Machine-Created Universal Language for Cross-Lingual Transfer

Yaobo Liang, Quanzhi Zhu, Junhe Zhao, Nan Duan

Microsoft Research Asia
{yaobo.liang, v-quanzhizhu, v-junhezhao, nanduan}@microsoft.com

Abstract

There are two primary approaches to addressing cross-lingual transfer: multilingual pre-training, which implicitly aligns the hidden representations of various languages, and translate-test, which explicitly translates different languages into an intermediate language, such as English. Translate-test offers better interpretability compared to multilingual pre-training. However, it has lower performance than multilingual pre-training and struggles with word-level tasks due to translation altering word order. As a result, we propose a new Machine-created Universal Language (MUL) as an alternative intermediate language. MUL comprises a set of discrete symbols forming a universal vocabulary and a natural language to MUL translator for converting multiple natural languages to MUL. MUL unifies shared concepts from various languages into a single universal word, enhancing cross-language transfer. Additionally, MUL retains language-specific words and word order, allowing the model to be easily applied to word-level tasks. Our experiments demonstrate that translating into MUL yields improved performance compared to multilingual pre-training, and our analysis indicates that MUL possesses strong interpretability. The code is at: https://github.com/microsoft/Unicoder/tree/master/MCUL.

Introduction

Cross-lingual transfer aims to tackle NLP tasks in multiple languages using training data from only one or a few languages, such as English. There are two primary approaches to addressing cross-lingual transfer: first, multilingual pre-training involves constructing a multilingual encoder, fine-tuning it in English, and directly testing it in other languages. The multilingual encoder combines words from all target languages to create a large vocabulary, and the hidden representations in the intermediate layers are implicitly aligned to facilitate the cross-lingual transfer. Second, the translate-test approach translates the test set of other languages into an intermediate language, typically English. This allows the model to use English as input for both training and testing, explicitly solving cross-lingual tasks.

Compared to multilingual pre-training, translate-test offers better interpretability by utilizing an intermediate language. However, it has two drawbacks: Translate-test yields

worse performance compared to cross-lingual transfer. For instance, its performance on XNLI is 3.1% lower than that of multilingual pre-training (Conneau et al. 2020). Translate-test cannot be applied to word-level tasks such as sequential labeling or machine reading comprehension, as translation alters the word order.

To retain the interpretability of an intermediate language while addressing its limitations, we propose to create a new language specifically designed for cross-lingual tasks. This language, created by machines without requiring human expertise, is called the Machine-created Universal Language (MUL). MUL consists of a set of discrete symbols that form a universal vocabulary and an NL-MUL translator for converting multiple natural languages (NL) to MUL. The NL-MUL translator maps shared concepts from different languages to the same universal words, facilitating better cross-lingual transfer. Additionally, it preserves word order and language-specific vocabulary, allowing for easy application to word-level tasks. This is consistent with the research presented by Chai, Liang, and Duan 2022, which indicates that word order does not affect cross-lingual abilities, thus allowing for the preservation of distinct word orders in different languages. To solve cross-lingual NLP tasks, we can translate both the English training dataset and the multilingual test dataset into MUL, enabling the model to use MUL as input for both training and testing.

To create MUL, our approach consists of three components: First, we pre-train the encoder using multilingual MLM loss and generate word alignment supervision on bilingual data, with the word alignment supervision being created through an unsupervised method. Second, we employ an inter-sentence contrastive learning approach to further enhance the alignment of contextualized word embeddings across languages. Lastly, we introduce vector quantization with cross-lingual alignment (VQ-CA) to improve the interpretability of the universal vocabulary.

We conduct experiments on XNLI, NER, MLQA, and Tatoeba using MUL as input. Compared to the combined vocabulary in multilingual pre-training, our model has a smaller vocabulary size and necessitates fewer parameters at the word embedding layer. We obtain comparable results to XLM-R with 50% fewer parameters and achieve superior results after redistributing the parameters from word embedding to the transformer's weights. Further analysis reveals

that MUL exhibits strong interpretability, as translating to MUL results in less ambiguity compared to translating to English.

Our work offers two significant contributions. First, we introduce a new universal language, MUL, along with a translator between multiple natural languages and MUL. Our experiments demonstrate that translating to MUL achieves strong cross-lingual transfer performance and exhibits good interpretability. Second, we propose an innovative approach to create MUL, which incorporates inter-sentence contrastive learning and vector quantization with cross-lingual alignment.

