Fine-grained Multi-lingual Disentangled Autoencoder for Language-agnostic Representation Learning

Zetian Wu*
Oregon State University
wuzet@oregonstate.edu

Zhongkai Sun
Amazon Alexa AI
zhongkas@amazon.com

Zhengyang Zhao
Amazon Alexa AI
zzhengya@amazon.com

Sixing Lu
Amazon Alexa AI
cynthilu@amazon.com

Chengyuan Ma
Amazon Alexa AI
mchengyu@amazon.com

Chenlei Guo
Amazon Alexa AI
guochenl@amazon.com

Abstract

Encoding both language-specific and language-agnostic information into a single high-dimensional space is a common practice of pre-trained Multi-lingual Language Models (pMLM). Such encoding has been shown to perform effectively on natural language tasks requiring semantics of the whole sentence (e.g., translation). However, its effectiveness appears to be limited on tasks requiring partial information of the utterance (e.g., multi-lingual entity retrieval, template retrieval, and semantic alignment). In this work, a novel Fine-grained Multilingual Disentangled Autoencoder (FMDA) is proposed to disentangle fine-grained semantic information from language-specific information in a multi-lingual setting. FMDA is capable of successfully extracting the disentangled template semantic and residual semantic representations. Experiments conducted on the MASSIVE dataset demonstrate that the disentangled encoding can boost each other during the training, thus consistently outperforming the original pMLM and the strong language disentanglement baseline on monolingual template retrieval and cross-lingual semantic retrieval tasks across multiple languages.

1 Introduction

Pre-trained multilingual language models such as multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-RoBERTa (XLM-R) (Conneau et al., 2020) have been extensively explored and used in academia and industry. These models encode both language-specific information (e.g., grammar, tense, syntax) and language-agnostic information (e.g., semantic, entity, sentiment) into one high-dimensional embedding. However, it has been demonstrated that such encoders perform poorly in some tasks due to a lack of capacity to disentangle fine-grained language-agnostic and language-specific information. (Tiyajamorn et al., 2021; Wieting et al., 2020; Roy et al., 2020; Ahuja et al., 2020; Lin et al., 2021; Asai et al., 2020).

Table 1 presents several application examples where disentangled language-specific or language-agnostic encoder might have better performance. The first example is cross-lingual retrieval, in which a English utterance with the same language-agnostic semantic can be retrieved by a German utterance. Note the semantics in both template and slot can be maintained. The second example is a template retrieval, in which "do you show me doing" can be replaced by another similar meaning template "can you show me how to do" while the slot value changes. The third example is paraphrase retrieval, in which the target utterance keeps slot text but rephrases the template part. Although the applications shown above are retrieval tasks but all of them can be used in query reformulation (Ponnusamy et al., 2020, 2022) and data augmentation (Xu et al., 2021; Kale and Rastogi, 2020; Liu et al., 2021; Gao et al., 2022). For instance, the source utterance "do you show me doing [exercise: backflip]" is a defective sentence with grammar error, and the disentangled encoder is able to retrieve a similar meaning but grammar correct utterance by ignoring the uncommon slot value "backflip".[1] The tasks of cross-lingual retrieval and paraphrase retrieval both are commonly used for data augmentation, especially for languages with data scarcity.

In this work, we proposed a lightweight encoding architecture called Fine-grained Multilingual Disentangled Autoencoder (FMDA) that can disentangle semantic representations at different aspects. The training of proposed encoder adopts reconstruction loss and contrastive learning. The contributions of our proposal are as follows:

*This work is finished during the internship at Amazon Alexa AI

[1]For this example, the golden reformulation for the defective query would be "can you show me how to do backflip". However, directly performing utterance-level retrieval may fail to find the golden reformulation because of data scarcity. Therefore, template-level retrieval is useful here.

Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22), pages 12 - 24

Application Tasks	Utterances	Language	Semantic	Template	Slot
cross-lingual retrieval	source: wecke mich um [time: fünf uhr] auf target: wake me up at [time: five am]	different	same	same/similar	same
template retrieval	source: do you show me doing [exercise: backflip] target: can you show me how to do [exercise: yoga]	same	related	same/similar	different
paraphrase retrieval	source: what can be seen inside [object: the basket] target: what does [object: the basket] mainly contain	same	same	different	same

Table 1: Applications of Disentangled Semantic

1. The FMDA is able to extract embedding of:

- language agnostic template representation that contains the semantic information related to the sentence backbone. E.g., in the sentence "Can you play the music Green Light", the template representation aims at encoding the semantic of the "Can you play the music []";
- language agnostic meaning representation which contains both template semantic and residual semantic (e.g., slot name "Green Light") information;
- language-specific non-semantic representation that contains unique language facts;

Visualizations of these fine-grained embedding representations are shown in Figure 4 in Section 4.5.

2. The FMDA designs multiple contrastive learning objectives to improve the performance of the disentanglement learning.

3. Compared with the original pMLM and a language-disentanglement SOTA (Tiyajamorn et al., 2021), FMDA achieves significant improvement on both monolingual template retrieval and cross-lingual meaning retrieval tasks, evaluated on the benchmark MASSIVE (FitzGerald et al., 2022).

