Unsupervised Neural Dialect Translation with Commonality and Diversity Modeling

Yu Wan,* Baosong Yang,* Derek F. Wong,† Lidia S. Chao, Haihua Du, Ben C.H. Ao
NLP^{2}CT Lab, Department of Computer and Information Science, University of Macau
{nlp2ct.ywan, nlp2ct.baosong, nlp2ct.duhaihua, nlp2ct.benao}@gmail.com, {derekfw,lidiasc}@um.edu.mo

Abstract

As a special machine translation task, dialect translation has two main characteristics: 1) lack of parallel training corpus; and 2) possessing similar grammar between two sides of the translation. In this paper, we investigate how to exploit the commonality and diversity between dialects thus to build unsupervised translation models merely accessing to monolingual data. Specifically, we leverage pivot-private embedding, layer coordination, as well as parameter sharing to sufficiently model commonality and diversity among source and target, ranging from lexical, through syntactic, to semantic levels. In order to examine the effectiveness of the proposed models, we collect 20 million monolingual corpus for each of Mandarin and Cantonese, which are official language and the most widely used dialect in China. Experimental results reveal that our methods outperform rule-based simplified and traditional Chinese conversion and conventional unsupervised translation models over 12 BLEU scores.

Introduction

Dialect refers to a variant of a given language, which could be defined by factors of regional speech patterns, social class or ethnicity (Lyons 1981). Except for pronunciation, a dialect is also distinguished by its textual expression (Wong and Lee 2018). For instance, Mandarin (MAN) and Cantonese (CAN) are the official language and the most widely used dialect of China, respectively (Lee and Wong 1998). As seen in Fig. 1, although the sentences have absolutely same semantic meaning, they have distinct attributes with respect to the expression on text. Correspondingly, in this task we attempt to build automatic translation system for dialects.

An intuitive way is to leverage advanced machine translation systems which have recently yielded human-level performance with the use of neural networks (Chen et al. 2018; Li et al. 2018). Nevertheless, contrast with traditional machine translation, there are two main challenges in dialect translation. First, the success of supervised neural machine translation depends on large-scale training parallel data,

CAN	nei	yaugeiho	faan	ngukkei	,	jamdaam	tong	sin	。
	你	有几何	返	屋企	，	饮啖	汤	先	。
MAN	ni	nande	hui	jia	,	xian	hekou	tang	。
	你	难得	回	家	，	先	喝口	汤	。

Figure 1: An example of CAN-MAN translation.

while dialect translation is not equipped such kind of prerequisite. This makes our task fall into unsupervised learning category (Artetxe et al. 2018; Lample et al. 2018a; 2018c). Second, dialects are closely related and, despite their differences, often share similar grammar, e.g. morphology and syntax (Chambers and Trudgill 1998). The extraction of *commonality* is beneficial to unsupervised mapping (Lample et al. 2018b) and model robustness (Firat, Cho, and Bengio 2016), in the meanwhile, preserving the explicit *diversity* plays a crucial role in our dialect translation. Consequently, it is challenging to balance the commonality and diversity for dialect translation thus to improve its performance.

We approach the mentioned problems by proposing unsupervised neural dialect translation model, which is merely trained using monolingual corpus and sufficiently leverage commonality and diversity of dialects. Specifically, we train an advanced NMT model TRANSFORMER (Vaswani et al. 2017) with denoising reconstruction (Vincent et al. 2008) and back-translation (Sennrich, Haddow, and Birch 2016a), which aim at building common language model and mapping different attributes, respectively. We introduce several strategies into translation model for balancing the commonality and diversity: 1) parameter-sharing that forces dialects to share the same latent space; 2) pivot-private embedding which models similarities and differences at lexical level; and 3) layer coordination which enhances the interaction of features between two sides of translation.

In order to evaluate the effectiveness of the proposed model, we propose monolingual dialect corpus which consists of 20 million colloquial sentences for each of MAN[1] and CAN. The sentences are extracted from conversations and comments in forums, social medias as well as subti-

*Equal contribution.
†Corresponding author.

[1]For simplification, we regard official language as a dialect.

tles, and carefully filtered during data preprocessing.[2] Empirical results on two directions of MAN-CAN translation task demonstrate that the proposed model significantly outperforms existing unsupervised NMT (Lample et al. 2018c) with even fewer parameters. The quantitative and qualitative analyses verified the necessity of commonality and diversity modeling for dialect translation.