Related Work

Multilingual pre-training was first proposed by mBERT (Devlin et al. 2019b), which extended the pre-training of BERT (Devlin et al. 2019a) to 100 languages by creating a large vocabulary for all languages and building a multilingual encoder. To improve the cross-lingual transfer performance, lots of works extended monolingual pre-training methods to multiple languages and achieved good cross-lingual performance, such as XLM-Roberta (Conneau et al. 2020) extended Roberta (Liu et al. 2019), mT5 (Xue et al. 2021) extended T5 (Raffel et al. 2020), XLM-E (Chi et al. 2022) extended Electra (Clark et al. 2020). These methods can be improved by introducing bilingual data (Conneau and Lample 2019; Huang et al. 2019) or multilingual knowledge (Jiang et al. 2022) to improve the implicitly cross-lingual alignment between different languages. All of these works take natural language as input and achieve cross-lingual transfer by implicitly cross-lingual alignment. Translate-test is a baseline of XNLI proposed by Conneau et al. 2018. Further experiments show that both XLM (Conneau and Lample 2019) and XLM-R (Conneau et al. 2020) can achieve better performance compared to the translate-test baseline. Our work achieves better performance compared to XLM-R by translating all data to MUL.

Abstract Meaning Representation(AMR) (Banarescu et al. 2013) targets to map natural language sentence to abstract graph, and can server as the transfer layer in MT system (Xue et al. 2014). Our work share the same motivation and propose new methods for cross-lingual pre-training.

VQ-VAE is proposed by van den Oord, Vinyals, and Kavukcuoglu 2017 to create discrete symbols in the neural network, which is usually used to create discrete symbols for image (Ramesh et al. 2021; Esser, Rombach, and Ommer 2021), video (Wu et al. 2022) and audio (Baevski et al. 2020). It's rare to be applied to natural language which is already discrete symbols. The symbols in our MUL have better interpretability than the symbols for other modalities.

Methodology

In this section, we begin by defining the Machine-created Universal Language (MUL) and providing an overview of its creation process. Following that, we present the detailed steps involved in creating MUL, including multilingual masked language modeling (MLM), inter-sentence contrastive learning, and vector quantization with cross-lingual alignment.

Machine-Created Universal Language (MUL)

MUL comprises a set of discrete symbols that form a universal vocabulary, along with an NL-MUL translator and a MUL-NL translator for translating between multiple natural languages and MUL.

Each symbol in the universal vocabulary is defined as a universal word. Each universal word corresponds to a concept identified by the model. Most universal words can be aligned with words in multiple natural languages, explicitly facilitating cross-lingual transfer. Some universal words correspond to specific words in certain languages, helping to understand linguistic features unique to those languages.

The NL-MUL translator aims to translate natural languages into MUL. It preserves the word order and generates one universal word for each natural word, which assists the model in solving word-level tasks such as sequential tagging and machine reading comprehension. The mapping relationship between natural words and universal words is context-dependent, meaning a single natural word may correspond to different universal words in varying contexts. Therefore, the translation from NL to MUL involves word disambiguation, which can reduce the model's difficulty in accomplishing specific tasks. The MUL-NL translator, on the other hand, restores NL from MUL and calculates the auto-encoder loss during the MUL creation process.

When addressing cross-lingual NLP tasks, we can employ the NL-MUL translator to convert both the English training dataset and the multilingual test dataset into MUL, which can then be used as input for the model.

Overview of MUL Training

In order to create MUL, we initially construct an encoder capable of generating contextualized word embeddings for each sentence. For two words with context in different languages, their embeddings are close to one another if and only if they share the same meaning. Subsequently, we create discrete symbols in the embedding space, to ensure that each symbol corresponds to a single concept.

Our approach comprises three components, and we demonstrate their impact on the embedding space with Figure 1. First, we pre-train the encoder using a multilingual masked language model (MLM) loss. The embeddings are depicted in Figure 1.a. Although different words with similar meanings do not have similar embeddings, the encoder can be employed to create unsupervised word alignment labels for bilingual sentence pairs (Dou and Neubig 2021).

Second, we implement an inter-sentence contrastive learning approach to enhance the alignment of contextualized word embeddings across languages. The results can be observed in Figure 1.b, which shows that different words with the same meanings now have similar embeddings.

