Assisting Language Learners: Automated Trans-Lingual Definition Generation via Contrastive Prompt Learning

Hengyuan Zhang[1], Dawei Li[2], Yanran Li[3†], Chenming Shang[1], Chufan Shi[1], Yong Jiang[1]
[1]Shenzhen International Graduate School, Tsinghua University
[2]Halicioğlu Data Science Institute, University of California, San Diego
[3]Independent Researcher
zhang-hy22@mails.tsinghua.edu.cn

Abstract

The standard definition generation task requires to automatically produce mono-lingual definitions (e.g., English definitions for English words), but ignores that the generated definitions may also consist of unfamiliar words for language learners. In this work, we propose a novel task of **T**rans-**L**ingual **D**efinition **G**eneration (TLDG), which aims to generate definitions in another language, i.e., the native speaker's language. Initially, we explore the unsupervised manner of this task and build up a simple implementation of fine-tuning the multilingual machine translation model. Then, we develop two novel methods, Prompt Combination and Contrastive Prompt Learning, for further enhancing the quality of the generation. Our methods are evaluated against the baseline Pipeline method in both rich- and low-resource settings, and we empirically establish its superiority in generating higher-quality trans-lingual definitions. The ablation studies and further analysis are also conducted to provide more hints on this new task.

1 Introduction

A significant area of research within Intelligent Computer-Assisted Language Learning (ICALL) is devoted to supporting language learners in understanding words (Enayati and Gilakjani, 2020; Lolita et al., 2020). This research is primarily motivated by two main issues: (1) Language learners often struggle to accurately identify the meaning of words with multiple definitions, as the cognitive process of differentiating each meaning can be challenging (Tyler and Evans, 2001); (2) On another note, lexicographers are responsible for manually updating predefined word-definition inventories for language learners, a process that may be time-consuming and not always able to keep up with the constantly evolving nature of language usage. To address these issues, researchers aim to benefit both language learners and lexicographers by automatically generating the definition for a given word based on its corresponding local context (Ni and Wang, 2017; Gadetsky et al., 2018; Ishiwatari et al., 2019; Bevilacqua et al., 2020).

Figure 1: The application scenes of a Chinese native speaker learning English and English native speaker learning Chinese. We also build a Chrome extension (in Appendix A) to better show the application scenes.

Previous works on definition generation mainly focus on mono-lingual generation scenarios, primarily due to the availability of parallel training and evaluation data (Yang et al., 2020; Huang et al., 2021; Zhang et al., 2022a). However, these works rarely notice a real-occurring problem that the generated definitions may also consist of unfamiliar words for language learners (Zhang, 2011). In other words, it is more applicable to generate definitions in the native language of foreign learners. As depicted in Figure 1, if a Chinese native speaker wants to know an English word's meaning, the definition in Chinese is easier to capture.

To this end, we propose a novel task called **T**rans-**L**ingual **D**efinition **G**eneration (**TLDG**). The TLDG task is challenging because there are no trans-lingual parallel datasets, e.g., the word and context are in Chinese, and the definition is

† Corresponding author

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 260–274

Context	Generated Definition	Error Type
This food revitalized the patient.	食物使病人恢复活力。 (*Food revitalizes patients*)	Ignore-task error
..., 各家各派对人性的看法极为不同。 (*...,Each party has a very different view of human nature.*)	形容 (*Describe*) a person's opinion about something.	Language-mix error

Table 1: Zh-En and En-Zh examples of the two error types in the unsupervised TLDG task. The target words are marked with underline in context.

in English. Also, building trans-lingual parallel datasets is labor-consuming. To address this, we leverage the data resources of mono-lingual definition generation and utilize translation model to explore the trans-lingual definition generation task in an unsupervised manner. During preliminary experiments, we find two typical types of errors in the generated results. As shown in Table 1[1], *Ignore-task error* means the model only translates the input's context but neglects the definition generation task. *Language-mix error* means words in different languages simultaneously appear in the generated definition.

To mitigate the problems, we develop two novel learning methods. For the Ignore-task error, we get inspired from task-oriented prompt learning (Chung et al., 2022; Akyürek et al., 2022), and design Prompt Combination method to force the models focus on generating trans-lingual definition rather than mere translation. In addition, we propose Contrastive Prompt Learning method based on an contrastive loss (Hadsell et al., 2006; Schroff et al., 2015), which separates language information from the task prompt and in turn acquires a better task prompt representation for definition generation. Due to the scarcity of definition generation data in numerous languages, we carry out extensive experiments in both rich- and low-resource situations. We demonstrate that the Contrastive Prompt Learning method is effective in addressing the two errors and capable of yielding higher-quality definitions when compared to the baselines in both scenarios.

In general, our contributions are as follows:

- To better assist language learners, we propose the task of TLDG in an unsupervised manner and identify two typical errors.
- We develop several methods to mitigate the problems and demonstrate the Contrastive Prompt Learning method yields promising performances in both rich- and low-resource scenarios.
- We analyze the methods through ablated and case studies, and provide several hints on this newly introduced task. Also, we build a Chrome extension (in Appendix A) to further show the application scene of our proposed task.

[1]In this paper, Zh-En means the input's word and context are in Chinese, and the expected generated definition is in English. Other language combinations are also similar.

2 Related Work

2.1 Definition Generation

The task of definition generation is first proposed by Noraset et al. (2017), which aims to generate definitions from corresponding word embeddings. Subsequent studies have investigated a broader range of application scenarios and model architectures for generating definitions. To generate appropriate definitions for polysemies, Ni and Wang (2017) first introduce the context and input the context with the target word to a bi-encoder model. Following them, Ishiwatari et al. (2019) develop a method that incorporates a gate mechanism in the decoding stage to integrate the information of the word and context. There are also some works that try to model the semantic representation in a more detailed way. Specifically, they break down the meaning of the target word into several components and provide a fine-grained word representation for the generation stage (Li et al., 2020; Reid et al., 2020a).