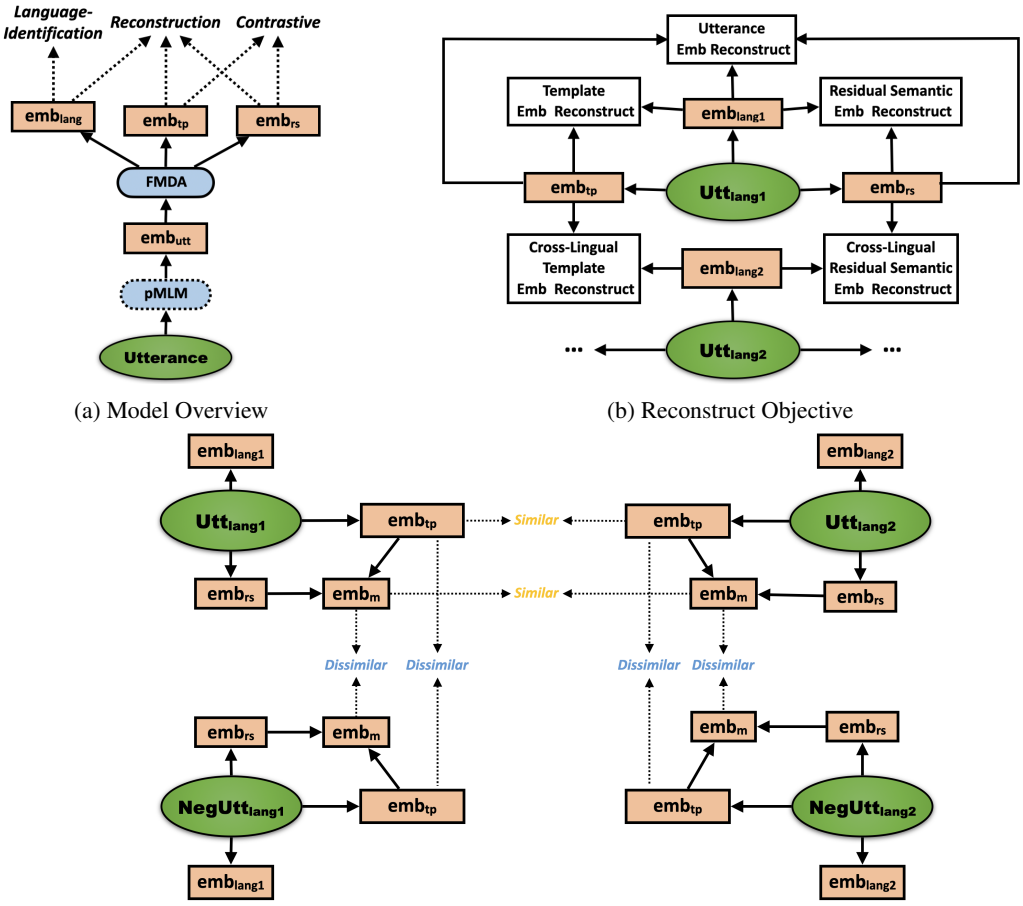
4. An ablation study further proves the effectiveness of our model, and a two-stage training experiment has been conducted to further study the effect of the fine-grained semantic.

2 Related Work

Multilingual sentence encoders are widely studied and applied to downstream tasks in recent years. Self-attention networks based multilingual sentence encoders, such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), are pretrained on multilingual corpora in over 100 languages. LaBSE (Feng et al., 2022) encodes text to multilingual sentence embedding by training with 100 million sentence pairs in bilingual corpora of 109+ languages. Libovický et al. (2020) proposes a centered embedding method that subtracts the mean embedding for each language from the sentence embedding, as well as a projection embedding method that projects bilingual using a parallel corpus. MUSE (Chidambaram et al., 2018; Yang et al., 2019) applies a translation based ranking task to one-billion weblab QA pairs to obtain a multilingual universal encoder. Multilingual SBERT (Reimers and Gurevych, 2020) extends pre-trained monolingual SBERT (Reimers and Gurevych, 2019) to the multi-lingual version by mapping translations and original utterances into the same space.

Beyond atomic encoding, some research also focus on disentangling language specific and language-agnostic embeddings. (Chen et al., 2019) learns to disentangle language syntax and semantic information by using aligned paraphrase data to train semantic and use word-order information to train syntax. BGT (Wieting et al., 2019) utilizes a deep variational probabilistic model together with transformers to learn better semantic embeddings in a bi-lingual setting by excluding language-specific information from the information shared across languages. Tiyajamorn et al. (2021) proposes a method for distilling language-agnostic meaning embeddings by removing language-specific information from sentence embeddings generated by off-the-shelf multilingual sentence encoders. Although these works extract both language-specific and language-agnostic embeddings, they are hard to support fine-grained semantic disentanglement.

Based on the work of Tiyajamorn et al. (2021), we further extend the semantic extraction to a fine-grained level. Specifically, our proposed method FMDA is able to extract not only the semantic information of the whole utterance but also part of it, i.e. template/carrier phrase semantic information by learning to disentangle language information at different levels.



(a) Model Overview

(b) Reconstruct Objective

(c) Contrastive Objective

Figure 1: The Overview of our model. (a) The FMDA outputs language-specific non-semantic embedding (emb_{lang}), the template semantic embedding (emb_{tp}), and the residual-semantic embedding (emb_{rs}). Three objectives are applied to train the FMDA: language identification, reconstruction, and contrastive learning. (b) The emb_{tp}, emb_{rs}, and emb_{lang1} are used to reconstruct the original pMLM's utterance embedding; the emb_{lang1} together with The emb_{tp} or emb_{rs} are used for template embedding or residual-semantic embedding reconstruction, respectively; the emb_{lang2} from the utterance in another language but with same meaning can be used with emb_{tp} and emb_{rs} for the cross-lingual reconstruction; (c) Contrastive learning objectives are applied to both emb_{tp} and emb_m (obtained from both emb_{tp} and emb_{rs}).