Preliminary

Neural machine translation (NMT) aims to use a neural network to build a translation model, which is trained to maximize the conditional distribution of sentence pairs (Bahdanau, Cho, and Bengio 2015; Sennrich, Haddow, and Birch 2016b; Vaswani et al. 2017). Given a source sentence $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_I\}$, conditional probability of its corresponding translation $\mathbf{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_J\}$ is defined as:

$$\mathbf{P}(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{|J|} \mathbf{P}(\mathbf{y}_j|\mathbf{Y}_{<j}, \mathbf{X}; \theta), \tag{1}$$

where $\mathbf{y}_j$ indicates the j-th target token. θ denotes the parameters of NMT model, which are optimized to minimize the following loss function over the training corpus $\mathbf{D}$:

$$\mathcal{L} = \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim\mathbf{D}}[-\log \mathbf{P}(\mathbf{Y}|\mathbf{X}; \theta)] \tag{2}$$

Such kind of auto-regressive translation process is generally achieved upon the encoder-decoder framework (Sutskever, Vinyals, and Le 2014). Specifically, the inputs of encoder $\mathbf{S}^0$ and decoder $\mathbf{T}^0$ are obtained by looking up source and target embeddings according to the input sentences $\mathbf{X}$ and $\mathbf{Y}$, respectively:

$$\mathbf{S}^0 = \mathrm{Emb}_{src}(\mathbf{X}) \qquad \in \mathbb{R}^{I\times d} \tag{3}$$

$$\mathbf{T}^0 = \mathrm{Emb}_{trg}(\mathbf{Y}) \qquad \in \mathbb{R}^{J\times d} \tag{4}$$

where d indicates the dimensionality. The encoder is composed of a stack of N identical layers. Given the input layer $\mathbf{S}^{n-1} \in \mathbb{R}^{I\times d}$, the output of the n-th layer can be formally expressed as:

$$\mathbf{S}^n = \mathrm{Layer}^n_{enc}(\mathbf{S}^{n-1}) \qquad \in \mathbb{R}^{I\times d} \tag{5}$$

The decoder is also composed of a stack of N identical layers. Contrary to the encoder which takes all the tokens into account, the decoder merely summarizes the forward representations in the input layer $\mathbf{T}^{n-1} \in \mathbb{R}^{J\times d}$ at each decoding step, since the subsequent representations are invisible. Besides, the generation process considers the contextual information of source sentence, by feeding the top layer of the encoder $\mathbf{S}^N$. Accordingly, the j-th representation in n-th decoding layer $\mathbf{T}^n = \{\mathbf{t}^n_1, \cdots, \mathbf{t}^n_J\}$ is calculated as:

$$\mathbf{t}^n_j = \mathrm{Layer}^n_{dec}(\mathbf{T}^{n-1}_{\leqslant j}, \mathrm{Att}^n(\mathbf{t}^{n-1}_j, \mathbf{S}^N)) \qquad \in \mathbb{R}^d \tag{6}$$

where $\mathrm{Att}(\cdot)$ indicates the attention model (Bahdanau, Cho, and Bengio 2015) which has recently been a basic module to allow a deep learning model to dynamically select related representations as needed. Finally, the conditional probability of the j-th target word $\mathbf{y}_j$ is calculated using a non-linear function $\mathrm{Softmax}(\cdot)$:

$$\mathbf{P}(\mathbf{y}_j|\mathbf{Y}_{<j}, \mathbf{X}; \theta) = \mathrm{Softmax}(\mathrm{Proj}(\mathbf{t}^N_j)) \tag{7}$$

In this section, we propose unsupervised neural dialect translation. We first serve the dialect translation as an unsupervised learning task to tackle with the low-resource problem. Moreover, concerning the commonality and diversity between dialects, we introduce pivot-private embedding and layer coordinating to improve the dialect translation model.

Dialect Translation with Unsupervised Learning

Despite the success of NMT over past years, the performance of a NMT model relies on large-scale parallel training corpus (Sennrich, Haddow, and Birch 2016a; Artetxe et al. 2018). As a low-resource translation task, dialect translation fails at leveraging conventional training strategy, since parallel resources are normally inaccessible. The scarcity of bilingual corpus leads to extraordinary challenging on building translation models for dialects. On the contrary, monolingual corpora is relatively easier to be collected. Partially inspired by recent studies on unsupervised NMT (Lample et al. 2018a; Artetxe et al. 2018; Lample et al. 2018c), we propose to build dialect translation model with unsupervised learning which merely depends on monolingual data. Generally, most of the features with respect to dialects are similar, while only a few of the surface information is different. To this end, we propose to divide the training process into two parts: 1) commonality modeling which learns to capture general features of all dialects; and 2) diversity modeling which builds connections between different expressions.