Recently, some works adopt pre-trained encoder-decoder models in definition generation and achieved great success. Huang et al. (2021) use a re-ranking strategy to obtain proper specific definitions. Zhang et al. (2022a) regard word and definition as a semantic equivalence pair to do contrastive learning. However, all the aforementioned works focus on improving the quality of the generated definitions, and the difficulty of understanding the definition itself for language learners has been ignored.

Although Kong et al. (2022a) design a multi-task framework to generate definitions with more simple words, we argue that other factors like language grammar will still hinder language learners to understand the definition. To mitigate it, we propose a novel task of trans-lingual definition generation

to generate definitions in the target language.

2.2 Prompt Learning

In recent years, numerous pre-trained models have been introduced, e.g., GPT (Radford et al., 2018), BART (Lewis et al., 2019). To adapt these models for different downstream tasks, prompt learning has been widely used. Schick and Schütze (2020a) manually design discrete template prompts to transform the downstream task into the text-infilling task, which is closer to the pre-trained paradigm. Besides, in the conditional text generation field, both Zhang et al. (2022b) and Xie et al. (2022) regard attribute keywords as hard prompts and fuse them into the model to control the generation result. However, Manually designing hard prompts can be both tedious and challenging, later works suggest using the soft prompts that consist of multiple learnable embeddings for the downstream tasks (Li and Liang, 2021; Liu et al., 2021; Han et al., 2022).

Furthermore, some works propose that rather than updating the entire PLM, it is more effective to fix its parameters and only update the soft prompts (Lester et al., 2021; Qin and Eisner, 2021a). When using large PLMs as the backbone, this method can achieve comparable results to fine-tuning the entire model. In the low-resource scenario, Gu et al. (2021) apply prompt initialization and use several tasks to obtain generalized prompts for different downstream tasks. Zheng and Huang (2021) and Zhang et al. (2021) use the prompt learning strategy to get different task-oriented prompts with corresponding task-specific objectives and achieve satisfactory results.

In this work, we use prompt learning to indicate the task and address the Ignore-task error. By developing a novel contrastive prompt learning loss, we finally achieve promising performances on both rich- and low-resource TLDG.

3 Method

One straightforward approach to generating trans-lingual definitions is to develop a pipeline that initially produces mono-lingual definitions and then translates them into the desired language. This intuitive approach serves as one naive baseline, which we elaborate in the experiment section (Section 4.2).

Besides, in this section, we introduce 3 methods to better fit our task: (1) a simple implementation of fine-tuning on multi-lingual translation model; (2) Prompt Combination method; and (3) Contrastive Prompt Learning method.

3.1 Task Formulation

The standard definition generation (DG) task is to generate the definition $D = \{d_0, ...d_t\}$ for a given word or phrase $W = \{w_i, ..., w_j\}$ and its corresponding context $C = \{w_0, ..., w_k\}(0 < i < j < k)$. Here, the context is a sentence containing the word. Note that standard DG is a mono-lingual task where the word, context, and definition are in the same language.

Distinguishedly, the task of trans-lingual definition generation (TLDG) is to generate trans-lingual definition D_{l_j} in language l_j for a given word W_{l_i} and context C_{l_i} in another language l_i. Since there does not exist TLDG example triplets $\{(W_{l_i}, C_{l_i}, D_{l_j})\}$, the only available resources are mono-lingual definition generation datasets. Hence, the TLDG task in this work can be regarded as a fully unsupervised task.

3.2 Simple Implementation of Directly Fine-tuning Translation Model

The newly introduced TLDG task aims to generate the trans-lingual definition without supervised parallel datasets. As neural machine translation (NMT) shows powerful performance in translation, as a preliminary attempt, we directly fine-tune multi-lingual NMT with existing mono-lingual DG datasets (G). Concretely, we concatenate language prompt (which is predefined in the multi-lingual NMT model to specify the source and target languages), target word, and context ($[L_{l_i}; W_{l_i}; C_{l_i}]$) as input X_{l_i} to the encoder. Similarly, we concatenate language prompt and definition ($[L_{l_i}; D_{l_i}]$) as ground-truth Y_{l_i} to train the model, which can be formulated as:

$$P(Y_{l_i}|X_{l_i}) = \prod_t p(y_t|y_{<t}, X_{l_i}; \theta) \quad (1)$$

where y_t is the t-th token of Y_{l_i}, θ is the model's parameters to be tuned. To optimize, a cross-entropy loss is utilized to assess the difference between the distribution generated by the model and the ground-truth distribution, and the loss function is as follows:

$$\mathcal{L}_{MLE} = - \sum_{\substack{(W_{l_i}, C_{l_i}, D_{l_i}) \in G_{l_i}, \\ l_i \in L}} logP(Y_{l_i}|X_{l_i}; \theta) \quad (2)$$