3 Method

This section describes the details of our proposed Fine-grained Multilingual Disentangled Autoencoder (FMDA). Figure 1(a) demonstrates the overview of the method. FMDA is trained to extract language-specific non-semantic embedding, template semantic embedding, and residual-semantic embedding from a pMLM with three objectives: language identification, embedding reconstruction, and contrastive learning. Figure 1(b) presents the objectives of the embedding reconstruction. Figure 1(c) shows the detailed contrastive learining design, in which FMDA tries to minimize embedding difference between two paired utterances in different languages and maximize the embedding difference between two non-related utterances in the same language. Each training step takes four utterances as inputs: utterance in language 1, utterance with same-semantic in language 2, and negative utterances with different semantic in language 1 and 2.

3.1 Fine-grained Model Details

Figure 2 demonstrates the detail of our proposed fine-grained model. Given an utterance, a pMLM is first utilized to extract the utterance embedding $emb_{utt} \in R^{L \times D}$, where L represents the utterance length and D represents the embedding dimension. Note that the pMLM's parameters are frozen and the emb_{utt} will remain unchanged during the training and inference. So that the original informa-

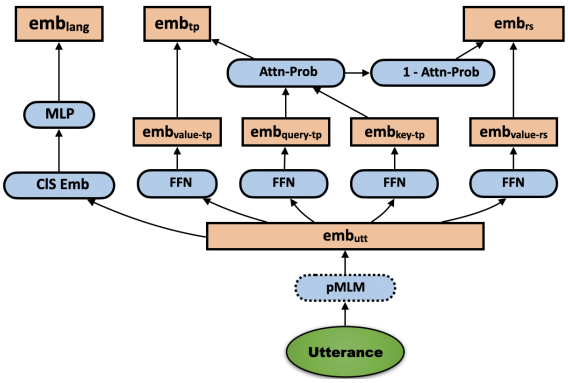


Figure 2: The details of the FMDA. (a) given the utterance embedding emb_{utt} obtained from the pMLM, a MLP layer is first applied to the $[CLS]$ token embedding to generate the emb_{lang}; four FFN layers are applied to the emb_{utt} to generate the template value-embedding ($Value_{tp}$), template query-embedding ($Query_{tp}$), template key-embedding (Key_{tp}), and the residual-semantic value-embedding ($Value_{rs}$). The $Query_{tp}$ and Key_{tp} are first used to generate the attention-probability $Attn - Prob$ for template, therefore the final template embedding can be generated using $Value_{rp}$ and $Attn - Prob$. Besides, the $1 - Attn - Prob$ is also calculated to represent the residual-semantic attention, which can then be used with the $Value_{rs}$ to generate the final emb_{rs}.

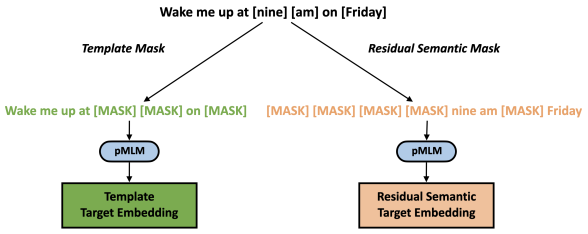


Figure 3: Overview of how to generate the template and residual-semantic reconstruction target embeddings.

tion encoded by the pMLM will be reserved and our method can be lightweight to extract the embeddings of interest. Next, FeedForward-Network (FFN) layers are applied to the emb_{utt} to extract three different embeddings:

Language-Specific Non-semantic Embedding The $[CLS]$ token embedding ($emb_{cls} \in R^{1\times D}$) of the emb_{utt} is extracted and input to a Multi-Layer-Perceptron (MLP) to obtain the language-specific non-semantic embedding $emb_{lang} \in R^{1\times D}$.

Language-agnostic Template Semantic embedding In order to encode the semantic information more effectively, an attention-based method is applied to the emb_{utt}. Specifically, three different FFN layers are applied to the emb_{utt} to extract the template value embeddings ($emb_{value-tp} \in R^{L\times D}$), template query-embeddings ($emb_{query-tp} \in R^{L\times D}$), and the template key-embeddings ($emb_{key-tp} \in R^{L\times D}$). The template attention-probability Attn-prob $\in L \times L$ can then be calculated as Softmax($emb_{query-tp} \cdot emb^T_{key-tp}$). The calculated Attn-prob is thereby used with the template value embeddings ($emb_{value-tp}$) to obtain the attention based template embeddings sequence $emb_{tp-seq} \in R^{L\times D}$): $emb_{tp-seq} = \text{Attn-prob} \cdot emb_{value-tp}$.

Finally, the $[CLS]$ token position of the emb_{tp-seq} is extracted as the final template semantic embedding $emb_{tp} \in R^{1\times D}$.

Language-agnostic residual-semantic embedding FMDA leverages the Attn-prob learned from the generation of template embedding to generate the embedding for residual semantic. The motivation is to disentangle the template semantic and residual-semantic as much as possible, i.e., the information that template does not pay attention to should be used more to generate the residual-semantic embedding.

Therefore, given the template Attn-prob learned above, we first calculate its opposite (1 - Attn-prob) to obtain the residual-semantic Attn-prob $\in R^{L\times L}$: RS Attn-prob $= \text{softmax}(1 - \text{Attn-prob})$. Then, the residual-semantic embedding sequence $emb_{rs} \in R^{1\times D}$ can be obtained.

3.2 Language Identification Objective

To ensure that the extracted language-specific embedding emb_{lang} contains correct language information, following the idea in Tiyajamorn et al. (2021), the emb_{lang} is used for a language-identification objective. Specifically, the emb_{lang} will be input to a MLP layer to achieve the language prediction P, which is used with the true language label L to calculate the language-identification loss $Loss_{lang} = \text{CrossEntropy}(P, L)$.