Lastly, we introduce vector-quantization with cross-lingual alignment (VQ-CA) to establish the universal word list in the universal language. Figure1.b and Figure1.c represent training without and with VQ-CA, respectively. The
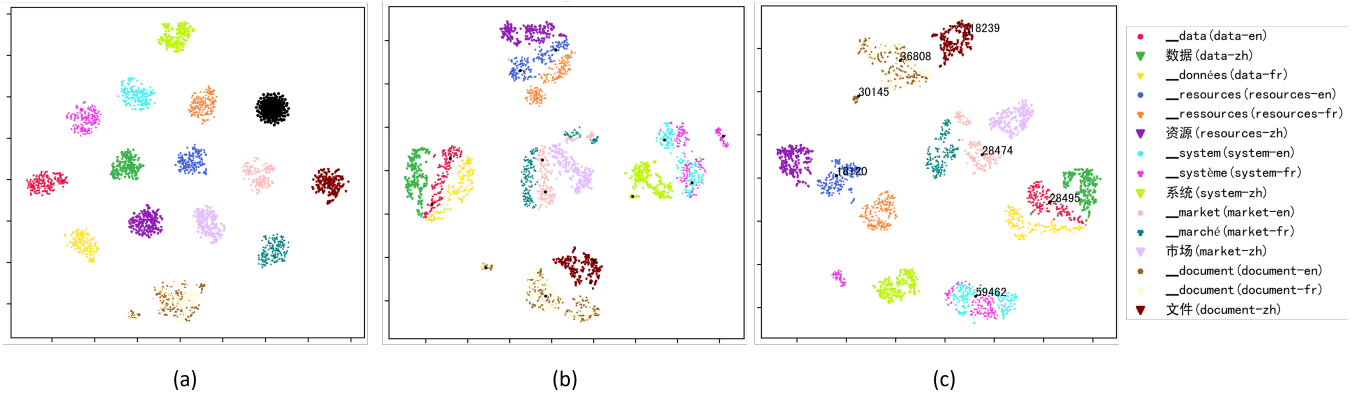
Figure 1: The visualization of contextualized word embeddings at various training stages. Each color represents a word and each point denotes the contextualized embedding of that word in different contexts. Figure 1.a displays the embeddings after pre-training with multilingual MLM, Figure 1.b exhibits the embeddings after inter-sentence contrastive learning, and Figure 1.c demonstrates the embeddings following VQ-CA.

black points in these figures are the embeddings of the created universal words. For each group of words with the same meanings, we observe that the model trained without VQ-CA generates multiple universal words, while the model trained with VQ-CA produces a single universal word in most instances.

Creating Word Alignment Supervision by Multilingual MLM

First, we pre-train our encoder $Encoder(x)$ using a multilingual Masked Language Model (MLM). This encoder has a vocabulary that includes words from all target languages, as well as a transformer encoder comprising 12 layers.

The contextualized word embeddings generated by the pre-trained encoder demonstrate good performance on the word alignment task (Dou and Neubig 2021). Specifically, the word alignment task involves processing two sentences, S_s and S_t, from different languages that have the same meanings. These sentences consist of n and m tokens, respectively, which can be represented as $S_s = s_1, s_2, ..., s_n$ and $S_t = t_1, t_2, ..., t_m$. The model's objective is to identify the aligned words or phrases between these two sentences.

We input the two sentences into the pre-trained model $Encoder(x)$ to obtain their contextualized representations, $H_s = Encoder(S_s) = h_{s_1}, h_{s_2}, ..., h_{s_n}$ and $H_t = Encoder(S_t) = h_{t_1}, h_{t_2}, ..., h_{t_m}$. The alignment matrix is then computed by $A = H_s H_t^T$.

Next, we apply the softmax function to the first and second dimensions to obtain A_{t2s} and A_{s2t}, respectively. The word alignment results are determined by $P = A_{t2s} > c \wedge A_{s2t} > c$, where c represents the threshold. Intuitively, this approach identifies the most similar words in the S_t sentence for each word s_i in S_s and vice versa. If both s_i and t_j are the most similar words to each other, they are predicted to be aligned words.

Inter-sentence Contrastive Learning

While the pre-trained contextualized word embeddings can achieve good cross-lingual word alignment performance, there are still two notable shortcomings. Firstly, the distance between aligned words is not close to zero, even though they are the most similar words between the source and target sentences, as illustrated in Figure1.a. Secondly, the distance between words of the same type is too close, and it becomes even closer when the model is trained with vanilla contrastive loss. For instance, words of the same type can include time-related terms such as "year", "month", "day", and "hour", or adverbs of frequency like "always", "never", and "sometimes". In bilingual sentences, there is typically only one or a few words for each type. Consequently, being adept at identifying words of the same type is sufficient for achieving good word alignment performance. However, such granularity is too coarse for MUL.

To address this issue, we propose inter-sentence contrastive learning. This approach has two main steps. First, we employ contrastive learning to minimize the distance between aligned words while maintaining a larger distance between non-aligned words. Second, we utilize words from other sentence pairs as negative samples to ensure that words of the same type remain distant from one another.