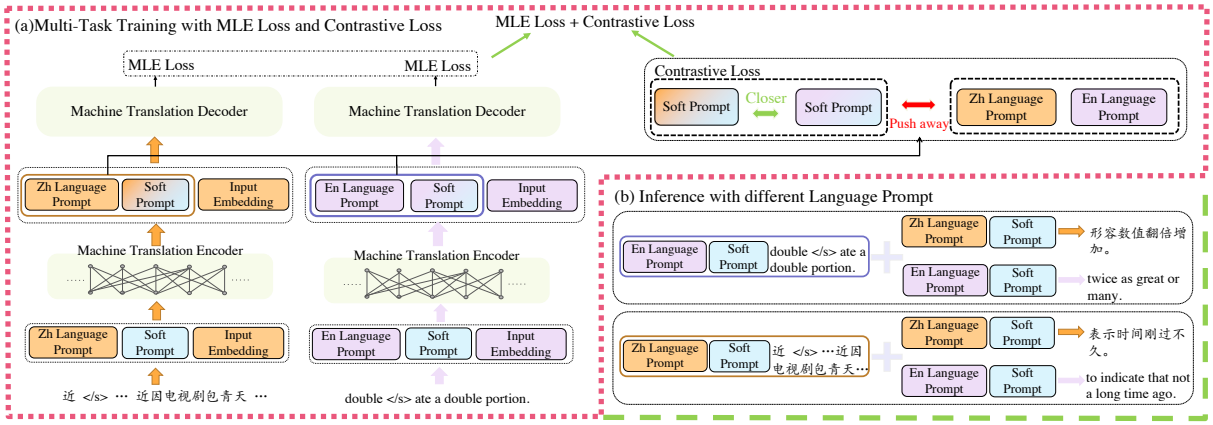

Figure 2: The area surrounded by the red dotted line represents the training process and the green dash line represents the inference process. In the training phase, (1) the task prompt will mix the language information from the language prompt (in blended color), and (2) the contrastive loss (upper right corner) is together applied with the MLE loss (upper left corner) to train the model jointly. At the inference stage, the language prompt could be set to any other languages used in the training stage for trans-lingual definition generation. Best viewed in color.

By concatenating the corresponding language prompt, the model is able to infer trans-lingual definitions in any language previously seen in the training stage ($< W_{l_i}, C_{l_i} > \rightarrow D_{l_j}, l_i, l_j \in L$).

3.3 Prompt Combination

Despite that fine-tuning the translation model seems plausible for trans-lingual definition generation, we find a plethora of Ignore-task cases in the generated definitions. We conjecture that the language prompt would still instinctively induce the translation model to perform the translation task, and thus leading to those Ignore-task errors.

To notify the model focus on the definition generation task, we add a specific task-oriented prompt after the language prompt. We adopt soft prompts for our task since they have been shown more flexible than hard prompts (Liu et al., 2022). In the training stage, we insert the task prompt $T = \{t_1, t_2, ..., t_n\}$ after the language prompt L_{l_i} for both encoder and decoder inputs, where n is the number of soft prompt tokens.

While this strategy mitigates the Ignore-task error in trans-lingual definition generation, we find adding the task-oriented soft prompt will lead to Language-mix errors. One possible explanation is that during the training stage, the task prompt is mixed up with the language information from the language prompt. During inference, such mixed task prompts will confuse the model to generate words in undesired languages.

3.4 Contrastive Prompt Learning

To tackle this problem, we propose a Contrastive Prompt Learning method. This method aims to obtain a more informative and representative task prompt by decoupling the language information inside within it. The overview of the proposed method is illustrated in Figure 2, where we take Chinese and English as examples.

In each batch, we randomly fetch training samples in two different languages (l_i and l_j) and separate them into two groups. After passing each group into the model, we extract the language prompt embedding $\mathbf{H}^{lp}_{l_i}$ and the task prompt embedding $\mathbf{H}^{tp}_{l_i}$ from each group's encoding $\mathbf{H}_{l_i}$ and $\mathbf{H}_{l_j}$ according to their positions:

$$\mathbf{H}^{tp}_{l_i}, \mathbf{H}^{lp}_{l_i} = \text{Extract}(\mathbf{H}_{l_i}) \quad (3)$$

$$\mathbf{H}^{tp}_{l_j}, \mathbf{H}^{lp}_{l_j} = \text{Extract}(\mathbf{H}_{l_j}) \quad (4)$$

Since the language prompt only has one token, we directly regard language prompt embedding as language prompt representation $\mathbf{h}^{lp}_{l_i}$. For multiple task prompt tokens, we apply the pooling function to $\mathbf{H}^{tp}_{l_i}$ and $\mathbf{H}^{tp}_{l_j}$ to get the task prompt representation $\mathbf{h}^{tp}_{l_i}$ and $\mathbf{h}^{tp}_{l_j}$. Without loss of generality, we implement attention-pooling, mean-pooling and max-pooling, and compare them in Section 4.5.

To build up contrastive loss, we regard task prompt representation in different languages as positive pairs ($\mathbf{h}^{tp}_{l_i}$, $\mathbf{h}^{tp}_{l_j}$), task prompt representation and different language prompt representation as negative pairs $\{(\mathbf{h}^{tp}_{l_i}, \mathbf{h}^{lp}_{l_j}),\ l_i, l_j \in L\}$. By doing

so, the language information in $\mathbf{h}_{l_i}^{tp}$ and $\mathbf{h}_{l_j}^{tp}$ can be effectively eliminated. Mathematically, the contrastive loss is formulated as:

$$\mathcal{L}_C = \max(d_p - d_n + \sigma, 0)/\tau \tag{5}$$

$$\begin{aligned} d_p &= \|\mathbf{h}_{l_i}^{tp} - \mathbf{h}_{l_j}^{tp}\| \\ d_n &= \sum_{a\in\{i,j\}} \frac{1}{2}\|\mathbf{h}_{l_i}^{tp} - \mathbf{h}_{l_a}^{lp}\| \end{aligned} \tag{6}$$

where d_p is the distance of positive pair, d_n is the average distance of negative pairs, σ is the margin and τ is the temperature to scale the contrastive loss.