At each training step, the $Loss_{lang}$ is calculated for both of the utterance in language 1 and the utterance with same semantic in language 2.

3.3 Reconstruction Objective

Utterance Embedding Reconstruction As shown in Figure 1(b), for the utterance in language 1 and its emb_{utt} obtained from the pMLM, the $emb_{lang1}, emb_{tp}, emb_{rs}$ learned from FMDA are expected to not lose any in-

formation in the original emb_{utt}. Therefore, $emb_{lang1}, emb_{tp}, emb_{rs}$ are used to reconstruct the emb_{utt}. MSE (mean-squared-error) is used as the loss function, thus the $Loss_{rec}^{utt-l1}$ for the utterance in language 1 can be defined as $Loss_{rec}^{utt-l1} = MSE(MLP(emb_{lang1} + emb_{tp} + emb_{rs}), emb_{utt})$

Similarly, the $Loss_{rec}^{utt-l2}$ is also calculated for the utterance with same semantic in language 2.

Template Embedding Reconstruction The emb_{lang} and emb_{tp} are used to reconstruct the template target embedding $emb_{tp}^{Tgt} \in R^{1\times D}$.

Figure 3 demonstrates how to obtain the target template embedding emb_{tp}^{Tgt-L1} together with the residual-semantic target embedding emb_{tp}^{Tgt-L1} for the utterance in language 1. Specifically, given an utterance with token labels indicating if a token belongs to the template or slots, the original utterance is then masked according to the labels and input to the pMLM to obtain the $[CLS]$ position's embeddings as the template target embedding and residual-semantic target embedding.

Both emb_{lang} and emb_{tp} are used to reconstruct the emb_{tp}^{Tgt} as emb_{tp} should contain language-agnostic template-semantic only, while the template target emb_{tp}^{Tgt} contains some language-specific information obtained from the pMLM.

In order to calculate the loss, the sum of the emb_{tp} and emb_{lang1} is obtained as $emb_{tpL1} \in R^{1\times D}$, which is then used to reconstruct the emb_{tp}^{Tgt-L1}. Therefore, the $Loss_{rec}^{cpL1}$ can be defined as:

$$Loss_{rec}^{tpL1} = MSE(emb_{tpL1}, emb_{tp}^{Tgt-L1}) \quad (1)$$

Besides, the emb_{tp} is also combined with emb_{lang2}, which is obtained from the utterance with the same semantic but in language 2, to build the to $emb_{tpL1-CL}$ to reconstruct the template target embedding in language2 (emb_{tp}^{Tgt-L2}). This objective further guarantees that the emb_{tp} contains language-agnostic template semantic information. The cross-lingual template loss $Loss_{rec}^{tpL1-CL}$ can then be defined as:

$$Loss_{rec}^{tpL1-CL} = MSE(emb_{tpL1-CL}, emb_{tp}^{Tgt-L2}) \quad (2)$$

The same process is also applied to the utterance with the same semantic in language 2, to calculate the $Loss_{rec}^{tpL2}$ and $Loss_{rec}^{tpL2-CL}$.

Residual-semantic Embedding Reconstruction Similar reconstruction objectives are applied to the residual-semantic. After obtaining the residual-semantic target embedding emb_{rs}^{Tgt} as illustrated in Figure 3, the $Loss_{rec}^{rsL1}$ and $Loss_{rec}^{rsL1-CL}$ can be calculated for the utterance in language 1 with the same process in the template reconstruction.

Similarly, the functions $Loss_{rec}^{rsL2}$ and $Loss_{rec}^{rsL2-CL}$ are calculated for the utterance with the same semantic in language 2.

3.4 Contrastive Objective

Figure 1 (c) demonstrates the idea of the contrastive learning objective. The contranstive learning is applied to both of the template semantic embeddings and the utterance-level semantic (meaning) embeddings. For the utterances with the same semantic in different languages, their template semantic embeddings and meaning embeddings should be similar; for the utterances in the same language but with different semantics, their template semantic embeddings and meaning embeddings should be different.

Given the inputs as the followings: utterance in language 1, utterance with same-semantic in language 2, and negative utterances with different semantic in language 1 and 2, the contrastive loss for the template semantic $Loss_{con}^{tp}$ can be calculated as:

$$\begin{aligned} Loss_{con}^{cp} = &- \text{Cos-sim}(emb_{tp}^{L1}, emb_{tp}^{L2}) \\ &+ \text{Cos-sim}(emb_{tp}^{L1}, emb_{neg-tp}^{L1}) \\ &+ \text{Cos-sim}(emb_{tp}^{L2}, emb_{neg-tp}^{L2}) \end{aligned} \quad (3)$$

where emb_{tp}^{L1} and emb_{tp}^{L2} come from the utterances with the same semantic but in language 1 and 2; emb_{neg-tp}^{L1} and emb_{neg-tp}^{L2} denote the template semantic embeddings of negative utterances that with different semantic in language 1 and 2.

To conduct the contrastive learning for the meaning embedding, a MLP is first used to generate the meaning embedding $emb_m \in R^{1\times D}$ using the sum of emb_{tp} and emb_{rs}. Then the contrastive loss for the meaning $Loss_{con}^{m}$ can be calculated as:

$$\begin{aligned} Loss_{con}^{m} = &- \text{Cos-sim}(emb_{m}^{L1}, emb_{m}^{L2}) \\ &+ \text{Cos-sim}(emb_{m}^{L1}, emb_{neg-m}^{L1}) \\ &+ \text{Cos-sim}(emb_{m}^{L2}, emb_{neg-m}^{L2}) \end{aligned} \quad (4)$$

Similarly, emb_{m}^{L1} and emb_{m}^{L2} represent the utterance-level semantic meaning of the utterances

with the same semantic but from language 1 and 2; emb^{L1}_{neg-m} and emb^{L2}_{neg-m} denote the meaning embeddings from negative utterances that with different semantic from language 1 and 2, respectively.