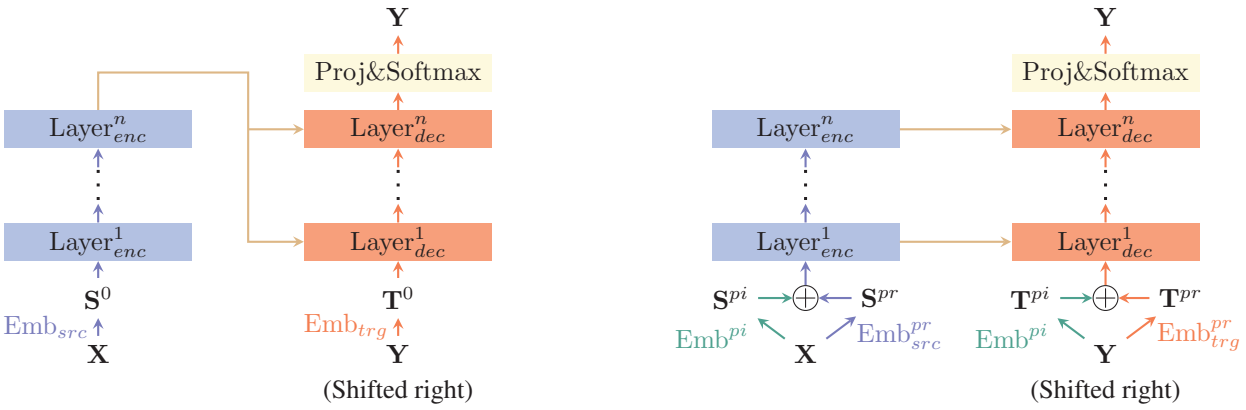
Commonality Modeling This procedure aims at offering our model the ability to extract the universal features of two dialects. Intuitively, the commonality modeling can be trained by reconstructing two dialects using one model. Artetxe et al. (2018) and Lample et al. (2018a) suggest that denoising autoencoding is beneficial to the language modeling. More importantly, it can avoid our model from severely copying the input sentence to the output. Contrary to Artetxe et al. (2018) and Lample et al. (2018a) who employ distinct model for each language, we train one model for both the two dialects, thus to encourage different dialects to be modeled under a common latent space. Consequently, the loss function is defined as:

$$\begin{aligned}\mathcal{L}_{com} =& \mathbb{E}_{\mathbf{X}\sim\mathbf{D}_X}[-\log \mathbf{P}(\mathbf{X}|\mathbf{X}^{noise}; \theta)] + \\ & \mathbb{E}_{\mathbf{Y}\sim\mathbf{D}_Y}[-\log \mathbf{P}(\mathbf{Y}|\mathbf{Y}^{noise}; \theta)]\end{aligned} \tag{8}$$

where $\mathbf{D}_X$ and $\mathbf{D}_Y$ are monolingual corpora for two dialects, $\mathbf{X}^{noise}$ and $\mathbf{Y}^{noise}$ denote noised inputs.[3] As seen, the two reconstruction models are shared with the same parameter θ.

[2] Our codes and data are released at: https://github.com/NLP2CT/Unsupervised_Dialect_Translation.

[3] We add noises to inputs by swapping, dropping and blanking words following Lample et al. (2018a), except that we swap two words rather than three, which shows better empirical results in our experiments.



(a) Conventional NMT model.

(b) The proposed model.

Figure 2: Illustration of (a) conventional NMT model and (b) the proposed model. As seen, we propose pivot-private embedding, which learns commonality (Emb^{pi}) and diversity (Emb^{pr}_{src} and Emb^{pr}_{trg}) at lexical level. Besides, the decoder attends to source representations layer by layer, rather than merely from the topmost layer.