As Figure 2 depicts, the proposed contrastive loss is combined with MLE loss to train the model:

$$\mathcal{L}_{Final} = \lambda * \mathcal{L}_{\mathcal{C}} + (1 - \lambda) * \mathcal{L}_{MLE} \tag{7}$$

where λ is a hyper-parameter to balance the two losses. In this way, our method is able to (1) separate the language information from the task prompt based on the novel contrastive loss, and (2) obtain a more oriented and pure task prompt representation for generating trans-lingual definition.

4 Experiments

In this section, we conduct extensive experiments and analyze the proposed methods carefully.

4.1 Datasets

Considering that many languages do not have sufficient definition generation data, we validate the proposed method in both rich- and low-resource scenarios. Note that all the datasets we use to train models are the mono-lingual definition generation datasets, which means the source language and target language are the same.

Rich-resource In the rich-resource scenario, we train and evaluate our models using English and Chinese definition generation datasets. For English, we use the Oxford dataset, collected using Oxford APIs of Oxford Dictionary[2] by Gadetsky et al. (2018). We follow Ishiwatari et al. (2019) to split them into training, validation, and test sets.

For Chinese, we follow Kong et al. (2022b) to use Chinese-WordNet (CWN) (Huang et al., 2010) and split them into training, validation, and test sets. It is a semantic lexicon aiming to provide a knowledge base of sense[3]. The statistics of these two datasets are shown in Appendix B. In the inference stage, we conduct En-Zh, Zh-En trans-lingual definition generation.

Low-resource In the low-resource scenario, we set the training data size to 256, validation data size to 200, and following Schick and Schütze (2020b); Perez et al. (2021) to use the validation set as test set.

In specific, we build few-shot mono-lingual training datasets in English, Chinese, and France. For English and Chinese, we randomly choose samples from Oxford and CWN. For France, as there doesn't exist any public France definition generation dataset, we follow Reid et al. (2020b) to collect data from Lerobert Dictionary[4]. In the inference stage, we conduct trans-lingual definition generation with 6 settings, i.e., En-Zh, Zh-En, En-Fr, Fr-En, Zh-Fr, and Fr-Zh.

4.2 Experimental Settings

In this work, we utilize two multi-lingual NMT models, namely mBART-many-to-many[5] (a model that fine-tuned on mBART (Liu et al., 2020) with downstream machine translation tasks) and M2M[6] (a model that directly trained on massive multi-lingual translation tasks from scratch), to implement our ideas. For convenience, we use mBART-T to represent mBART-many-to-many in this paper.

For all experiments, we set the batch size to 16 and use Adam optimizer to update parameters. We train all of our models on a V100 GPU. Following Lester et al. (2021), we adopt 100 tunable soft prompt tokens. For the Contrastive Prompt Learning method, we set the temperature as 0.16 to scale the contrastive loss. The best performances in Section 4.4 adopt the attention-pooling function.

Compared Methods We compare with four methods: (1) A naive **Pipeline** method; (2) **Directly Fine-tuning** (Section 3.2); (3) **Prompt Combination** (Section 3.3); (4) **Contrastive Prompt** (Section 3.4). Specifically, the Pipeline method consists of generation and translation procedures. We begin with fine-tuning the pre-trained

[2] https://developer.oxforddictionaries.com

[3] https://lope.linguistics.ntu.edu.tw/cwn2/

[4] https://dictionnaire.lerobert.com

[5] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[6] https://huggingface.co/facebook/m2m100_418M

Model	Method	Semantic Sim	
		En-Zh	Zh-En
mBART + M2M	Pipeline	47.58	52.24
M2M	w/ Directly Fine-tuning	45.69	51.68
	w/ Prompt Combination	47.56	52.63
	w/ Contrastive Prompt Learning	**49.42**	**55.19**
mBART-T	w/ Directly Fine-tuning	43.32	51.16
	w/ Prompt Combination	45.13	51.85
	w/ Contrastive Prompt Learning	47.79	53.92

Table 2: Automatic evaluation results on the rich-resource test dataset. The best results are in **bold**.

mBART (Liu et al., 2020) model with mono-lingual datasets to generate mono-lingual definitions rather than trans-lingual definitions. Subsequently, we utilize the M2M model to translate the generated definitions into the target language.

Rich-resource In the rich-resource scenario, we fine-tune all the parameters (including soft prompt tokens) of the model with 10 epochs. We set the learning rate 5e-5 for M2M, and 1e-5 for mBART-T and mBART.

Low-resource In the low-resource scenario, we use the prompt-tuning strategy only to tune the soft prompt tokens as suggested by Li and Liang (2021); Qin and Eisner (2021b). Following (Gu et al., 2021), we set training epochs to 30 and learning rate to 1e-2 for all models.

4.3 Evaluation Metrics

Automatic Metrics To measure the semantic quality of generated trans-lingual definitions, we apply the sentence-transformer toolkit (Reimers and Gurevych, 2020) to calculate the semantic similarity between the generated definition in the target language and the golden reference in its original language (e.g., for En-Zh, we calculate semantic similarity between generated Chinese definition and the golden English definition).

Manual Evaluation We also perform manual evaluation on the test set of 200 examples in low-resource setting. Based on the automatic evaluation results from Table 4.4, we only manually assess M2M model with three methods (Directly Fine-tuning, Prompt Combination, Prompt Contrastive Learning) in rich-resource setting, and M2M model with Prompt Contrastive Learning method in low-resource setting.

We ask six college students who achieved a score above 580 in the College English Test 6 level (CET-6) as annotators. Three of these students will be responsible for annotating En-Zh results, while the remaining three will focus on Zh-En results. Similarly, we recruit six annotators who have passed Test national du français enseigné à titre de spécialité, niveau IV (TFS-4). Three of these annotators will be assigned to annotate En-Fr and Fr-En results, the remaining three will be responsible for Zh-Fr and Fr-Zh results.

