5	0.1	2.2	7.4	1.2	2.6
INV	1.1	13.2	7.4	6.3	5.6	2.5	13.6	6.1	12.8	2.9	2.1	18.7
NEG	19.7	18.7	21.4	**15.5**	13.7	**15.3**	15.4	**17**	18.9	18.8	**19.4**	24.1
IMP	-	-	-	-	-	-	-	-	-	-	-	-
COMP	**26**	**31.3**	**37.6**	14.2	**14.3**	14.4	**27.3**	13.4	**29.9**	**30.3**	2	**32.6**

Table 4.2: Mean Precision@1 of zero-shot transfer of symbolic rules from source language to target language for single source Latin-languages with general relations. Note: IMP was not above chance, so results are omitted.

4.1 Experimental Setup

We hypothesize that MLLMs such as mBERT have the ability to use symbolic reasoning (SR) to infer factual knowledge across languages. The model can either (i) share symbolic rules associated with a relation from a source to a target language and then infer new facts through their application (*rule-transfer*), or (ii) combine factual knowledge transfer (FKT) with the application of the symbolic rule (*SR-FKT*). For this mBERT is trained on a training set consisting of two different set of facts: train-facts and test-facts. Train-facts are used to demonstrate the symbolic rule to the model in a source language. Test-facts are facts memorized by the model, so that the correct application of the rule can be tested at inference time. For the equivalence-rule a test-fact could be (e, r, f) and at inference time (e, s, f) would need to be inferred. Therefore the test-facts act as the premise of the symbolic rule. During evaluation the model should then be able to infer the conclusions, i.e. the target-facts corresponding to the application of the rule on the test-facts. These target-facts form the test set. We test this by using cloze-like statements. If the model correctly predicts the object of the target-fact, it is considered to have inferred it from its corresponding test-fact. While the target-fact is always in the target language, the test-facts can either be in the source or target language. By that we can analyse different abilities of the model. Test-facts in the target language are to analyse if the model has the ability to transfer the rule for a relation across languages (Fig. 4.1). Test-facts in the source language are to analyse the model's capability of combining factual knowledge transfer with the learned rule (Fig. 4.2).

We follow a setup similar to our previous experiments with factual knowledge transfer. Again facts are created by sampling relations and entities from our synthetic dataset and mean Precision@1 is used as metric. Since our experiments with multilingual entities performed rather poorly, we will focus only on language-agnostic entities. In addition to the relations we previously used, we collect relations that likely exhibit symbolic rules as part of the pretraining process. For

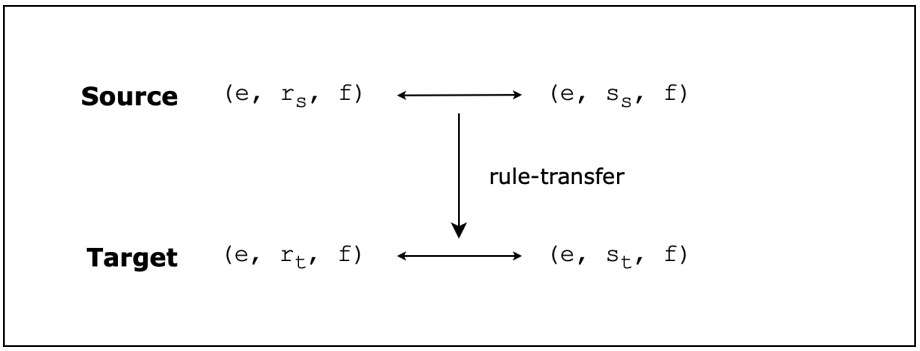

Figure 4.1: Overview of *rule-transfer* by the example of equivalence. A symbolic rule that is demonstrated in the source language is transferred in a zero-shot manner to a target language.

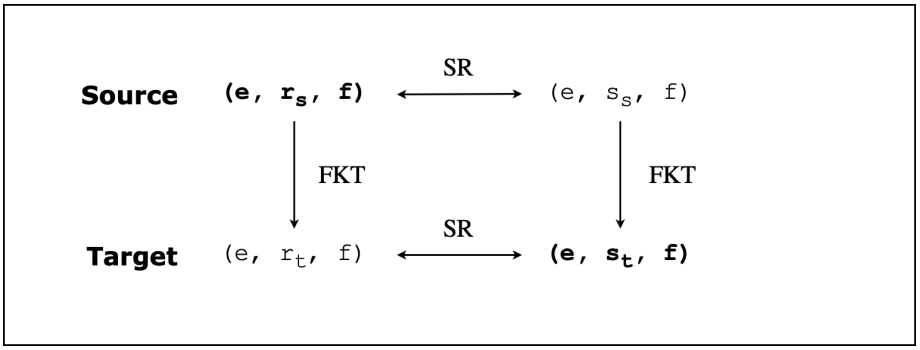

Figure 4.2: Possible ways of combining factual knowledge transfer (FKT) with symbolic reasoning (SR) by the model to transfer test-fact (e, r_s, f) from source language s to target-fact (e, s_t, f) in target language t using the equivalence rule.

instance "is married to" displays symmetry in the example above. We refer to these relations as **rule-specific relations** compared to **general relations** that do not display the rule during pretraining. For instance, "is capital of" is not a symmetric relation, therefore counted as general in a symmetric context. By training on either of them, we can compare if mBERT is able to learn rules easier with relations that possibly exhibited the behaviour already during pretraining. The training set is divided into 90% of facts which are dedicated to learning the symbolic rule in the source language (train-facts) and 10% to memorize facts for evaluation in either source or target language (test-facts). Similar to our previous experiments, we also investigate the impact of training with multiple source languages.