Diversity Modeling Although there is marginal difference between dialects, the transfer of diversity is the key problem of dialect translation. Contrast to supervised NMT model which learns relevance between source and target using parallel data, dialect translation model fails to directly establish the functional mapping from source latent space to target one. An alternative way is to exploiting back-translation (Sennrich, Haddow, and Birch 2016a; Edunov et al. 2018). Specifically, $\mathbf{X}$ and $\mathbf{Y}$ are first translated to their candidate translation $\mathbf{Y}^{bak}$ and $\mathbf{X}^{bak}$, respectively. The mapping of cross-dialect latent spaces can be learned by minimizing:

$$\begin{aligned}\mathcal{L}_{div} =& \mathbb{E}_{\mathbf{X}\sim\mathbf{D}_X}[-\log \mathbf{P}(\mathbf{X}|\mathbf{Y}^{bak};\theta)]+ \\ & \mathbb{E}_{\mathbf{Y}\sim\mathbf{D}_Y}[-\log \mathbf{P}(\mathbf{Y}|\mathbf{X}^{bak};\theta)]\end{aligned} \tag{9}$$

Finally, the loss function in Equation 2 is modified as:

$$\mathcal{L} = \lambda_{com}\mathcal{L}_{com} + \lambda_{div}\mathcal{L}_{div} \tag{10}$$

where $\lambda_{com}, \lambda_{div}$ are hyper-parameters balancing the importance of commonality and diversity modeling, respectively.

Pivot-Private Embedding

An open problem in unsupervised NMT is the initialization of the translation model, which plays a crucial role in the iteratively training (Lample et al. 2018a; Artetxe et al. 2018) and affects the final performance of the unsupervised learning (Lample et al. 2018c). For two languages with different vocabularies, an usual solution in recent studies is to map the same tokens which are then cast as seeds for aligning other words (Artetxe et al. 2018; Lample et al. 2018a). For example, Artetxe et al. (2018) employ unsupervised bilingual word embeddings (Artetxe, Labaka, and Agirre 2017), while Lample et al. (2018c) utilize the representations of shared tokens (Mikolov et al. 2013) in different languages to initialize the lookup tables. Fortunately, dialect translation dispels this problem since most of tokens are shared among dialects. Therefore, we propose pivot and private embedding, in which, the former learns to share a part of the features while the latter captures the word-level characteristics in different dialects.

Pivot Embedding Since vocabularies in different dialects are almost the same, we join monolingual corpora of two dialects and extract all the tokens in it. In order to build the connections between source and target, we assign pivot embedding with d_s dimensions as the initial alignments:

$$\mathbf{S}^{pi} = \text{Emb}^{pi}(\mathbf{X}) \qquad \in \mathbb{R}^{I\times d_s} \tag{11}$$

$$\mathbf{T}^{pi} = \text{Emb}^{pi}(\mathbf{Y}) \qquad \in \mathbb{R}^{J\times d_s} \tag{12}$$

where the function of looking up embedding $\text{Emb}^{pi}(\cdot)$ shares parameters across dialects.

Private Embedding Except the common features, there also exists differences between dialects. We argue that such kind of difference mainly lies in the word-level surface information. To this end, we introduce private embedding for each translation side to distinguish and maintain the characteristics in dialects:

$$\mathbf{S}^{pr} = \text{Emb}^{pr}_{src}(\mathbf{X}) \qquad \in \mathbb{R}^{I\times(d-d_s)} \tag{13}$$

$$\mathbf{T}^{pr} = \text{Emb}^{pr}_{trg}(\mathbf{Y}) \qquad \in \mathbb{R}^{J\times(d-d_s)} \tag{14}$$

Contrary to pivot embedding, $\text{Emb}^{pr}_{src}(\cdot)$ and $\text{Emb}^{pr}_{trg}(\cdot)$ are assigned distinct parameters. Thus, the final input embedding in Equation 3 and 4 are modified as:

$$\mathbf{S}^0 = \mathbf{S}^{pr} \oplus \mathbf{S}^{pi} \qquad \in \mathbb{R}^{I\times d} \tag{15}$$

$$\mathbf{T}^0 = \mathbf{T}^{pr} \oplus \mathbf{T}^{pi} \qquad \in \mathbb{R}^{J\times d} \tag{16}$$

where $\oplus$ is the concatenation operator. Note that, since each token has d_s and $d - d_s$ dimensions for the associate pivot embedding and private embedding, the final input is still

composed of d-dimensional vector. $\mathrm{Emb}^{pi}(\cdot)$, $\mathrm{Emb}^{pr}_{src}(\cdot)$ and $\mathrm{Emb}^{pr}_{trg}(\cdot)$ are all pretrained, and co-optimized under the translation objective. In this way, we expect that pivot embedding can enhance the commonality of translation model, while private embedding raises the ability to capture the diversity of different dialects (Liu et al. 2019).