Each annotator is asked to evaluate the generated trans-lingual definitions on two aspects: (1) Accuracy (Acc.) is a measure of the semantic relevance of the definitions to the word; (2) Fluency (Flu.) evaluates their readability without considering semantic alignment. Both criteria have a range of 1-5. In addition, the annotators are asked to rate the Ignore-task error and Language-mix error. We average the scores as the final score, and the agreements among the annotators of En-Zh, Zh-En, En-Fr & Fr-En, and Zh-Fr & Fr-Zh are ICC 0.937 ($p<0.001$), ICC 0.932 ($p<0.001$), ICC 0.904 ($p<0.001$) and 0.929 ($p<0.001$) respectively.

4.4 Main Results

We begin by examining the automatic evaluation results in rich-resource settings. As shown in Table 2, applying Contrastive Prompt Learning method on M2M and mBART-T models outperform other strategies across En-Zh and Zh-En scenarios. Furthermore, the baseline Pipeline method exhibits a performance degradation of 1.84 (En-Zh) and 2.95 (Zh-En) on the Semantic Sim metric when compared to our best method. This suggests that **the proposed Trans-lingual Definition Generation (TLDG) task cannot be simply addressed with a naive pipeline method**, which can be attributed to the errors accumulated during the pipeline.

Comparing the rows of M2M and mBART-T, M2M-based is superior on TLDG. We conjecture the superior performance comes from M2M's translation ability, which is empirically validated in Fan et al. (2021). Since M2M model is trained

with massive parallel translation data and equipped with the Language-Specific Sparse technique, it is shown more powerful than mBART-T on translation tasks. The comparison between M2M and mBART-T gives us a hint that **model's translation ability has an impact on our TLDG task**, which we analyze in later sections.

When checking the manual evaluation results in Table 3, it is notable that the proposed Contrastive Prompt Learning method obtains the highest scores on both Acc. and Flu. metrics. Comparing baseline Pipeline method with Contrastive Prompt Learning method in the Zh-En trans-lingual scenario (row 2 and row 8), we can see that Contrastive Prompt Learning method significantly improves trans-lingual quality, as it achieves 7.2% relative increase on Acc and 7.1% relative increase on Flu. A similar result in low-resource setting can refer to Appendix C.

Method	Language Combination	Acc. ↑	Flu. ↑
Pipeline (rich-resource)	En-Zh	3.09	3.34
	Zh-En	3.18	3.52
w/ Directly Fine-tuning (rich-resource)	En-Zh	3.02	3.37
	Zh-En	3.08	3.61
w/ Prompt Combination (rich-resource)	En-Zh	3.13	3.45
	Zh-En	3.17	3.67
w/ Contrastive Prompt (rich-resource)	En-Zh	**3.29**$_{(+6.4\%)}$	**3.51**$_{(+5.1\%)}$
	Zh-En	**3.41**$_{(+7.2\%)}$	**3.77**$_{(+7.1\%)}$
w/ Contrastive Prompt (low-resource)	En-Zh	2.98	3.31
	Zh-En	3.08	3.59
	En-Fr	3.04	3.48
	Zh-Fr	3.07	3.45
	Fr-En	3.11	3.62
	Fr-Zh	3.02	3.32

Table 3: Manual evaluation for quality assessment of trans-lingual definitions generated by M2M in low-resource test datasets

Another interesting finding comes when we compare the performances in rich- and low-resource scenarios. Take Zh-En trans-lingual task for example. It is observed that leveraging Contrastive Prompt Learning method in low-resource setting (row 10) is comparable to the simple implementation of directly fine-tuning (row 4) in rich-resource settings. Similar findings can also be found on the rows of En-Zh trans-lingual task. These findings greatly show the potential of the proposed method in the low-resource scenario. The results presented in Table 4 demonstrate that **our Contrastive Prompt Learning method effectively mitigates the two types of errors**. Specifically, when compared to directly fine-tuning implementation in the En-Zh scenario (row 1 and row 5), the Contrastive Prompt Learning method achieves a relative decrease of 77.8% in Language-mix error rate and perform well in Ignore-task error rate.

Method	Language Combination	Language-mix error rate↓	Ignore-task error rate↓
w/ Direct Fine-tuning (rich-resource)	En-Zh	-	11.25%
	Zh-En	-	9.50%
w/ Prompt Combination (rich-resource)	En-Zh	3.50%	7.50%$_{(-33.3\%)}$
	Zh-En	4.00%	6.00%$_{(-36.8\%)}$
w/ Contrastive Prompt (rich-resource)	En-Zh	-	2.50%$_{(-77.8\%)}$
	Zh-En	-	2.00%$_{(-78.9\%)}$

Table 4: Manual evaluation results of the two errors in trans-lingual definition generated by M2M in low-resource test datasets.

4.5 Ablation Study

Pooling Function To examine the variants of pooling functions as introduced in Section 3.4, we then conduct an ablation study on M2M model with the best task-ratio 0.2 obtained in Section 4.5.

As Table 5 shows, the attention-pooling function outperforms mean- and max- pooling functions on all the metrics. The reason lies in the distinctness of how these pooling functions gather token information. When constructing task prompt representation, the attention-pooling function aggregates all the task prompt tokens with the attention weight between the task and language prompt. Intuitively, the attention weight measures the degree of language information in each token of the task prompt. As a result, the task prompt representation based on attention-pooling contains more precise mixed language information, and in turn aids in separating language information when implementing Prompt Contrastive Learning. The variations observed in different pooling functions suggest that **the approach used to obtain an accurate representation is crucial in contrastive learning**.