Next, we discuss the acquisition process for rule-specific relations. For equivalence (EQUI), we need relations that are, as the name suggests, equivalent to each other. That means they need to be synonymous or simply: aliases. As pre-

viously mentioned, Wikidata provides aliases for most relations. The equivalent relations are obtained by collecting these aliases for all relations in our language-pairs. The rule-specific relations for symmetry (SYM) are more challenging to acquire. Since Wikidata represents a directed graph, we can load the dataset into the graph database system Neo4j[1]. Now every Wikidata entity-relation-entity triple is seen as a connection between vertices (entities) over edges (relations). We then query the database for all symmetric connections in the graph and filter these to collect the symmetric relations. For inversion, we use the SPARQL query service provided by Wikidata[2] that can be used to query the data directly through their endpoint. We use the service to query all relations that have the inversion property[3] to collect the rule-specific relations for Inversion (INV). Negation (NEG) relations are created by adding a simple negation (*'not'*) before the relation. To make relations more grammatically correct, we manually negate pre-selected relations and translate them to other languages with GoogleTranslate. These are our rule-specific relations. Additionally, we collected antonym entities for negation. Antonym pairs such as 'sky' and 'earth' have the property 'opposite of'[4] in Wikidata. Querying them through the SPARQL query service results in 23,077 entity pairs but manual filtering brought the number down to only 30. Finally, implication (IMP) and composition (COMP) both do not have rule-specific relations, so only general relations are used. Note that we only collect rule-specific relations for Latin-languages as not enough were available for Non-Latin languages.

4.2 Equivalence

In our experiments with equivalence (EQUI), we investigate if associations of relations and their aliases are shared across languages:

$$(e, r, f) \iff (e, s, f)$$

More specifically we ask: Does mBERT share the equivalence of a relation and its alias from a source language to a target language?

The model is trained on the equivalence of general relations. Here, general relations are unrelated relation-pairs that are trained as if they were aliases. For those, the model can not leverage existing pretraining knowledge. We find that the model can learn to transfer the equivalence of these pairs from source to the target language, although not with high performance. The results with Latin-European languages are shown in Table 4.3. The model can infer up to 27%

[1] https://neo4j.com
[2] query.wikidata.org
[3] www.wikidata.org/wiki/Property:P1696
[4] www.wikidata.org/wiki/Property:P461

Source	en			de			fr			es		
Target	de	fr	es	en	fr	es	en	de	es	en	de	fr
rule-transfer												
general	8.1	20.8	21.5	14	11.7	8.7	17.4	8.3	18	23.3	7.8	27.4
rule-specific	**30.9**	**51.9**	**60**	**53.2**	**61**	**61.9**	48.8	**46.2**	**60**	**62.1**	**49.6**	**60**
SR+FKT												
general	29.3	46.6	47.1	29.6	39.4	31.1	**48.9**	31.9	49.4	23.3	27.8	52.7
rule-specific	15.8	36.1	30.3	17.7	22.6	21.9	29	25.9	38.2	38.9	29.5	53.2

Table 4.3: Mean P@1 of zero-shot cross-lingual transfer of equivalence and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) for general and rule-specific relations on Latin-languages. In bold is the overall best transfer performance to the target language.

Source	en			ru		zh		ja	
Target	ru	zh	ja	zh	ja	ja	ru	zh	ru
rule-transfer	6.1	3.4	5.4	7.2	7.5	24.3	2.7	15.6	3.2
SR+FKT	22.2	16.5	12.5	26.2	17.1	67.9	32	61.9	11.4

Table 4.4: Mean P@1 of zero-shot cross-lingual transfer of equivalence and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) for general relations on Non-Latin-languages with Latin entities.

of target-facts (Spanish-to-French). German as source language has the poorest equivalence-transfer with an average mP@1 of 11.5%. In contrast, Spanish exhibits the best transfer among the Latin-European-languages with an average mP@1 of 19.5%. For Non-Latin languages (Table 4.4) the best transfer is Chinese-to-Japanese and the converse Japanese-to-Chinese with 24.3% and 15.6% respectively. This might be partially because of their lexical overlap, as they have a comparable performance to Latin-European-languages that also have a high lexical overlap. Other language-pairs have a performance of only 3 to 8% with English-centric language-pairs performing the worst. If we compare the results with our previous experiments, we find a surprising correlation between the performance of language-pairs for factual knowledge transfer and equivalence-transfer.

To investigate the utilization of pretraining knowledge for equivalence transfer, we repeat the experiments with rule-specific relations. For equivalence, these relations are pairs of respective aliases. Our experiments demonstrate that the model can leverage the pretraining knowledge and infer up to 62% of target-facts. However, we already found in Section 3.3 that factual knowledge transfer is not restricted to the relation's target label but aliases often have comparable performance. Possibly due to having appeared in a similar context during pretraining. We suspect that the same holds for aliases within a language. This