Layer Coordination

Recent studies have pointed out that multiple neural network layers are able to capture different types of syntactic and semantic information (Peters et al. 2018; Li et al. 2019). For example, Peters et al. (2018) demonstrate that higher-level layer states capture the context-dependent aspects of word meaning while lower-level states model the aspects of syntax, and simultaneously exposing all of these signals is highly beneficial. To sufficiently interact these features, an alternative way is to perform attention from a decoder layer to its corresponding encoder layer, rather than merely from the topmost layer. Accordingly, the n-th decoding layer (Equation 6) is changed to:

$$\mathbf{t}^n_j = \mathrm{Layer}^n_{dec}(\mathbf{T}^{n-1}_{\leqslant j}, \mathrm{Att}^n(\mathbf{t}^{n-1}_j, \mathbf{S}^n)) \quad \in \mathbb{R}^d \qquad (17)$$

This technique has been proven effective (He et al. 2018; Yang et al. 2019a; Hao et al. 2019) upon NMT tasks via shortening the path of gradient propagation, thus stabilizes the training of a extremely deep model. However, the improvements on traditional translation tasks become marginal when we apply layer coordination to the models with less than 6 layers (He et al. 2018). We attribute this to the fact that directly interacting lexical and syntactic level information between different languages affects the diversity modeling of them, since it forces the two languages to share the same latent space layer by layer. Different from prior studies, our work focuses on a pair of languages which have extremely similar grammar. We examine whether layer coordination is conductive to commonality modeling of dialects and the translation quality.

Datasets

In this section, we first introduce the CAN and MAN datasets collected for our experiments, then show adequate rudimentary statistical results upon training corpora.

Monolingual Corpora The lack of CAN monolingual corpora with strong colloquial features is serious obstacle in our research. Existing CAN corpora, such as HKCanCor (Luke and Wong 2015) and CANCORP (Lee and Wong 1998), all have the following shortcomings: 1) they were collected in rather early years, the linguistic features of which vary from the current ones due to language evolution; and 2) they are scarce for data-intensive unsupervised training. Due to the fact that colloquial corpora possess more distinguished linguistic features of CAN, we collect CAN sentences among domains including talks, comments and dialogues from scratch.[4] In order to maintain the consistency of training sets, MAN corpora are also derived from same domains as CAN from ChineseNlpCorpus and Large Scale Chinese Corpus for NLP.[5]

Dialect	# Sents	Vocab size	Unique
CAN	20M	9,025	541
MAN	20M	8,856	372

Table 1: Statistics of two monolingual corpora after preprocessing. We conduct experiment at character-based level, and the joint vocabulary size is exactly 9,397.

Parallel Corpus We collect adequate parallel corpora for the development and evaluation of models. Parallel sentence pairs from dialogues are manually selected by native CAN and MAN speakers. Consequently, 1,227 and 1,085 sentence pairs are selected as development and test set, respectively.

Data Preprocessing & Statistics As there is no well-performed CAN segment toolkit, we conduct all the experiments at character level. In order to share the commonality of both languages and reduce the size of vocabularies, we convert all the texts into simplified Chinese.[6] For reasons of computational efficiency, we keep the sentences whose length lies between 4 and 32, and remove sentences composing characters with low frequencies. Finally, each of MAN and CAN monolingual training corpora consists of 20M sentences. The statistics of training set are concluded in Tab. 1. As seen, CAN has larger vocabulary size and more unique characters than MAN. To identify the commonality and diversity of CAN and MAN, we compute the Spearman's rank correlation coefficient (Zhelezniak et al. 2019) between two vocabulary rankings by their frequencies within each corpus. The coefficient score of two full vocabularies is 0.81 ($p < 0.001$), meaning that the overall relation is significantly strong. While the coefficient score of the 250 most frequent tokens is 0.26 ($p < 0.001$), indicating that the relation is significantly weak. These results cater to our hypothesis that dialects share considerate commonality with each other, but possess diversity upon most frequent tokens.

Experiments

Experimental Setting

We use TRANSFORMER (Vaswani et al. 2017) as our model architecture, and follow the base model setting for our model dimensionalities. We refer to the parameter setting of Lample et al. (2018c), and implement our approach on top of their source code.[7] We use BLEU score as the evaluation

[4] https://www.wikipedia.org, https://www.cyberctm.com, http://discuss.hk and https://lihkg.com.

[5] https://github.com/brightmart/nlp_chinese_corpus and https://github.com/SophonPlus/ChineseNlpCorpus.