Model & Method	Pooling Function	Semantic Sim	
		En-Zh	Zh-En
M2M /w Contrastive Prompt /w Task Ratio 0.2	attention	**49.42**	**55.19**
	mean	48.91	54.75
	max	48.83	54.68

Table 5: Ablation study results on the pooling functions. The best numbers are in **bold**.

Hyper-Parameter Another influential factor in our method is hyper-parameter λ in Eq. 7. To ex-

Model & Method	Task Ratio	Semantic Sim	
		En-Zh	Zh-En
M2M /w Contrastive Prompt /w Attention Pooling	0.1	48.81	55.12
	0.2	**49.42**	**55.19**
	0.3	48.24	54.76
	0.4	47.63	53.81
	0.5	47.87	53.79

Table 6: Hyper-parameter analysis results on the task ratio. The best results are in **bold**.

plore its effect, we keep using attention-pooling in all settings and set different λ for each model to observe the performance change.

As Table 6 shows, when the task ratio is set to 0.2, the proposed method yields the best performance. When the task ratio is lower or higher than 0.2, the performances deteriorate. We conjecture that our model requires more generation loss to guide contrastive learning in the right way.

4.6 Case Study

For better understanding, we present some cases under the rich-resource setting to vividly analyze the superiority of our Contrastive Prompt Learning method. Table 7 compares all methods on two trans-lingual scenarios. After examining the definitions produced by the directly fine-tuning implementation, we find undesired words like *"经济"* (*economy*) (in the En-Zh case), as well as the words *"interdependence"* and *"country"* (in the Zh-En case). All these words are the direct translations of the context words rather than the definitions. In the Zh-En case, it is clear that the definition from the Prompt Combination method contains Language-mix error, as it includes a Chinese word "形容" (*describe*). In the En-Zh case, the definition produced by the baseline Pipeline method includes an unsuitable explanation word "上升运动" (*upward movement*), which might be resulted from the limited definition style's data in the translation model's training corpus. In contrast, the Contrastive Prompt Learning method's output, which includes "正面发展" (*positive development*) and *"fewer or greater"*, accurately represents the meaning of the target words. Drawing on the highest scores in Table 2 and Table 3, we safely conclude that the proposed **Prompt Contrastive Learning is more effective in trans-lingual definition generation**.

We also conduct case studies on the choice of multi-lingual translation model, as a complementary assessment to the results in Table 2. As shown in Table 8, the generated definitions of mBART-T

Word	upturn
Context	... in response to the economic upturn helped by a recovery of key western export markets.
Pipeline	某人或某物的状况中的上升运动 (*The upward movement in the condition of someone or something.*)
Directly Fine-tuning	经济的好转。 (*The improvement of the economy.*)
Prompt Combination	形容 (*Describing*) a rising trend of something.
Contrastive Prompt	比喻特定事件向正面发展。 (*The specific event is developing towards a positive direction*)
Word	日益 (*day by day*)
Context	... 各国相互依赖程度日益加深。 (*... the degree of interdependence among countries is increasingly deepening.*)
Pipeline	the degree is deepening.
Directly Fine-tuning	increasing interdependence of country.
Prompt Combination	in a gradual and increasing degree.
Contrastive Prompt	to an ever greater or fewer degree.

Table 7: Generated result comparison between four methods on M2M model.

Word	accent
Context	... cobalt blue was used to accent certain elements including ...
M2M	强调特定对象。(*Emphasize specific objects.*)
mBART-T	强调的重点。(*Key points to emphasize.*)
Word	珍惜 (*cherish*)
Context	..., 什么又是值得你去珍惜的? (*..., what is worth cherishing for you?*)
M2M	deeply regard the value of something.
mBART-T	regard with great appreciation.

Table 8: Generated result comparison between M2M based and mBART-T based models.

contain "重点" (*key*) and *"appreciation"*, which are not accurate for explaining the corresponding words' meanings. However, the M2M model handles these cases well. This case study further demonstrates the hint that **the translation capability of the backbone model is crucial for trans-lingual definition generation**. For more cases in both rich- and low-resource scenarios, please kindly refer to Appendix D.

5 Conclusions

In this work, we propose a novel and challenging task TLDG that generates the trans-lingual definition in an unsupervised manner. To tackle the task, we leverage multi-lingual translation models and propose an effective method of Contrastive Prompt Learning for the task. Through extensive experiments, we validate the method is capable of addressing typical errors and promising in both

rich- and low-resource scenarios. In the future, we will develop more strategies to improve the quality of trans-lingual definitions.

Limitations

Our work has several limitations. In terms of method generalization, the proposed method depends on multi-lingual neural machine translation models to generate trans-lingual definitions, and hence limits its application scope to those languages rarely supported by translation models. Moreover, our findings are based on three languages, but different families of languages may exhibit distinct phenomenon that even challenges our conclusions.

References

Afra Feyza Akyürek, Sejin Paik, Muhammed Kocyigit, Seda Akbiyik, Serife Leman Runyun, and Derry Wijaya. 2022. On measuring social biases in prompt-based multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 551–564, Seattle, United States. Association for Computational Linguistics.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Fatemeh Enayati and Abbas Pourhosein Gilakjani. 2020. The impact of computer assisted language learning (call) on improving intermediate efl learners' vocabulary learning. *International Journal of Language Education*, 4(1):96–112.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(107):1–48.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2021. Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022. Ptr: Prompt tuning with rules for text classification. *AI Open*, 3:182–192.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese Information Processing*, 24(2):14–23.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. 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 3467–3476.

Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022a. Multitasking framework for unsupervised simple definition generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5934–5943.

Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022b. Multitasking framework for unsupervised simple definition generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5934–5943, Dublin, Ireland. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiahuan Li, Yu Bao, Shujian Huang, Xinyu Dai, and Jiajun Chen. 2020. Explicit semantic decomposition for definition generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 708–717.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yuri Lolita, Endry Boeriswati, and Ninuk Lustyantie. 2020. The impact of computer assisted language learning (call) use of english vocabulary enhancement. *Linguistic, English Education and Art (LEEA) Journal*, 4(1):206–221.

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in Neural Information Processing Systems*, 34:11054–11070.

Guanghui Qin and Jason Eisner. 2021a. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*.

Guanghui Qin and Jason Eisner. 2021b. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2020a. Vcdm: Leveraging variational bi-encoding and deep contextualized word representations for improved definition modeling. *arXiv preprint arXiv:2010.03124*.

Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2020b. Vcdm: Leveraging variational bi-encoding and deep contextualized word representations for improved definition modeling. *arXiv preprint arXiv:2010.03124*.

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

Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Timo Schick and Hinrich Schütze. 2020b. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Andrea Tyler and Vyvyan Evans. 2001. Reconsidering prepositional polysemy networks: The case of over. *Language*, pages 724–765.

Yuqiang Xie, Yue Hu, Yunpeng Li, Guanqun Bi, Luxi Xing, and Wei Peng. 2022. Psychology-guided controllable story generation. *arXiv preprint arXiv:2210.07493*.

Liner Yang, Cunliang Kong, Yun Chen, Yang Liu, Qinan Fan, and Erhong Yang. 2020. Incorporating sememes into chinese definition modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1669–1677.

Hengyuan Zhang, Dawei Li, Shiping Yang, and Yanran Li. 2022a. Fine-grained contrastive learning for definition generation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 1001–1012.

Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Differentiable prompt makes pre-trained language models better few-shot learners. *arXiv preprint arXiv:2108.13161*.

Yihua Zhang. 2011. Discussion on the definitions in chinese learner's dictionaries: Comparative study of domestic and foreign learner dictionaries (translated from chinese). *Chinese Teaching in the World*.

Zhexin Zhang, Jiaxin Wen, Jian Guan, and Minlie Huang. 2022b. Persona-guided planning for controlling the protagonist's persona in story generation. *arXiv preprint arXiv:2204.10703*.

Chujie Zheng and Minlie Huang. 2021. Exploring prompt-based few-shot learning for grounded dialog generation. *arXiv preprint arXiv:2109.06513*.

A Chrome Extension Application Scene

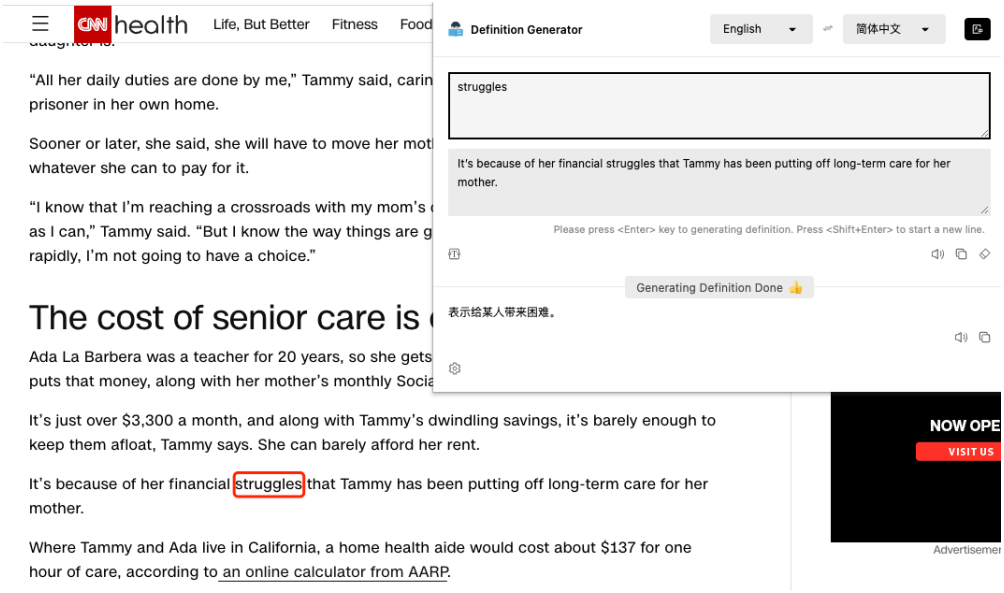

Figure 3: The application scene of Learning English words based on our best method. Given the word "struggles" and press the shortcut key, the application will identify its corresponding context and output the definition "表示给某人带来困难。" (To make someone difficult.).