means that the high performance is not sufficient evidence for the language transfer of equivalence but instead the model could already have the ability to show similar knowledge for aliases. We train the model with rule-specific relations to further investigate this. In particular, we only train on the test-facts in the target language but without train-facts, i.e. the model does not learn equivalence explicitly. When trained in German the model achieves a mean P@1 of 13.6% solely based on its pretraining. While this is high, it shows that not all of the performance can be explained by the pretraining. Further note that the difference of the model's transfer from German with rule-specific relations (30.9% (en), 46.2% (fr), 49.6% (es)) and only relying on pretraining (13.6%) is higher than the transfer performance with general relations (8%). This indicates that the model can exploit the pretraining for a more efficient rule-transfer for rule-specific relations. An even more extreme example is Spanish-to-English. We find that the model achieves a performance of 11.5% in English by relying only on its pretraining. In comparison, this is significantly less than the transfer with rule-specific relations of 62.1%. We conclude that while relations and their aliases share similar contexts within a language through the pretraining, this can not fully explain the high performance of rule-specific relations. Instead, we find that cross-lingual rule-transfer benefits from the pretraining as well. Similar to our previous experiments on language-agnostic factual knowledge transfer, combining multiple source languages increases the rule-transfer performance (Table 4.5). While the model achieves the highest performance with 81.2% mP@1 on the transfer to French, adding more source languages to the transfer to German does not compensate for its already weaker performance. Comparing rule-specific and general relations, rule-specific relations perform worse in the transfer to English and French while German and Spanish seem benefit more from their pretraining.

Further, we find that the model can utilize both factual knowledge transfer and the equivalence rule to transfer factual knowledge to other languages. For this we train the model on test-facts in the source language instead of the target language. There are two ways to combine factual knowledge transfer and the symbolic rule to transfer factual knowledge to other languages (SR-FKT) as illustrated in Fig. 4.2. Either the test-fact in the source language is (i) transferred from source to target language via factual knowledge transfer and symbolic reasoning is applied on the result ($FKT \rightarrow SR$) or, alternatively, (ii) symbolic reasoning is applied to the test-fact in the source language and then the result is transferred to the target language ($SR \rightarrow FKT$) to infer the target-fact. We refer to these as the *paths* the model can take. Surprisingly, we find that the model is capable of combining symbolic reasoning and factual knowledge transfer to create new, unseen facts in other languages. Furthermore, the results are higher than with *rule-transfer*. We also find that general relations perform better than rule-specific relations in this setup. For Non-Latin languages the highest performance is again achieved with the Chinese-Japanese language-pair. When using multiple source languages, all performances are above single source language per-

Source Target	de-fr-es en	en-fr-es de	en-de-es fr	en-de-fr es
rule-transfer				
baseline	23.3	8.3	27.4	21.5
general	72.3	25.6	**81.2**	54.5
rule-specific	66.2	**56.3**	64	**77.1**
SR+FKT				
baseline	48.9	31.9	52.7	49.4
general	**78**	54.9	71.6	65.2
rule-specific	49.2	35.1	62.1	65.5

Table 4.5: Mean P@1 of zero-shot cross-lingual transfer of equivalence and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) with multiple source languages. The baseline is the best transfer from a single source to the target language with general relations. In bold is the overall best transfer performance to the target language.

formance with English-French-Spanish performing the worst with 54.9%, while German-French-Spanish transfer to English performs the best with 78%.

We analyse which path the model chooses by examining the performance of applying the equivalence rule $(e, r_\mathrm{s}, f) \rightarrow (e, s_\mathrm{s}, f)$ to the test-fact and the performance of factual knowledge transfer of the test-fact $(e, r_\mathrm{s}, f) \rightarrow (e, r_\mathrm{t}, f)$. Using English-to-German, we find that the average symbolic reasoning performance in the source language is very high with 94%, comparable to the results of Kassner et al. (2020). The factual knowledge transfer is much lower with a mean P@1 of 33.7%. Analysing the overlap between them, we find that almost all test-facts that have been transferred are also successfully reasoned upon with the equivalence rule. If we assume that all correctly inferred target-facts could have come from symbolic reasoning on the transferred test-facts (e, r_t, f), we can compute the overlap between them to find the performance upper-bound on $(FKT \rightarrow SR)$. We find this to be 16.7%. With a total mean P@1 of 32.3% on the target-facts, this shows that 15.6% are unaccounted for and therefore the lower-bound on the $(SR \rightarrow FKT)$-path. Accordingly, we hypothesize that it is much easier for the model to apply the symbolic rule in the source language, the language it was explicitly learned, and then use factual knowledge transfer to transfer the results to the target language $(SR \rightarrow FKT)$.

Source	en			de			fr			es		
Target	de	fr	es	en	fr	es	en	de	es	en	de	fr
rule-transfer	0	2.6	7.4	0	1	0	0.5	0.1	2.2	7.4	1.2	2.6
SR+FKT	36.3	47.4	48.2	36.5	42.6	40.1	41.1	43.8	44.2	37.8	39.5	46.3

Table 4.6: Mean P@1 of zero-shot cross-lingual transfer of symmetry and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) with Latin-European-languages.

Source	en			ru		zh		ja	
Target	ru	zh	ja	zh	ja	ja	ru	zh	ru
rule-transfer	0	0	1	0	0	1.4	0	3.9	0
SR+FKT	40.1	50.1	44	42.1	39.7	59.2	28.1	54.5	39.6

Table 4.7: Mean P@1 of zero-shot cross-lingual transfer of symmetry and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) with Non-Latin-European-languages with Latin entities.