3.5 Total Training Objective

During the training stage, all of the objectives' loss functions will be optimized together. Therefore, the total loss can be written as:

$$\begin{aligned} Loss_{total} = & \, Loss^{L1}_{lang} + Loss^{L2}_{lang} \\ & + Loss^{utt-L1}_{rec} + Loss^{utt-L2}_{rec} \\ & + Loss^{tpL1}_{rec} + Loss^{tpL2}_{rec} \\ & + Loss^{tpL1-CL}_{rec} + Loss^{tpL2-CL}_{rec} \\ & + Loss^{rsL1}_{rec} + Loss^{rsL2}_{rec} \\ & + Loss^{rsL1-CL}_{rec} + Loss^{rsL2-CL}_{rec} \\ & + Loss^{tp}_{con} + Loss^{m}_{con} \end{aligned} \quad (5)$$

4 Experiments

This section describes the experiments conducted on various language pairs using the multilingual natural language understanding dataset MASSIVE (FitzGerald et al., 2022). To evaluate the proposed FMDA, two retrieval tasks introduced in Table 1 are used: (1) cross-lingual semantic retrieval, where the goal is to find the best semantically matching utterance pairs from two languages; and (2) monolinugal template retrieval, where the goal is to find utterances with different slot values but similar template in one language.

4.1 Dataset

Both training and evaluation of our experiment were conducted using the MASSIVE dataset (FitzGerald et al., 2022), which is a cross-lingual corpus that contains virtual assistant utterances across 51 languages. Domains, intents, and slots have been labeled for each utterance.

We chose four languages from MASSIVE - English (EN), German (DE), Spanish (ES) and Japanese (JA) - to form three language pairs (EN-DE, EN-ES, EN-JA) to conduct the experiment. Such selection covers both languages that are similar (e.g. EN-DE) and languages that belongs to distant families (e.g. EN-JA). Our training and evaluation sets were prepared by pre-processing on MASSIVE's train split (containing 11k utterances in each language) and test split (containing 2974 utterances in each language), respectively.

4.2 Setup

XLM-R (base) (Conneau et al., 2020) is used as the backbone encoder to train our proposed FMDA model on three language pairs: EN-DE, EN-ES and EN-JA. As described in Section 3, each training step takes four utterances as inputs: a pair of parallel utterances from language 1 and language 2, and negative utterances with different semantic in language 1 and 2. The following is an example of the training data for EN-DE language pair:

utt_en: Wake me up at nine am on Friday.
utt_de: Weck mich am freitag um neun uhr auf.
neg_utt_en: Quiet.
neg_utt_de: Zeit zu schlafen. (Time to sleep.)

The parallel utterances are directly from the MASSIVE dataset. Negative utterances, on the other hand, are sampled from negative utterance pools. We built a negative utterance pool for each language from the whole training set. For each utterance in the training set, we calculated BLEU scores between it and all other utterances in the same language. We add an utterance into the negative pool if its scores are all smaller than 0.1, which guarantees that utterances in the pool are dissimilar from all other utterances in the training set except itself.

During training, the weights of XLM-R were frozen and only the layers in the FMDA were fine-tuned. The development set from MASSIVE was used to determine the best stop point of training. The other hyperparameters were similar with those used in Tiyajamorn et al., 2021.

To evaluate the output embeddings of our FMDA model, we performed two retrieval tasks as described in Section 4.3.1 and 4.3.2, and compared our results with (a) XLM-R's original $[CLS]$ embedding, and (b) the SOTA language-disentanglement model (Tiyajamorn et al., 2021) trained with our data.

4.3 Results

4.3.1 Cross-lingual Semantic Retrieval Task

We used MASSIVE's test split (containing 2974 sets of parallel utterances) to conduct cross-lingual semantic retrieval evaluation. Given one utterance (query) in the source language, we expect to locate its exact translation from the 2974 candidates in the target language. This was done by calculating cosine similarity between each query and all candidates in the embedding space as the ranking score. The retrieval performance was measure by accu-

Model	Embedding	EN-DE	DE-EN	EN-ES	ES-EN	EN-JA	JA-EN
XLM-R	cls	0.182	0.203	0.194	0.198	0.050	0.037
Tiyajamorn et al. (2021)	meaning	0.550	0.575	0.583	0.602	0.364	0.359
Our model	meaning	**0.594**	**0.605**	**0.645**	**0.650**	**0.400**	**0.380**
	Template(TP)	0.583	0.589	0.630	0.643	0.388	0.371
	Residual(RS)	0.283	0.268	0.369	0.321	0.070	0.075

Table 2: Results of cross-lingual semantic retrieval. **X-Y** in the column headers represents the language pairs for evaluation, where **X** is the source language and **Y** is the target language. The training language pair is the same as the corresponding evluation language pair for each column. The retrieval performance is measured by accuracy@1.

Model	Embedding	$\textbf{EN}_{EN\_DE}$	$\textbf{EN}_{EN\_ES}$	$\textbf{EN}_{EN\_JA}$	DE	ES	JA
XLM-R	cls	0.371	0.371	0.371	0.351	0.427	0.120
Tiyajamorn et al. (2021)	meaning	0.393	0.392	0.299	0.381	0.434	0.267
Our model	meaning	0.427	0.437	0.356	0.387	**0.441**	0.314
	TP	**0.427**	**0.439**	**0.441**	**0.396**	0.436	**0.340**
	RS	0.330	0.348	0.241	0.354	0.420	0.094

Table 3: Results of mono-lingual template retrieval. Column headers show the training and evaluation languages. For example, $\textbf{EN}_{EN\_DE}$ means the model is trained on EN-DE language pair and evaluated on EN. For language X other than EN, the training language pair is EN-X. The retrieval performance is measured by accuracy@1.

racy@1, i.e. the fraction that the top-1 retrieval matches the target.