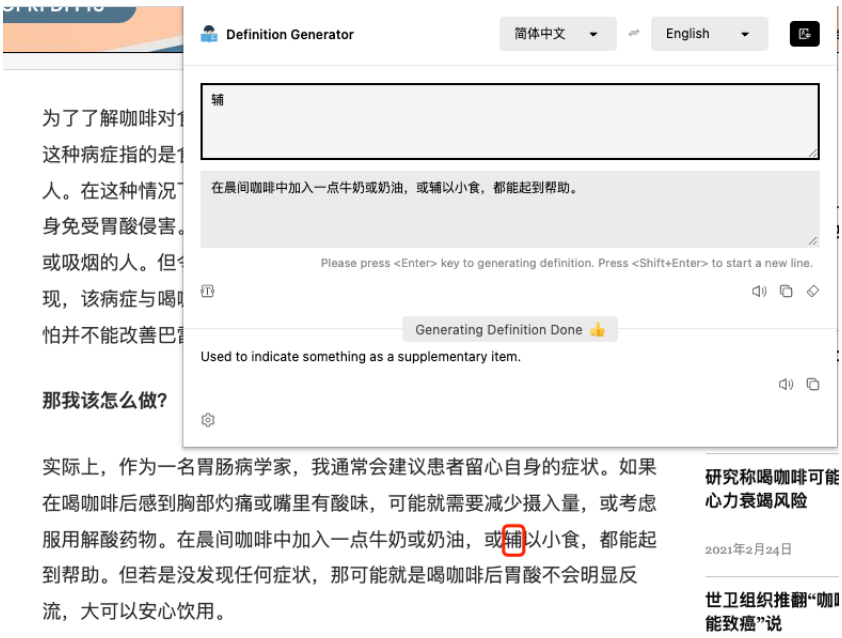

Figure 4: The application scene of Learning Chinese words based on our best method. Select the word "辅" (supplement) and press the shortcut key, the application will identify its corresponding context and output the definition "Describing something as an accessory or auxiliary item".

B Rich-resource Detailed Dataset Setting

	Oxford			CWN		
	Train	Valid	Test	Train	Valid	Test
Words	33128	8,867	3881	6574	823	823
Entries	97,855	12,232	5111	67861	8082	8599
Context length	17.74	17.80	16.24	34.49	34.73	34.04
Desc. length	11.02	10.99	10.03	14.76	14.60	14.72

Table 9: Statistics of the Oxford (English) dataset and CWN (Chinese) dataset.

We use Oxford and CWN definition generation datasets in rich-resource setting experiment, the statistics of Oxford and CWN are shown in Table 9.

C Human Evaluation of Pipeline method in Low-resource Setting

Language Combination	Method			
	/w Contrastive Prompt		Pipeline	
	Acc	Flu	Acc	Flu
En-Zh	2.98	3.31	2.73	3.09
Zh-En	3.08	3.59	2.91	3.34
En-Fr	3.04	3.48	2.88	3.31
Fr-En	3.11	3.62	2.98	3.46
Zh-Fr	3.07	3.45	2.92	3.27
Fr-Zh	3.02	3.32	2.74	3.13

Table 10: Human evaluation results of M2M /w Contrastive Prompt Learning method and baseline Pipeline method in low-resource setting.

We also compare our proposed M2M /w Contrastive Prompt Learning method with baseline Pipeline method in low-resource setting, the results are shown in Table 10.

D Generated Results

D.1 Rich-Resource Generated Results

Word	telex
Context	they telexed the company denying breach of contract.
Generated Result	以电传方式传送讯息。
Word	bulky
Context	radio could communicate between cities, but they were too bulky to be man-carried.
Generated Result	形容体积大的。
Word	concession
Context	a corona and one adverb of resignation - or is it concession?
Generated Result	承认或授权后述对象。
Word	electronic
Context	1987 was an early but fertile time for electronic dance music.
Generated Result	以电子方式进行演奏。
Word	spiral
Context	tensions have spiraled between pyongyang and the us.
Generated Result	比喻特定事件在一段长时间内持续进行。
Word	fortune
Context	I have had the good fortune to see the piece several times.
Generated Result	形容运气好。
Word	revitalize
Context	this food revitalized the patient.
Prompt Combination	使后述对象恢复生命力。
Word	意外
Context	好在我们都已买了保险，如果发生意外，一切都由保险公司理赔。
Generated Result	an unfortunate or unexpected occurrence of something.
Word	学术
Context	国立大学及所有私校没必要一窝蜂搞学术，现在学生所学和社会往往都是脱节的。
Generated Result	an academic activity of the university or community.
Word	立国
Context	立国精神、民族意识的观念如果不在军训课中提醒学生，根本没有机会。
Generated Result	the establishment of state.
Word	近
Context	包公庙..., 近因电视剧包青天脍炙人口而引来参拜人潮。
Generated Result	to indicate that not a long time ago.
Word	维
Context	怪手及人员到市场附近巡视，凡发现摊架，则一律予以铲除，以维公权力的威信。
Generated Result	maintain the state of (something).

Table 11: The generated results of M2M model with Contrastive Prompt Learning method under rich-resource setting.

D.2 Low-Resource Generated Results

Word	concession
Context	a corona and one adverb of resignation - or is it concession?
Generated Result	形容被授权的。
Word	antithesis
Context	his theory is the antithesis of mine.
Generated Result	形容与特定事件相反的。
Word	conditional
Context	the conditional sale will not be complete until the full purchase price is paid.
Generated Result	形容有条件的。
Word	lame
Context	the comedy aspect is a little lame, with too many one-liners
Generated Result	形容缺乏活力的。
Word	surge
Context	the testing equipment-maker 's shares surged as sales rose for the first time in six quarters.
Generated Result	形容特定对象数量增加。
Word	近
Context	我认为太阳在清早刚出来的时候离人近，中午的时候离人远。
Generated Result	close to or nearby.
Word	意外
Context	好在我们都已买了保险，如果发生意外，一切都由保险公司理赔。
Generated Result	an accidental occurrence.
Word	看法
Context	我希望七月初开院士会议时，能够再提出在这方面一些具体的看法。
Generated Result	the opinion of a person.
Word	终究
Context	走在错误的路上，终究是要输的。
Generated Result	Décrivez le résultat final de l'événement.
Word	日益
Context	融入实际生活的经验中，人生经验便日益丰富。
Generated Result	Le degré de description est approfondi.
Word	revitalize
Context	this food revitalized the patient.
Generated Result	Donner une nouvelle vitalité.
Word	enter
Context	enter a drug treatment program.
Generated Result	Participer à un programme ou un projet.

Table 12: The generated results of M2M model with Contrastive Prompt Learning method under low-resource setting.