4.3 Symmetry

In this section, we investigate if the symmetry-rule can be shared across languages. Here, symmetry can be viewed as a property of relations.

$$(e, r, f) \iff (f, r, e)$$

Again the model is trained on general and rule-specific relations with test-facts in either source or target language. Since we find that rule-specific and general relations have very similar performance for single source languages, we only report the results for general relations. This indicates that the model barely learns symmetry during pretraining for our rule-specific relations. The results for Latin-European languages can be found in Table 4.6. Surprisingly, we find that the performance is barely above 2% for rule-transfer in most language-pairs. Non-Latin-languages perform similarly poor as shown in Table 4.7. The highest transfer has the language-pair English-Spanish with 7.4% in either direction and Japanese-to-Chinese with 3.9% performing better than most Latin-European languages. In contrast, when the symmetry rule is combined with factual knowledge transfer, the model infers 40 to 48% of target-facts. Similar to equivalence, Chinese-to-Japanese performs the best with 59.2% of all Non-Latin-language-pairs. However, the transfer English-to-Chinese is much better for symmetry than with equivalence. All other languages perform around the same with Russian being seemingly worse at symmetry in either transfer direction. When multiple source languages are used, the symmetry-rule can be transferred much easier. The rule-transfer performance increases for every target language with Spanish even achieving 36.7% mP@1 (Table 4.8) compared to the best single source per-

Source Target	de-fr-es en	en-fr-es de	en-de-es fr	en-de-fr es
rule-transfer				
baseline	7.4	1.2	2.6	7.4
general	19.5	20.4	21.2	36.7
rule-specific	14	28.6	13.9	23.9
SR+FKT				
baseline	41.1	43.8	47.4	48.2
general	49.6	**60.2**	**60.8**	53.6
rule-specific	**58.2**	52.5	57.6	**59.8**

Table 4.8: Mean P@1 of zero-shot cross-lingual transfer of symmetry and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) with multiple source languages. The baseline is the best transfer from a single source to the target language with general relations. In bold is the overall best transfer performance to the target language.

formance of 7.4%. Similarly, rule-transfer to German increased from only 1.4% to 20.4%. Further, we find that rule-specific relations perform worse than general relations, most likely due to not displaying symmetry during pretraining and having unfavorable alignment except for the transfer to German where mean P@1 increases by 8.2% over general relations.

Since there is a substantial gap in performance between source test-facts and target test-facts, the model can apply symmetry much better in the source language than in the target language. This means that when mBERT is provided with source test-facts, it infers target-facts by ($SR \rightarrow FKT$). We analyse this behaviour more in-depth for English-to-German and find that applying the symmetry-rule has a performance of 75% in the source language. We also analyse the overlap of facts between factual knowledge transfer from source to target language $(e, r_{\text{en}}, f) \rightarrow (e, r_{\text{de}}, f)$ and symbolic reasoning in the source language $(e, r_{\text{en}}, f) \rightarrow (f, r_{\text{en}}, e)$. Similar to equivalence, we find that for almost all test-facts that are successfully transferred, the model has also applied symbolic reasoning. However, symbolic reasoning creates almost three times more facts. Repeating our experiments with German-to-English confirms these findings.

Kassner et al. (2020) find that in their experiments BERT tends to overgeneralize symmetry. While we find that mBERT partially overgeneralizes as well, with up to 20% of random relations displaying symmetry, the model can mostly distinguish between symmetric and non-symmetric relations. This is likely due to BERT being trained from scratch on a vastly higher percentage of symmetric relations and not on a natural language corpus, while we only fine-tune mBERT.

Source	en			de			fr			es		
Target	de	fr	es	en	fr	es	en	de	es	en	de	fr
rule-transfer												
general	1.1	13.2	7.4	6.3	5.6	2.5	13.6	6.1	12.8	2.9	2.1	18.7
rule-specific	11.4	6.7	18.8	14	6.8	13.8	10.1	8.8	12.6	10.5	14.6	6.3
SR+FKT												
general	**28.3**	**44.4**	**39.9**	**29.9**	**33.1**	**28.9**	**40.6**	**29.5**	**41.4**	**45.4**	25.1	**46.5**
rule-specific	19.6	21.7	23.7	22.5	16.1	24.5	27.8	17	38.2	29.8	**33.1**	44.1

Table 4.9: Mean P@1 of zero-shot cross-lingual transfer of inversion and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) with Latin-European-languages.

4.4 Inversion

Inversion (INV) can be viewed as a combination of symmetry and equivalence, therefore adding complexity to both rules:

$$(e, r, f) \iff (f, s, e)$$

Our results for inversion-transfer in Table 4.9 illustrate that mBERT is able to share the rule from source to target language better than symmetry. However, while some language-pairs perform well like Spanish-to-French with 18.7%, other language-pairs like English-to-German have barely any rule-transfer. In contrast, source test-facts achieve better performance, indicating that mBERT utilizes symbolic reasoning better in the source language than in the target language. The same holds for Non-Latin languages (Table 4.10), where again rule-transfer is on average higher than for symmetry. Similar to Latin-languages, Non-Latin languages can achieve a higher performance by combining factual knowledge transfer with symbolic reasoning. For this, Russian, Chinese and Japanese mostly perform equally well ranging from 20% to 25%. The exceptions are English-to-Chinese with almost 30% and the Chinese-Japanese language-pair exhibiting vastly better transfer performance again with 53% and 48.3% respectively. Our results with multiple source languages for inversion are shown in Table 4.11. Again in line with prior rules, utilizing multiple source languages makes rule-transfer easier, possibly due to the added information to generalize from. Surprisingly, when we investigated the effect of rule-specific relations and therefore the impact of pretraining on rule-transfer, we found that these perform in almost the opposite way to general relations with large gaps between their performances. This indicates that our collected rule-specific relations likely did not exhibit inversion during pretraining and the performance gap is solely due to alignment differences, i.e. due to some relations simply performing better or worse than others.