Table 2 shows the cross-lingual retrieval result of different models and embeddings. The first row shows the performance of XLM-R's original $[CLS]$ embedding, and the second row shows the performance of the language-agnostic meaning embedding by training the network from Tiyajamorn et al., 2021. The meaning embedding from our FMDA model constantly outperforms both baselines. Diving deeper, we notice that the meaning embedding from Tiyajamorn et al., 2021 may retrieve an utterance with related semantic but of different template and slots; whereas our meaning embedding, which is reconstructed using the fine-grained components, is able to capture the exact translation (as demonstrated by the case in Table 7 of Appendix). The ablation study in Section 4.4 also proves the importance of the fine-grained reconstruction for cross-lingual retrieval.

Comparing the three embedding representations of our model, we find the meaning embedding outperforms the template embedding (TP) as expected, since the former contains more semantic information than the latter (as shown by the case in Table 8 in Appendix). Residual-semantic embedding (RS) in the bottom row has the worst performance because it encodes the least semantic information.

We also notice the differences between language pairs when comparing the columns in Table 2. All embeddings perform much worse on EN-JA than EN-DE/ES, because Japanese belongs to a language family distant from the others. We will further discuss this in Appendix C.

4.3.2 Mono-lingual Template Retrieval Task

To validate the capacity of our model for extracting the carrier phrase/template information from an utterance, we further carried out the mono-lingual template retrieval as the second evaluation task. The evaluation pairs were generated from MASSIVE's test split by manually replacing the slot value of utterances, such that the source and target utterances are from the same language, share the same template, but differs in their slot values (for utterances without a slot labelled, we just discarded them). This pre-processing resulted in about 1.9k evaluation pairs for each language. The following is one of the evaluation pair from EN:

source_utt: I like Senatra songs.

target_utt: I like Taylor Swift songs.

Similar with the cross-lingual semantic retrieval, given a source utterance we expect to find the target utterance from the pool. The performance of the retrieval measured by accuracy@1 is shown in Table 3. As can be seen, the template embedding from our model have consistent better performance than the embedding from the baseline model by Tiyajamorn et al. (2021). Besides, our meaning embedding has the similar performance compared to the template embedding, which means that the meaning embedding is able to contain most of the information from the template embedding.

To make the conclusion of the template retrieval experiment more solid, we prepared an alternative evaluation set, where multiple target utterances were generated from one source, all sharing the same template. Then the retrieval performance measured by mean average precision (MAP) is shown in Table 6 in the Appendix, which matches the observation from Table 3.

4.4 Ablation Study

To understand the importance of each set of losses in the FMDA model, we conducted an ablation study for the cross-lingual retrieval as shown in Table 4. Removing the contrastive loss leads to a significant drop on the retrieval performance, since such loss between a pair of parallel utterances is essential to build up the language alignment.

In addition, among the three sets of reconstruction losses, we find the utterance reconstruction loss $Loss_{rec}^{utt}$ brings significant benefit, while the residual-semantic reconstruction loss $Loss_{rec}^{rs}$ has little function. It needs to noted that when removing all three reconstruction losses, the performance is worse than the original FMDA model, but better than removing $Loss_{rec}^{utt}$ only. This is because in the latter setting, the model with partial reconstruction may lead to a sub-optimal by learning partial information of the utterance. This further proves the effectiveness of the interaction among each reconstruction loss.

Model	EN-DE	DE-EN
XLM-R	0.182	0.203
Tiyajamorn et al. (2021)	0.550	0.575
Our model	**0.594**	0.605
w/o all reconstruction losses	0.580	0.603
w/o utterance reconstruction loss	0.569	0.586
w/o template reconstruction loss	0.583	0.600
w/o residual reconstruction loss	0.584	**0.610**
w/o contrastive loss	0.241	0.234

Table 4: The performance of models with different training settings on cross-lingual semantic retrieval tasks (measured by accuracy@1).

4.5 Visualization

The fine-grained embeddings from our FMDA model are visualized using t-SNE plotting as shown in Figure 4. Figure 4a and 4b show the language embeddings and meaning embeddings of 800 EN-DE utterance pairs, respectively. These embeddings are generated from the FMDA model described in Section 3. Clearly, the language embeddings shows separated language clusters. While the meaning embeddings shows the translation alignment between two languages. In addition, Figure 4b contains multiple clusters, which correspond to different domains/intents in the corpus.