In the contrastive learning process, we consider all aligned words in matrix P as positive pairs. We perform post-processing on the unaligned words in P and represent the negative matrix as $N \in {0, 1}^{n \times m}$. The contrastive loss is defined as

$$loss_{cts} = -log \sum_{i,j} P_{ij} \exp H_{s_i} H_{t_j}^T$$
$$+ log \sum_{i,j} \left(P_{ij} \exp H_{s_i} H_{t_j}^T + N_{ij} \exp H_{s_i} H_{t_j}^T \right)$$

In inter-sentence contrastive learning, we sample multiple bilingual pairs and generate a new pair by concatenating the source and target sentences, respectively. For example,

consider two pairs: (S_s^1, S_t^1) and (S_s^2, S_t^2). The new pair is $([S_s^1, S_s^2], [S_t^1, S_t^2])$. We create the positive alignment matrices P^1 and P^2 for the two pairs separately. Subsequently, we merge the two positive alignment matrices and construct a positive matrix for the concatenated sentence pair:

$$P_{inter} = \begin{bmatrix} P^1 & 0 \\ 0 & P^2 \end{bmatrix}$$

This means that we won't treat any pairs between (S_s^1, S_t^2) and (S_s^2, S_t^1) as positive alignment. We avoid concatenating the two sentences initially to generate P_{inter} directly, as this could introduce additional interference in word alignment and diminish alignment quality. By employing this method, we can effectively push words of the same type further apart. For negative pairs, we apply the same postprocessing technique.

Vector Quantization with Cross-lingual Alignment (VQ-CA)

To create a universal vocabulary, one option is using VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2017) to learn a set of discrete symbols. However, the symbols generated by VQ-VAE lack clear meanings and are difficult for humans to comprehend. For instance, in Figure1.b, multiple symbols are created for each meaning, and each symbol lacks a precise definition. So we propose Vector-Quantization with Cross-Lingual Alignment (VQ-CA) to guide the learning of discrete symbols by aligning them with multiple languages simultaneously. In most cases, the symbols produced by VQ-CA correspond to a single concept, making them easier to understand compared to those created by VQ-VAE.

We define the embedding of universal vocabulary as $e = \{e_1, e_2, ..., e_K\}$, where $e_i \in R^D$ is the embedding of discrete symbol i. K is the size of the universal vocabulary and D is the dimension of hidden representation.

Our model comprises an $Encoder(x)$ and a $Decoder(x)$. The $Encoder(x)$ contains word embedding layers and multiple transformer layers. For a sentence $S = s_1, s_2, ..., s_n$, we map it to contextualized word embeddings $H = Encoder(S) = h_1, h_2, ..., h_n$. We generate the sentence in the universal language by mapping each contextualized word representation h_i to symbol $k_i = Quantize(h_i) = \arg\min_j \|e_j - h_i\|_2$. The sentence in MUL is $S_u = k_1, k_2, ..., k_n$ and its embedding is $E = e_{k_1}, e_{k_2}, ..., e_{k_n}$. The $Encoder(x)$ and $Quantize(x)$ together form the NL-MUL translator. The $Decoder(x)$ consists of several transformer layers and a softmax layer, which can generate the probability of mapping the sentence embedding in MUL back to natural language as $P(S|E) = Decoder(E)$. The $Decoder(x)$ serves as the MUL-NL translator.

To train the $Encoder(x)$, $Decoder(x)$ and universal vocabulary embedding e, our loss is:

$$loss_{VQ-CA} = logP(S|E) + \|sg(E) - H\|_2 \\ + \lambda\|E - sg(H)\|_2 + loss_{CA}$$

The notation $sg(x)$ represents the stop gradient operation. The first three losses are derived from VQ-VAE. The
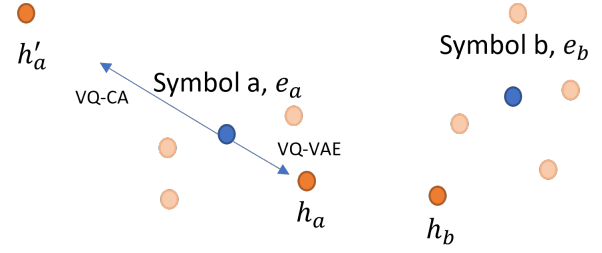

Figure 2: Visualization of VQ-CA. The orange dots shows the embeddings related to a pair of aligned words. The light orange dots shows the embeddings that map to symbol a and b.

first term is the auto-encoder loss, which aims to recover the original natural language sentences from the MUL sentences. The second term constraint contextualized word embeddings to be close to universal language embeddings. In our experiment The third term constraints universal language embeddings to be close to contextualized word embeddings. In our experiments, we find that the update speed for the embeddings of the universal language is too slow. Consequently, we replace the third loss with exponential moving averages, following the approach of van den Oord, Vinyals, and Kavukcuoglu 2017.