To better understand the divergence of performance between *rule-transfer* and *SR-FKT*, we analyse the model's predictive behaviour again. We find that mBERT is barely able to transfer source test-facts directly to the target language

Source	en			ru		zh		ja	
Target	ru	zh	ja	zh	ja	ja	ru	zh	ru
rule-transfer	1.4	2.3	2.1	2.3	4.5	4.1	1.4	4.8	3.4
SR+FKT	21.8	29.4	21.2	26.1	25.3	53	19.5	48.3	18.2

Table 4.10: Mean P@1 of zero-shot cross-lingual transfer of inversion and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) for general relations on Non-Latin-languages with Latin entities.

and instead uses inversion and subsequent knowledge transfer (SR $\rightarrow$ FKT). We also investigate possible overgeneralization of inversion. Similar to Symmetry, we find that the model overgeneralizes the symmetric component of inversion. In particular, the model seems to have difficulties differentiating between symmetry and inversion. We find that for a test-fact (e, r_s, f) in the training set, the model predicts (f, r_s, e) and (f, r_t, e) for 20% of facts, i.e. inferring symmetric facts wihtout changing the relation.

Source	de-fr-es	en-fr-es	en-de-es	en-de-fr
Target	en	de	fr	es
rule-transfer				
baseline	13.6	6.1	18.7	12.8
general	25.7	11.1	36.3	29.5
rule-specific	29.8	24.1	18.8	33.5
SR+FKT				
baseline	45.4	29.5	46.5	41.4
general	**62.6**	**37.8**	**60.8**	**54.8**
rule-specific	44.4	30.6	38.1	34.2

Table 4.11: Mean P@1 of zero-shot cross-lingual transfer of inversion and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) with multiple source languages. The baseline is the best transfer from a single source to the target language with general relations. In bold is the overall best transfer performance to the target language.

4.5 Negation

In negation (NEG) the model needs to form a connection between the antonyms `a` and `b` through a relation **r** and its negation **not r**, generalizing across subjects:

$$(e, r, a) \iff (e, not\ r, b)$$

Therefore the model needs to see antonym-pairs (a, b) in a similar context during training to learn their association, especially with synthetic antonyms - entities that are used as if they were antonyms. For each antonym-pair two facts are created, one with the relation and one with its negation. As an example, facts in the training set might be (jupiter, is, big) and (jupiter, is not, small) where 'jupiter' is the subject, 'is' and 'is not' are the relation and the respective negation. 'big' and 'small' are the antonyms. Note that we use different subject-entities for train- and test-facts to prevent the model from learning to just associate entities with antonyms, independent of the relation. For each relation we create a number of antonym-pairs p, which are only used with that relation. The number of antonym-pairs is a trade-off between generalization and overfitting to co-occurrences of relation and antonym. The fewer antonyms we have, the more the model can learn to generalize across different subjects since we have more facts per antonym. However, it is also more susceptible to only learning co-occurrences. As a result, the model could guess the correct antonym based on the relation instead of inferring based on the memorized antonym-fact. The probability of the model guessing antonyms associated with the relation is $1/p$. We choose $p = 20$ for our experiments, so the highest mean P@1 that can be achieved by exploiting co-occurrences is 5%.

We find that mBERT has high transfer performance with the negation-rule for single source languages as shown in Table 4.12. Here, rule-specific relations are relation-negation-pairs that were manually corrected for grammar mistakes. However, we found that there is little difference in performance between general and rule-specific relations. When we use multiple source languages (Table 4.13), we find that the increase is not as high as for other symbolic rules and mP@1 even decreases for the transfer to Spanish. While rule-transfer is high compared to other rules, mBERT is not capable of combining symbolic reasoning and factual knowledge transfer. When investigating this further, we find that the model does not learn the negation-rule very well in the source language but seems to rather have high rule-transfer efficiency. Therefore, when the model tries to combine factual knowledge transfer with symbolic reasoning, the cumulative transfer loss is too high.

Since relations and negations are fairly similar, we suspected that the model would tend to overgeneralize the negation-rule. However, we found that for (e, r, a) the model never predicts $(e, not\ r, a)$ and therefore does not overgeneralize. In contrast, Kassner et al. (2020) and Kassner and Schütze (2020) report an overgeneralization of negation. The different results are possibly due to the different settings. While Kassner et al. train BERT from scratch and Kassner and Schütze analyse negation learned during pretraining, we fine-tune mBERT explicitly on a negation corpus. Future work could analyse if the different results are in fact due to mBERT learning negation better than BERT or if our results are solely due to fine-tuning.

Source	en			de			fr			es		
Target	de	fr	es	en	fr	es	en	de	es	en	de	fr
rule-transfer												
general	19.7	18.7	21.4	15.5	13.7	15.3	15.4	17	18.9	18.8	19.4	24.1
rule-specific	16.5	16.3	18.9	16	17.3	19	17.6	17.7	15.6	14.7	18.1	13.5

Table 4.12: Mean P@1 of zero-shot cross-lingual transfer of negation and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) with Latin-European-languages. We use 20 antonym-pairs; as a result, learning co-occurrences has a performance of 5% at best.

Source	de-fr-es	en-fr-es	en-de-es	en-de-fr
Target	en	de	fr	es
baseline	18.8	19.7	24.1	21.4
rule-transfer	26	29.6	25.7	15.8

Table 4.13: Mean P@1 of zero-shot cross-lingual transfer of negation and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) with multiple source languages. The baseline is the best transfer from a single source to the target language with general relations. In bold is the overall best transfer performance to the target language. We use 20 antonym-pairs; as a result, learning co-occurrences has a performance of 5% at best.