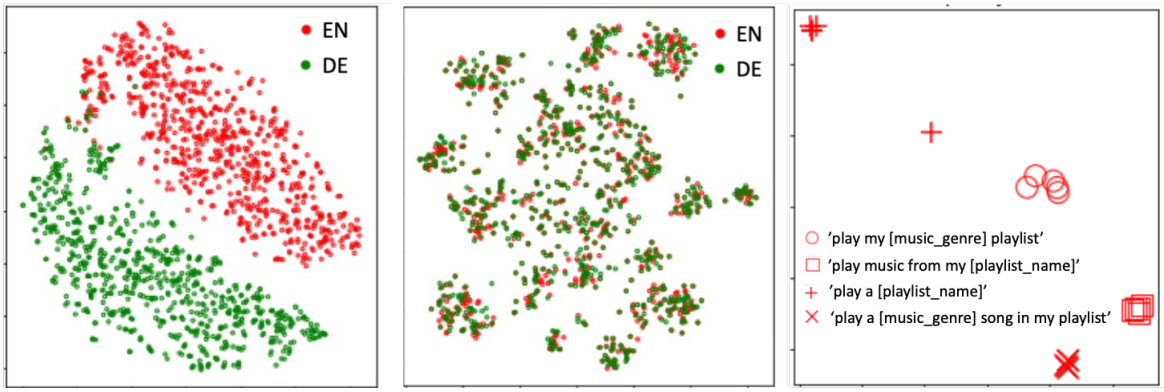
Figure 4c visualizes the template embeddings (generated from the two-stage FMDA model as described in Section 4.6) of 20 English utterances. All of them are from *play-music* intent, but of 4 different templates. The plot shows clearly that our template embedding is efficient in extracting the template information from different sentences.

4.6 Further Exploration with the Two-Stage FMDA

Former experiments demonstrate the effectiveness of fine-grained decomposition and reconstruction of embedding representations using our proposed FMDA model, and its benefit for different applications. However, the training of the FDMA involves multiple different loss functions, which may affect the optimization of each component. Therefore, we would like to investigate if training different components of FDMA in separating steps can lead to better embedding representations.

Here we conducted a two-stage training procedure to obtain better template representation. For the first stage, we focused on template encoder in FMDA and template related loss terms, i.e. the template reconstruction loss and the template contrastive loss. Mono-lingual template pairs data, in the same format as that described in Section 4.3.2, were built as positive pairs for the training. Therefore, the template encoder in FMDA can be better learned on this pure template data. In the second stage, the template encoder in the FMDA was frozen and all other losses except template-related ones were used together on the dual-lingual pair training data described in section 4.

Results of the model trained in two-stage setting are shown in Tabel 5. For mono-lingual template retrieval tasks (columns **EN** and **DE**), the template embedding (TP) obtained through two-stage training is far better than that from the original FMDA model. The meaning embedding also benefits from the boost of TP. For cross-lingual semantic retrieval tasks (columns **EN-DE** and **DE-EN**), although the performance of two-stage TP is low (since the template encoder hasn't been trained with dual-lingual pair data in the two-stage setting), the performance



(a) Language embedding (b) Meaning embedding (c) Template embedding

Figure 4: Visualisation of fine-grained embeddings from our FMDA model.

of the meaning embedding is similar with that from the original FMDA model.

This experiment demonstrates that the two-stage based FMDA is able to learn a much better template embedding while the meaning embedding still effectively encodes the whole semantic.

Method	Embedding	EN-DE	DE-EN	EN	DE
All-together	meaning	0.594	**0.605**	0.427	0.387
	TP	0.583	0.589	0.427	0.396
Two-stage	meaning	**0.595**	0.601	0.584	0.566
	TP	0.241	0.251	**0.811**	**0.802**

Table 5: Comparison of all-together training and two-stage training. The numbers of all-together training are from Table 2 and Table 3.

5 Conclusion

In this paper, we introduced FMDA, a lightweight encoding architecture that is able to disentangle fine-grained semantic information from language-specific information in a multilingual setting. Compared with previous works, the FMDA distils 1) language embedding emb_{lang} to encode the language-specific information, 2) template embedding emb_{tp} to encode the the backbone template of the sentence, and 3) the residual embedding emb_{rs} to encode the residual information such as slot. Such fine-grained representations allow retrieval applications at different levels under the NLU setting.

Two retrieval tasks conducted on the MASSIVE dataset demonstrate that FMDA's meaning embedding achieves the best performance on the cross-lingual semantic retrieval task and FMDA's template embedding achieves the best performance on the mono-lingual template retrieval task. Both constantly outperform the SOTA language-disentanglement baseline across multiple languages.

References

Aman Ahuja, Nikhil Rao, Sumeet Katariya, Karthik Subbian, and Chandan K Reddy. 2020. Language-agnostic representation learning for product search on e-commerce platforms. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 7–15.

Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. Xor qa: Cross-lingual open-retrieval question answering. *arXiv preprint arXiv:2010.11856*.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. A multi-task approach for disentangling syntax and semantics in sentence representations. *arXiv preprint arXiv:1904.01173*.

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning cross-lingual sentence representations via a multi-task dual-encoder model. *arXiv preprint arXiv:1810.12836*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Yifan Gao, Qingyu Yin, Zheng Li, Rui Meng, Tong Zhao, Bing Yin, Irwin King, and Michael Lyu. 2022. Retrieval-augmented multilingual keyphrase generation with retriever-generator iterative training. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1233–1246. Association for Computational Linguistics.

Mihir Kale and Abhinav Rastogi. 2020. Template guided text generation for task-oriented dialogue. *arXiv preprint arXiv:2004.15006*.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. *arXiv preprint arXiv:2106.06937*.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846. Association for Computational Linguistics.

Pragaash Ponnusamy, Alireza Roshan Ghias, Chenlei Guo, and Ruhi Sarikaya. 2020. Feedback-based self-learning in large-scale conversational ai agents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13180–13187.

Pragaash Ponnusamy, Clint Solomon Mathialagan, Gustavo Aguilar, Chengyuan Ma, and Chenlei Guo. 2022. Self-aware feedback-based self-learning in large-scale conversational AI. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 324–333.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. Lareqa: Language-agnostic answer retrieval from a multilingual pool. *arXiv preprint arXiv:2004.05484*.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages

7764–7774, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Wieting, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. A bilingual generative transformer for semantic sentence embedding. *arXiv preprint arXiv:1911.03895*.