The fourth loss, L_{CA}, constrains the aligned words to map to the same symbol. For aligned words that map to different symbols, the loss L_{CA} pushes one symbol away and retains only the other symbol in the nearby region. Consequently, both words can be mapped to the preserved symbol, ensuring that aligned words share the same symbol in the universal language representation.

We illustrate $loss_{CA}$ in Figure 2. Formally, let's consider two aligned words with embeddings h_a and h_b. We quantize them to two symbols $a = quantize(h_a)$ and $b = quantize(h_b)$. The original VQ-VAE loss requires the symbol a to move towards h_a and symbol b to move towards h_b. However, in $loss_{CA}$, one symbol should be pushed away. The selection of which symbol to push away is determined by the number of natural language words that are mapped to it. Without loss of generality, we assume that a should be pushed away. Then we create an embedding $h_a' = e_a + \lambda * (e_a - h_a)$ at the opposition direction of h_a, and add a new loss $loss_{CA} = \|e_a - h_a'\|_2$ for it. We also use exponential moving averages to update e_a. This loss moves e_a in the direction opposite to that of the VQ-VAE. Once a has been moved far away from h_a, the nearest symbol of h_a may change to b. As training progresses, symbol b will dominate the region of symbols a and b, while symbol a will fade away.

Experiments

In this section, we begin by presenting the training details, followed by experiments on four diverse cross-lingual tasks. Lastly, we conduct the ablation study to examine the different components of our method.

Model	Parameter	en	de	fr	es	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	avg
mBERT	178M	82.1	73.8	74.3	71.1	66.4	68.9	69	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
XLM	250M	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	**76.5**	69.6	68.4	67.3	75.1
XLM-R Base	278M	85.3	78.3	79.2	79.9	77.3	78.6	76.1	74.7	73.8	75.6	**73.3**	74.6	71.7	68.6	68.2	75.7
mT5 Base	580M	84.7	77.4	79.1	80.3	77.1	78.6	77.1	72.8	73.3	74.2	73.2	74.1	70.8	69.4	68.3	75.4
Unicoder	278M	85.4	78.2	79.2	79.8	77.3	78.5	76.7	73.8	73.9	75.9	71.8	74.7	70.1	67.4	66.3	75.3
InfoXLM	278M	86.4	79.3	80.3	80.9	77.8	79.3	77.6	75.6	74.2	77.1	74.6	77.0	72.2	67.5	67.3	76.5
MUL Small	132M	84.0	78.5	79.5	79.9	78.4	79.0	75.8	74.4	74.8	75.8	70.9	73.8	70.8	71.1	68.1	75.7
MUL Base	277M	**85.5**	**80.5**	**81.1**	**81.4**	**79.8**	**80.6**	**78.4**	**75.9**	**77.4**	**78.4**	72.8	76.0	**73.8**	**72.9**	**69.9**	**77.6**

Table 1: Evaluation results on XNLI.

Model	NER	MLQA	Tatoeba
XLM-R Base	61.9	65.6 / 47.9	63.4
mT5 Base	59.5	64.4 / 45.0	-
InfoXLM	-	68.1 / 49.7	77.8
MUL Small	60.8	65.6 / 47.4	74.6
MUL Base	**63.0**	**69.4 / 50.8**	**79.3**

Table 2: Evaluation results on three cross-lingual tasks.

Training Details

In the first stage, we pre-train the encoder with a multilingual MLM objective on 15 languages of XNLI. The vocabulary size is 250K, and the model contains 12 layers and 768 hidden states, identical to XLM-R base. Limited by resources, we pre-train the model for 500K steps with a batch size of 8192, which is less than XLM-R Base. The pre-training corpus is CC-Net (Wenzek et al. 2020).

In the second stage, we train our model on bilingual data OPUS-100 (Zhang et al. 2020). The encoder has 8 layers, and the decoder has 4 layers. They are initialized by the first 8 and last 4 layers of the encoder pre-trained in the first stage. We select 8 as encoder layers because previous research(Dou and Neubig 2021) shows that outputs of the 8th layer have the best cross-lingual alignment quality. The size of the universal vocabulary K is set to 60K, as the vocabulary size of GPT is 50K.