4.6 Implication

An implication (IMP) consists of a premise and one or multiple conclusions:

$$(e, r, f) \Longrightarrow (e, s, a), (e, s, b), (e, s, c)$$

Unlike prior rules, this can result in possibly multiple implied facts. Note that IMP and NEG are highly similar since the relation and its negation are just a special case of $\mathbf{r}$ and $\mathbf{s}$ and the entity-tuples (f, a, b, c) are essentially analogous to antonym-pairs. Only the number of conclusions is an additional parameter. For simplicity we choose to use only one conclusion and 20 entity-tuples in line with the antonym-pairs in the NEG experiments. Due to their similarity and the added complexity of having dissimilar relations, we suspected to have similar or worse results on cross-lingual implication-transfer. Instead, we find that with a training accuracy of over 99% the rule is not learned in the source language for any language-pair and barely even learned with multiple source languages. We hypothesize that NEG can benefit from the similarity between the relation and its negation while implication lacks proper structure to be learned. In contrast, Kassner et al. (2020) report that BERT learns implication surprisingly well. The difference in performance is likely due to the different settings. While they train BERT from scratch, we fine-tune mBERT. However, mBERT should be capable

of benefiting from its pretraining and therefore be able to learn implication. We leave further investigations to future work.

4.7 Composition

Finally, we discuss composition (COMP), a two-hop rule where two premises have to be associated with each other to create a conclusion as so:

$$(e, r, f) \wedge (f, s, g) \implies (e, t, g)$$

This is particularly difficult since not only one but two premises need to be memorized and then reasoned upon, especially since mBERT already had difficulties with not only cross-lingual transfer of IMP but also learning it in the source language.

We find that the model is able to apply composition in the source language and even capable of transferring composition to the target language. This is surprising since mBERT could not learn IMP in the source language, although it has a high similarity with COMP. Furthermore, the model can even combine applying the composition-rule and factual knowledge transfer to infer new facts across languages. Our setup is similar to IMP and NEG. However, we choose 100 entity-pairs (f, g) per relation. The results for composition with single source languages and Latin-European languages are shown in Table 4.14. mBERT achieves up to 37.6% mean P@1 for composition-transfer with English-to-Spanish and 33.3% when combining composition and factual knowledge transfer. Transfer to German performs poorly and is barely above chance. When we investigate this further we find that German tends to overfit very heavily to the point where performance collapses. Reducing the training epochs from 200 to 100, we get a higher mP@1 with 20.5% (en-de), 11% (es-de). However, since we did our hyperparameter search over the average performance of Latin-languages, we do not report the improvements in our Table. In general, composition seems very unstable and sensitive to hyperparameter tuning. Among Non-Latin languages (Table 4.15), the highest composition-transfer is achieved by English-to-Russian with 50.6% but in general are comparable to Latin-European language performance. We also investigated the combination of multiple source languages. The results can be seen in Table 4.16. German-French-Spanish-to-English obtains with 65% the highest rule-transfer and English-Spanish-French to German again the lowest with 30.5%. These results are remarkable since we suspected that COMP is comparably challenging to IMP but strangely, the model seems to be able to learn COMP much easier although showing high instability during training. Furthermore, we find that when we use the same relation for the premises and conclusion $r = s = t$, performance collapses. This indicates that the model relies on the relations as an important signal during transfer.

Source	en			de			fr			es		
Target	de	fr	es	en	fr	es	en	de	es	en	de	fr
rule-transfer	26	31.3	37.6	14.2	14.3	14.4	27.3	13.4	29.9	30.3	2	32.6
SR+FKT	20.3	23.7	33.3	9.9	16.1	10.7	30.9	2	32.6	32.6	1	28

Table 4.14: Mean P@1 of zero-shot cross-lingual transfer of composition and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) with Latin-European-languages.

Source	en			ru		zh		ja	
Target	ru	zh	ja	zh	ja	ja	ru	zh	ru
rule-transfer	50.6	12.5	20.2	12	4.4	32.1	26.8	25	12.2
SR+FKT	19.1	4.5	12.4	9.5	15.6	37.3	18.9	26.9	12.4

Table 4.15: Mean P@1 of zero-shot cross-lingual transfer of composition and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) for general relations on Non-Latin-languages with Latin entities.

In accordance with Kassner et al. (2020), we also investigate if adding semantic structure facilitates mBERT's ability to learn composition (*COMP Enhanced*). We divide the entities **e**, **f**, **g** into 3 separate groups. For each group, connecting-facts are added to the training set such as 'e1 connected e2', where e1 and e2 belong to the same group. However, we find that adding these semantic structures only adds noise to the training process. mBERT appears to be able to better identify the actual relevant structure without. Kassner et al.'s success with their version of *COMP Enhanced* may be due to the fact that their model has not yet learned to easily identify structure in the training data since they train their model from scratch and the data lacks natural variation.

4.8 Discussion

We studied the ability of MLLMs to utilize symbolic reasoning across languages to infer previously unseen factual knowledge. For this we trained mBERT on facts that demonstrate a particular symbolic rule in the source language and facts that are used to evaluate its ability to transfer the rule. Composition, negation and equivalence show the highest transfer performance, while symmetry and inversion are severely lacking with inversion performing slightly better. In contrast, implication is neither learned in source or target language, although training accuracy is high. This is surprising since we expected implication, negation and composition to have comparable performance due to their similarity. However, composition performs vastly better than implication. We speculate that it benefits from the additional structure of its rule and from the model already knowing

Source Target	de-fr-es en	en-fr-es de	en-de-es fr	en-de-fr es
rule-transfer				
baseline	30.3	26	32.6	37.6
general	41.2	**30.5**	40.1	49
enhanced	8.8	5.8	6.6	7.3
SR+FKT				
baseline	32.6	20.3	28	33.3
general	**50**	15.9	**41.7**	**58.5**
enhanced	9.3	6.3	6.9	8

Table 4.16: Mean P@1 of zero-shot cross-lingual transfer of composition and the combined application of symbolic reasoning and factual knowledge transfer (SR+FKT) with multiple source languages. The baseline is the best transfer from a single source to the target language with general relations. In bold is the overall best transfer performance to the target language.

all necessary entity-information through the premises in the test-facts. As a result, the model only needs to assemble the target-fact entities and infer the relation. This might be easier than inferring entities and explain why equivalence performs vastly better than implication, although they are similar as well. With the negation-rule as a special case of implication, we hypothesize that the good performance is due to the model exploiting the similarity between relation and negation.