John Wieting, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A bilingual generative transformer for semantic sentence embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1581–1594, Online. Association for Computational Linguistics.

Xinnuo Xu, Guoyin Wang, Young-Bum Kim, and Sungjin Lee. 2021. Augnlg: Few-shot natural language generation using self-trained data augmentation. *arXiv preprint arXiv:2106.05589*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

A Multi-target Template Retrieval

In Section 4.3.2, we introduced the mono-lingual template retrieval experiment, which was evaluated using source-target utterance pairs generated by slot replacement. In order to make the conclusion of the experiment more solid, here we prepared an alternative evaluation set, where five target utterances were generated from each source, all of them sharing the same template but with different slot values. Then, mean average precision (MAP) was used to evaluate whether all of the ground-truth targets can be retrieved at high rank. The result is shown in Table 6, which is consistent with the observations from Table 3.

B Case Study

For better understanding of the cross-lingual semantic retrieval results (Section 4.3.1), we pulled out some examples from the EN-DE retrieval experiment to demonstrate the different behaviors of different embeddings.

Table 7 shows a case where the result from our model is different from that of the baseline model (Tiyajamorn et al. (2021)). The meaning embedding from our FMDA model is able to retrieve the correct target, which is the exact translation of the source query. Whereas the embedding from the baseline model retrieves a wrong answer, which has the same intent as the source but differs in template and slot. This is because the semantic representation from our FMDA model is reconstructed from finer grains (template and residual-semantic/slots) and is able to capture detailed information in the sentence more accurately.

Table 8 shows an example where the meaning embedding from our FMDA model captures the correct target while the template embedding from the same model retrieves a wrong one – though its template is same as the query, the slot doesn't match. This is as expected, since the residual-semantic information (slot) is decoupled from the template embedding.

In conclusion, our template embeddings are able to capture necessary template information an utterance, while our meaning embeddings are able to cover the whole semantic of the utterance.

C Extend FMDA to a Unified Multilingual Model

In the former experiments, the models were all trained on dual-lingual pairs, e.g. EN-DE, EN-ES. To further validate if the model can benefit from training multiple languages together, a multi-lingual training experiment is conducted and analyzed. The model architecture remains the same, while the input is not only a dual-lingual pair, but multiple dual-lingual pairs together, i.e., the model is trained on the mixture of EN-DE, EN-ES and EN-JA pairs in one epoch.

Table 9 demonstrates the results of the multi-language pairs training. First, for languages like English (EN), German (DE), and Spanish (ES), the performance of the multi-lingual training model is actually worse than the dual-lingual training model. However, the multi-lingual model performs better on Japanese (JA). Second, for all languages, the multi-lingual training model still outperforms the baseline model from Tiyajamorn et al. (2021).

We argue that the reason of this results is: languages from similar families like EN, DE, and ES have been well learned in the original XLM-R and may have more in common. However, language like JA is a single-family language which does not share common scripts nor in the same genre with others, and is not well-studied in the original XLM-R. Therefore, the performance of the EN, DE, and ES part in the model trained under the multi-lingual setting is affected by JA so that to be worse , while JA, which is insufficient leanred in the XLM-R, can benefit more from other well-learned languages.

Model	Embedding	EN_{EN_DE}	EN_{EN_ES}	EN_{EN_JA}	DE	ES	JA
XLM-R	cls	0.370	0.370	0.370	0.355	0.379	0.117
Tiyajamorn et al. (2021)	meaning	0.397	0.386	0.310	0.386	0.380	0.283
Our model	meaning	0.421	0.432	0.360	**0.409**	0.423	0.350
	TP	**0.421**	**0.434**	**0.395**	0.405	**0.425**	**0.375**
	RS	0.328	0.344	0.234	0.360	0.379	0.088

Table 6: Results of mono-lingual template retrieval under multi-target retrieval setup. The retrieval performance is measured by mean average precision (MAP). Other settings are the same as Table 3.

Query	wake me up at five am this week
Top-1 retrieval	Our model (meaning): wecke mich in dieser woche um fünf uhr auf (wake me up at five am this week)
	Tiyajamorn et al. (2021) model (meaning): ich muss morgen um zehn uhr aufstehen (i need to get up at ten tomorrow)

Table 7: An example from the EN-DE cross-lingual retrieval experiment, for which our meaning embedding retrieved the correct target, whereas the embedding from Tiyajamorn et al. (2021) retrieved a wrong answer.

Query	what's the time in sweden
Top-1 retrieval	Our model (meaning): wie spät ist es in schweden (what's the time in sweden)
	Our model (template): welche uhrzeit ist es in einer stadt (what time is it in a city)

Table 8: An example from the EN-DE cross-lingual retrieval experiment, for which our meaning embedding retrieved the correct target, whereas our template embedding retrieved a wrong answer.

Eval. Task	Tiyajamorn et al. (2021)	Dual-lingual Training	Multi-lingual Training
DE-EN	0.575	**0.605**	0.572
ES-EN	0.602	**0.650**	0.617
JA-EN*	0.359	0.380	**0.397**
EN	0.361	**0.436**	0.429
DE	0.381	**0.396**	0.395
ES	0.434	**0.436**	0.433
JA*	0.267	0.340	**0.341**

Table 9: Comparison of dual-lingual training with multi-lingual training. Results shown are the performance of cross-lingual semantic retrieval (top rows) and mono-lingual template retrieval (bottom rows) respectively, measured by accuracy@1. For dual-lingual training, the model is trained on one language pair and evaluated on the corresponding language (pair), as described in Table 2 and Table 3. For multi-lingual training, a unified model is trained using data from multiple language pairs. Specifically, JA data is not included in the training set except for the rows marked by *.