Once we have the encoder, decoder, and universal vocabulary, we proceed with pre-training and fine-tuning on MUL. As both pre-training and fine-tuning require multiple epochs, we translate the corpus into MUL during the pre-processing stage, saving significant time. The vocabulary size is reduced from 250K to 60K. We try two sets of model sizes: MUL Small and MUL Base. The small model has the same layer number and hidden size as XLM-R Base, with the total parameter number being only half of XLM-R Base, due to the reduction in vocabulary size. The base model reallocates parameters from embedding layers to transformer layers, keeping the total parameter number unchanged. The hyper-parameters in pre-training and fine-tuning are the same as those of natural language. We run all fine-tuning experiments four times and report the average of the results.

Performance on Cross-lingual Tasks

We test MUL on four diverse cross-lingual tasks: cross-lingual Natural Language Inference (XNLI)(Conneau et al. 2018) is a sentence classification task; NER(Pan et al. 2017)

is a sequential labeling task; MLQA (Lewis et al. 2020) is a machine reading comprehension task; Tatoeba (Artetxe and Schwenk 2019) is a cross-lingual sentence retrieval task. We only use English training data in the first three tasks and don't use any training data in Tatoeba.

We compare our model with six baseline models that use natural language as input. The first three models are pre-trained exclusively on monolingual datasets: mBERT (Devlin et al. 2019b) and XLM-R Base (Conneau et al. 2020) share the same pre-training objective as ours, while mT5 (Xue et al. 2021) is pre-trained using a denoising objective. The last three models are pre-trained on both monolingual and bilingual datasets: XLM (Conneau and Lample 2019) employs multilingual MLM and TLM in the 15 languages of XNLI. Unicoder (Liang et al. 2020) and InfoXLM (Chi et al. 2021) introduce new bilingual objectives; their monolingual datasets are the same as our model, but their bilingual datasets are larger. For a fair comparison, we continue to pre-train XLM-R Base on the 15 languages of XNLI.

We show the performance of XNLI for each language in Table 1, and present the results on NER, MLQA, and Tatoeba in Table 2. Based on the results, we can draw three conclusions: 1) MUL Base achieves the best performance on all tasks with the same parameter number as XLM-R Base, Unicoder, and InfoXLM. This demonstrates that taking MUL as input can achieve excellent cross-lingual transfer performance. 2) MUL Small also achieves comparable performance to baselines with minimal parameters. On Tatoeba, it achieves better performance compared to XLM-R Base and slightly lower than InfoXLM, which introduces sentence-level contrastive learning. On XNLI, MLQA, and NER, MUL Small can achieve comparable results to baselines. 3) On XNLI, both MUL Small and Base achieve good performance on low-resource languages, such as Swahili (sw) and Urdu (ur).

Ablation Study

We evaluate the quality of MUL using performance on word alignment and XNLI:

Word alignment with MUL We translate natural language sentences into MUL and predict aligned words by checking if they correspond to the same universal word. Word alignment can help us understand whether words with the same meanings are mapped to the same universal word. We report three metrics: precision, recall, and alignment error rate (AER). We don't train our model on the word alignment training dataset and directly evaluate it on the test

Setting	Precision ↑	Recall ↑	AER ↓	XNLI ↑
MUL (pair=4)	90.0	**51.2**	**35.3**	**74.0**
w/o VQ-CA	89.1	42.3	43.4	73.7
w/o contrastive loss + VQ-CA	69.2	11.5	80.9	69.7
w/o inter-sentence contrastive loss (pair=1)	90.1	46.0	39.8	72.9
inter-sentence contrastive loss (pair=2)	**90.5**	49.5	36.6	73.4

Table 3: The ablation study of MUL. The first row is the best setting in our paper which uses inter-sentence contrastive on 4 pairs of sentences. We skip pre-training and only fine-tuning in these experiments to reduce computational costs.

Language	Sentence in Natural Language and MUL
English	He is sitting in a chair _He/45816 _is/36575 _sitting/44530 _in/43023 _a/29017 _chair/43227
French	Il est assis sur une chaise _Il/45816 _est/55230 _assis/44530 _sur/7419 _une/29017 _chaise/43227
Chinese	他正坐在椅子上 _他/45816 正/36575 坐在/44530 椅/43227 子/32533 上/17777
English	She serves as the chair in our committee _She/28168 _serves/13352 _as/45554 _the/50140 _chair/38789 _in/43267 _our/10816 _committee/59378
French	Elle est la présidente de notre comité _Elle/28168 _est/22312 _la/50140 _présidente/38789 _de/29699 _notre/10816 _comité/59378
Chinese	她在我们的委员会中担任主席 _她/28168 在/50000 我们的/10816 委员会/59378 中/10923 担任/43518 主席/38789

Table 4: The examples to translate natural language sentences into universal language. For each example, we show the results of tokenization and the universal word corresponding to each token.

dataset. We evaluate our model on German-English (de-en), French-English (fr-en), and Chinese-English (zh-en) and report the averaged results. The test datasets come from Mihalcea and Pedersen 2003; Vilar, Popović, and Ney 2006; Liu and Sun 2015 respectively.