The performance of language-pairs for rule-transfer correlates with our results from factual knowledge transfer. Especially Spanish-to-French and English-to-Spanish performed well across symbolic rules and had high transfer performance in factual knowledge transfer. Transfer to and from German performed generally the worst as they did in factual knowledge transfer. Further analyses should investigate if this is due to the model or a bias in the data. Non-Latin languages often performed worse than Latin-languages. Since we used language-agnostic facts, we opted for Latin-entities with Non-Latin relations as this was more consistent across languages than our alternative approach of using source entities. The Chinese-Japanese language-pair performed much closer to the Latin-languages, possibly due to the lexical overlap between Chinese and Japanese. In line with the results of the previous chapter, we also found that using multiple source languages increased the performance of rule-transfer. Only negation seems to barely benefit from the additional information. We distinguished between general and rule-specific relations in our experiments. By using rule-specific relations, the model could leverage its pretraining knowledge for higher performance. However, we found that this only worked for equivalence.

In our analysis of the predictive behaviour of each symbolic rule, we found that symmetry and inversion suffer from overgeneralization which seems to relate to the symmetric component of both relations. While the model tends to predict that some relations are symmetric without evidence, the model mostly distinguishes symmetric from non-symmetric relations. The overgeneralization of symmetry was also observed by Kassner et al. (2020), although much stronger. Further, they find that BERT can not properly distinguish between negated and positive facts in NEG. In contrast, mBERT showed no signs of overgeneralization for negation in our experiments. This is surprising since relation and its negation are almost identical. We hypothesize that this might be due to differences in our setup but needs further investigation.

Moreover, we investigated mBERT's ability to combine symbolic reasoning and factual knowledge transfer. For this we trained the model on test-facts in the source language instead of the target language. Therefore when we evaluate the model on corresponding target-facts in the target language, it has to infer them through a combined application of symbolic reasoning and factual knowledge transfer. As previously discussed, we identify two paths that the model can take (Fig. 4.2). Either it transfers the facts to the target language and then uses symbolic reasoning (FKT $\rightarrow$ SR) or the model applies symbolic reasoning in the source language and then uses factual knowledge transfer (SR $\rightarrow$ FKT) to infer the facts in the target language. Our results have revealed that the model has a substantially higher transfer performance when it combines symbolic reasoning and factual knowledge transfer compared to using rule-transfer. When combining them, the model seems to apply the symbolic rule in the source language and then transfer the result through factual knowledge transfer to the target language (SR $\rightarrow$ FKT). In general, we find that almost all symbolic rules can be combined with factual knowledge transfer. Only negation exhibits deficiencies in its application in both source and target language. When combined with factual knowledge transfer, these deficiencies lead to cumulative transfer loss and therefore the target-facts can not be inferred. However, compared to other symbolic rules negation exhibits a rather high rule-transfer efficiency in our experiments, i.e. the model shows almost the same performance for applying the negation rule in source and target language.

Chapter 5

Conclusion

In this work, we presented a comprehensive study on the emergence of factual knowledge in pretrained MLLMs.

Our results revealed that mBERT is capable of zero-shot cross-lingual factual knowledge transfer. More specifically, the model has the ability to share memorized factual knowledge across languages, possibly due to shared language-agnostic representations. While the transfer performance for language-agnostic entities is high, transfer for multilingual entities is rather poor. We hypothesize that this is due to deficiencies in the alignment of language-specific sub-spaces. To improve alignment, we can use multiple source languages that act as a parallel corpus. Even more effective is adding a small amount of parallel facts in the source and target language. Furthermore, by replacing multilingual entities with language-agnostic entities the transfer performance substantially increases. These likely function as anchor points and by that aid alignment. When using language-agnostic entities, cross-lingual factual knowledge transfer is even possible for languages in different scripts such as Russian, Chinese and Japanese. We analysed the various factors that influence the zero-shot cross-lingual transfer and discovered that mBERT's transfer capabilities and therefore its factual knowledge is often underreported in our experiments since the choice of relation label leads to a selection bias. Instead probing the model not only on the relation but also on its aliases increases the lower-bound on the transfer performance substantially. Furthermore, our results show that a matching word order between the facts in the source and target language is necessary for cross-lingual transfer. If the word order diverges, transfer fails. We also find that adding a dot at the end of our facts can have an impact on the transfer performance, likely because mBERT is more confident about its predictions due to the sentence having ended.

Besides mBERT, we studied the zero-shot cross-lingual factual knowledge transfer on monolingual BERT. Surprisingly, we found that the model has the ability to learn factual knowledge through different languages than it was trained on and even languages in different scripts. We hypothesize this is due to the pretraining corpus containing artifacts from other languages leading monolingual BERT to create language-agnostic representations.