XNLI results We report the results on XNLI to evaluate the quality of using MUL as input to solve cross-lingual tasks. We don't conduct pre-training on MUL in the ablation study limited by resources. In fine-tuning, we load the transformer weight of the pre-trained encoder. The word embedding of each universal word is the weighted sum of its corresponding natural words, and the weights are the frequency of the corresponding natural words.

The results of the ablation study are presented in Table 3 and include two aspects:

Ablation of loss After removing VQ-CA, the recall of word alignment drops about 10 percent. This is because the model without VQ-CA often generates multiple universal words for the same concept. As a result, aligned words are mapped to different concepts even if they have similar contextualized word embeddings. After removing both of them, the performance on word alignment becomes very poor, and the performance on XNLI drops significantly. This is because the embeddings of aligned words are far from each other and are mapped to different universal words.

Ablation of inter-sentence contrastive learning The inter-sentence contrastive learning leverages multiple sentence pairs, and we report the performance of 1, 2, and 4 sentence pairs. Using one sentence pair means vanilla contrastive and removes the inter-sentence strategy. We find that a larger number of sentence pairs leads to better performance both on word alignment and XNLI. However, increasing the sentence pair numbers also increases GPU memory usage and training time, so we can only set it to 4.

Analysis

We conduct the analysis focusing on three aspects: the interpretability of MUL, the word disambiguation in NL-MUL translation, and the language-specific words in MUL.

The Interpretability of MUL

To better understand MUL, we show two groups of examples in Table 4. Each group contains three sentences in English, French, and Chinese, all with the same meanings. We first tokenize these sentences and then translate them into MUL.

To understand the meaning of each universal word, we can summarize the natural words that are often translated into it. In Table 5, we list the top 2 natural words that correspond to the universal word in three languages. For example, the universal word "43227" corresponds to "chaise" in French and "chair in Chinese", which helps us to know that "43227" means a chair, which is a piece of furniture for one person to sit on. Similarly, we can deduce that "38789" means the person in charge of the meeting based on "président" in French.

For most words in different languages with the same meanings, their universal words are the same as each other. By mapping to the same universal words, knowledge can

Universal Word	English Word	French Word	Chinese Word
43227	_chair:46%, _wheelchair:33%	_fauteuil:44%, _chaise:35%	椅:95%, 凳:3%
38789	_Chair:40%, _Chairman:28%	_président:48%, président:21%	主席:98%, 院长:1%
36575	_is:68%, s:32%	_est:49%, _a:18%	正在:64%, 正:22%
29017	_a:92%, _an:7%	_un:37%, _une:28%	一个:69%, 一:11%
45816	_He:94%, _he:4%	_Il:89%, _il:5%	_他:95%, _He:2%
28168	_She:77%, _she:21%	_Elle:76%, _elle:19%	_她:81%, 她:17%
10816	_our:100%, _Our:0%	_notre:57%, _nos:41%	我们:81%, 我国:9%

Table 5: For each universal word, we list the top 2 natural words that correspond to it in three languages.

be easily transferred between languages, enabling effective cross-lingual learning and understanding.

chair	chairman	Chair (seat)
30320	2	0
38789	100	13
43227	11	102
53430	2	0

Table 6: The statistics of relation between universal words and different meanings of "chair".

apple	apple_inc	apple (fruit)
18766	668	44
20027	224	848

Table 7: The statistics of the relation between universal words and different meanings of "apple".

club	club	nightclub	Club (weapon)
50064	54	54	54

Table 8: The statistics of the relation between universal words and different meanings of "club".

Word Disambiguation of NL-MUL Translation

For a word that has different meanings in different contexts, it may correspond to different universal words during the NL-MUL translation. For example, in Table 4, "chair" means "furniture" in the first group and means "a person" in the second group, so it corresponds to different universal words. Compared to natural words, the meaning of universal words is closer to concepts shared across multiple languages. This makes them less ambiguous. For example, we can distinguish the meaning of 43227 and 38789, while we can't distinguish the meanings of two instances of "chair" without context.

We conduct more statistical experiments on the CoarseWSD-20 dataset (Loureiro et al. 2021) and present the results in Table 6, Table 7, and Table 8. We can find that the universal words of "chair" and "apple" have a good correlation to concepts, while the universal words for "club" are the same in most cases. This is because most

of the "club" instances in bilingual data correspond to the first concept, only 3% of "club" means "nightclub," and almost no "club" means "club (weapon)." This shows that translating to MUL can disambiguate parts of words, but the disambiguation is not good enough due to the unbalanced distribution of concepts in our data.