Furthermore, we investigated if the model can apply symbolic reasoning across languages to infer previously unseen factual knowledge. This can be done by either directly sharing the symbolic rule or through a combined application of symbolic reasoning and factual knowledge transfer. Our results revealed that mBERT can transfer most symbolic rules from a source language to a target language. While equivalence, negation and composition show high transfer performance, inversion performs not as good and symmetry rather poorly. The last two rules also tend to be overgeneralized by the model, in particular their symmetric component. Additionally, we found that implication is not learned in the source or target language. This is surprising due to the similarity of the rule with negation and composition. In contrast, Kassner et al. (2020) report that BERT can learn implication rather well but is not capable of learning symmetry and inversion correctly. While they report the overgeneralization of symmetry as well, our model can distinguish between symmetric and non-symmetric relations much better. They also report that BERT is not capable of learning negation properly and can not differentiate between positive and negative facts, unlike our model. Furthermore, BERT could learn composition only with added structural information. This is likely due to differences in our setup. While Kassner et al. train BERT from scratch, we only fine-tune mBERT and therefore can utilize its pretraining. When we analysed the predictive behaviour for the symbolic rules, we found that the model can even combine factual knowledge transfer and symbolic reasoning. In particular, the model applies symbolic reasoning in the source language and then uses factual knowledge on the result to infer new facts in the target language.

Overall, mBERT shows surprising abilities to infer unseen factual knowledge across languages. While the model displays deficiencies in its transfer, these can be partially overcome in our experiments by improving alignment through several different methods. However, if these were overcome in a more systematic manner for pretraining, it will likely lead to more robust language-agnostic representations and therefore more sharing across languages. In addition, the model could use its parameters more efficiently and increase overall memorization. Similarly, improved reasoning capabilities could lead to a higher inference of new factual knowledge. Our experiments also show the necessity of alternative prompt generation to reduce selection bias and reduce the performance variance as the model can be very susceptible to small changes such as adding a dot. This increases the lower-bound on overall knowledge. Finally, our results lead us to agree with Dufter and Schütze (2020) that training on comparable corpora like Wikipedia could explain mBERT's surprising cross-lingual abilities.

5.1 Future Work

We leave it to future work to find a more systematic way of improving the alignment of MLLMs, e.g. by implementing knowledge adapters (Wang et al., 2021) that are trained on entity alignment. Additionally, deficiencies could also be alleviated by using a multilingual objective to promote language-agnostic representations during pretraining. This has the potential to vastly improve cross-lingual performance and alleviate the lack of data for low-resource languages. While our work was done through fine-tuning mBERT, a more intricate analysis could be possible by having full control over the pretraining process. Extending our investigation to a multi-token setting would reduce not only selection bias but also lead to more general results. We would also like to see an extension of our reasoning investigation to more difficult reasoning tasks, MLLMs could have the ability to learn more complicated reasoning which the model could then utilize to create more factual knowledge. Lastly, while using fact triples is a common setup, we would like to see studies that investigate the transfer of factual knowledge with natural language since this would give us more insights into the behaviour on natural text corpora.

Bibliography

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the Cross-lingual Transferability of Monolingual Representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, 2020. doi: 10.18653/v1/2020.acl-main.421.

Alexis Conneau and Guillaume Lample. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale, April 2020.

Leandro Rodrigues de Souza, Rodrigo Nogueira, and Roberto Lotufo. On the ability of monolingual models to learn language-agnostic representations. *arXiv:2109.01942 [cs]*, October 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019.

Philipp Dufter and Hinrich Schütze. Identifying Elements Essential for BERT's Multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.358.

Yoav Goldberg. Assessing BERT's Syntactic Abilities, January 2019.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. *arXiv:2003.11080 [cs]*, September 2020.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models. *arXiv:2010.06189 [cs]*, October 2020a.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How Can We Know What Language Models Know? *arXiv:1911.12543 [cs]*, May 2020b.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. *arXiv:1912.07840 [cs]*, February 2020.

Nora Kassner and Hinrich Schütze. Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. *arXiv:1911.03343 [cs]*, May 2020.

Nora Kassner, Benno Krojer, and Hinrich Schütze. Are Pretrained Language Models Symbolic Reasoners Over Knowledge? *arXiv:2006.10413 [cs]*, October 2020.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models. *arXiv:2102.00894 [cs]*, February 2021.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers. *arXiv:2005.00633 [cs]*, May 2020.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation. *arXiv:2004.01401 [cs]*, May 2020.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. How Language-Neutral is Multilingual BERT? *arXiv:1911.03310 [cs]*, November 2019.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. On the Language Neutrality of Pre-trained Multilingual Representations. *arXiv:2004.05160 [cs]*, September 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language Models as Knowledge Bases? *arXiv:1909.01066 [cs]*, September 2019.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is Multilingual BERT? *arXiv:1906.01502 [cs]*, June 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. page 24, 2019.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. oLMpics – On what Language Model Pre-training Captures. *arXiv:1912.13283 [cs]*, November 2020.

Marc Tanti, Lonneke van der Plas, Claudia Borg, and Albert Gatt. On the Language-specificity of Multilingual BERT and the Impact of Fine-tuning. *arXiv:2109.06935 [cs]*, September 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Technical report, 2017.

Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57:78–85, 2014. URL http://cacm.acm.org/magazines/2014/10/178785-wikidata/fulltext.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.121.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July 2020.

Shijie Wu and Mark Dredze. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. *arXiv:1904.09077 [cs]*, October 2019.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging Cross-lingual Structure in Pretrained Language Models. *arXiv:1911.01464 [cs]*, May 2020.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Emergence of Factual Knowledge in Pretrained Multilingual Language Models

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Paech	Laurin

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
01.08.2022	Laurin

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.