During translation, two different words in non-English languages may be translated into the same word in English. This increases the ambiguity of words. However, when translated into different universal words, this ambiguity is reduced. By using universal words, the difficulty of solving NLP tasks is decreased.

Conclusion

In this work, we present a new universal language MUL created by machines, which can serve as an intermediate language and solve cross-lingual tasks by translating all languages into MUL. We introduced inter-sentence contrastive learning and VQ-CA which are critical to creating MUL. The experiments show that the model with MUL as input achieves excellent cross-lingual performance and greatly reduces the size of vocabulary size. Further analysis shows the good interpretability of MUL and the capability for word disambiguation.

References

Artetxe, M.; and Schwenk, H. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the ACL 2019*.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33: 12449–12460.

Banarescu, L.; Bonial, C.; Cai, S.; Georgescu, M.; Griffitt, K.; Hermjakob, U.; Knight, K.; Koehn, P.; Palmer, M.; and Schneider, N. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, 178–186.

Chai, Y.; Liang, Y.; and Duan, N. 2022. Cross-Lingual Ability of Multilingual Masked Language Models: A Study of Language Structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4702–4712.

Chi, Z.; Dong, L.; Wei, F.; Yang, N.; Singhal, S.; Wang, W.; Song, X.; Mao, X.-L.; Huang, H.-Y.; and Zhou, M.

2021. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3576–3588.

Chi, Z.; Huang, S.; Dong, L.; Ma, S.; Zheng, B.; Singhal, S.; Bajaj, P.; Song, X.; Mao, X.-L.; Huang, H.-Y.; et al. 2022. XLM-E: Cross-lingual Language Model Pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6170–6182.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, É.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

Conneau, A.; and Lample, G. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Conneau, A.; Rinott, R.; Lample, G.; Williams, A.; Bowman, S.; Schwenk, H.; and Stoyanov, V. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2475–2485.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019a. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019b. Multilingual Bert.

Dou, Z.; and Neubig, G. 2021. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, 2112–2128. Association for Computational Linguistics.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Huang, H.; Liang, Y.; Duan, N.; Gong, M.; Shou, L.; Jiang, D.; and Zhou, M. 2019. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2485–2494.

Jiang, X.; Liang, Y.; Chen, W.; and Duan, N. 2022. XLM-K: Improving Cross-Lingual Language Model Pre-training with Multilingual Knowledge.

Lewis, P.; Oğuz, B.; Rinott, R.; Riedel, S.; and Schwenk, H. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proceedings of ACL 2020*.

Liang, Y.; Duan, N.; Gong, Y.; Wu, N.; Guo, F.; Qi, W.; Gong, M.; Shou, L.; Jiang, D.; Cao, G.; et al. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6008–6018.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y.; and Sun, M. 2015. Contrastive unsupervised word alignment with non-local features. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Loureiro, D.; Rezaee, K.; Pilehvar, M. T.; and Camacho-Collados, J. 2021. Analysis and Evaluation of Language Models for Word Sense Disambiguation. *Computational Linguistics*, 47(2): 387–443.

Mihalcea, R.; and Pedersen, T. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*, 1–10.

Pan, X.; Zhang, B.; May, J.; Nothman, J.; Knight, K.; and Ji, H. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of ACL 2017*, 1946–1958.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P. J.; et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140): 1–67.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 6306–6315.

Vilar, D.; Popović, M.; and Ney, H. 2006. AER: Do we need to "improve" our alignments? In *Proceedings of the Third International Workshop on Spoken Language Translation: Papers*.

Wenzek, G.; Lachaux, M.-A.; Conneau, A.; Chaudhary, V.; Guzmán, F.; Joulin, A.; and Grave, É. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4003–4012.

Wu, C.; Liang, J.; Ji, L.; Yang, F.; Fang, Y.; Jiang, D.; and Duan, N. 2022. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European Conference on Computer Vision*, 720–736. Springer.

Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; and Raffel, C. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 483–498.

Xue, N.; Bojar, O.; Hajič, J.; Palmer, M.; Urešová, Z.; and Zhang, X. 2014. Not an Interlingua, But Close: Comparison of English AMRs to Chinese and Czech. In Calzolari, N.; Choukri, K.; Declerck, T.; Loftsson, H.; Maegaard, B.; Mariani, J.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Zhang, B.; Williams, P.; Titov, I.; and Sennrich, R. 2020. Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1628–1639.