[6] We also attempt to transform all the texts into traditional characters. It does not work well since some simplified characters has multiple corresponding traditional characters and such kind of one-to-many mapping results in ambiguity and data sparsity.

[7] https://github.com/facebookresearch/UnsupervisedMT

Model	CAN⇒MAN	MAN⇒CAN	# Params (M)
Baseline			
Character-level Rule-based Transition	42.18	42.27	-
Unsupervised Style Transfer (Hu et al. 2017)	41.97	42.03	14.40
Unsupervised PB-SMT (Lample et al. 2018c)	42.12	42.20	-
Unsupervised NMT (Lample et al. 2018c)	42.90	42.39	39.08
Ours			
Layer Coordination	48.45	43.11	39.08
Pivot-Private Embedding	52.74	46.69	36.65
Pivot-Private Embedding + Layer Coordination	**54.95**	**47.45**	36.65

Table 2: Experimental results on unsupervised dialect neural machine translation. # Params (M): number of parameters in million. We can see that layer coordination provides improvement over baseline on both directions, and pivot-private embedding improves the result further by almost 10 BLEU scores on CAN⇒MAN. Combining both layer coordination and pivot-private embedding gives the best result, exceeding 12 and 5 BLEU scores than baseline NMT system on two directions, respectively.

metric. The training of each model was early-stopped to maximize BLEU score on the development set.

All the embeddings are pretrained using fasttext (Bojanowski et al. 2017),[8] and pivot embeddings are derived from concatenated training corpora. In the procedure of training, λ_{div} is set to 1.0, while λ_{com} is linearly decayed from 1.0 at the beginning to 0.0 at the step being 200k.

Baseline We compare our model with four systems:

- We collect simple CAN-MAN conversion rules and regard character-level transition as one of our baseline systems.
- Our model is built upon unsupervised NMT methods, we choose one of the most widely used architecture (Lample et al. 2018c) as our baseline system.
- Moreover, unsupervised phrase-based statistical MT (Lample et al. 2018c) has comparable performance to its NMT counterpart. Therefore, we also take unsupervised PB-SMT model into account.
- For reference, we also examine whether a style transfer system (Hu et al. 2017) can handle dialect translation task.

Overall Performances

Tab. 2 lists the experimental results. As seen, character-level rule-based translation system performs comparably with conventional unsupervised NMT system. This is in accord with Lample et al. (2018c) that training process of unsupervised NMT is vulnerable, because no aligned information between languages can be afforded to model training. Relatively, character transition rules offer adequate aligned references to conduct the fairish results. Besides, the unsupervised PB-SMT model performs slightly worse than NMT system, a possible reason is that the model is hard to extract a well-performed phrase table from colloquial data (Laurens et al. 1997). We also evaluate a style transfer system (Hu et al. 2017). The model underperforms unsupervised NMT baseline, indicating that, to some extent, style transfer is not adequate for dialect translation.

[8]https://github.com/facebookresearch/fastText

Model	CAN⇒MAN	MAN⇒CAN
Baseline	1.80 ± 0.44	2.57 ± 0.50
Our Model	$2.50 \pm 0.87 \uparrow$	$3.16 \pm 0.61 \uparrow$

Table 3: Human assessment on our experimental results. ↑: improvement is strongly significant ($p < 0.01$).

As to our proposed methods, layer coordination improves the performance by more than 5 BLEU scores at CAN⇒MAN direction, proving that sharing coordinate information at the same semantic level among dialects is effective. Besides, using pivot-private embedding further gives a higher increase of nearly 10 BLEU scores as well as reducing the model size, verifying that jointly modeling commonality and diversity of both dialects is both effective and efficient. Furthermore, combining both of above can give us more than 12 BLEU scores improvement than baseline NMT system, revealing that both pivot-private embedding and layer coordination are complementary to each other. As to the MAN⇒CAN direction, we can also observe improvements of our proposed methods. Translating MAN to CAN is more difficult since it contains more one-to-many character-level transition cases than its reversed direction. Despite this, our best approach still gains 5 BLEU scores improvement than baseline systems on MAN⇒CAN translation, revealing the universal effectiveness of our proposed method.

Human Assessment Since BLEU metric may be insufficient to reflect the quality of oral sentences, we randomly extract 50 CAN ⇒ MAN and 50 MAN ⇒ CAN examples from test set for human evaluation, respectively. Each example contains source sentence, translated sentences from Unsupervised NMT model ("baseline") and our proposed model. Each native speaker is asked to present a score ranging from 1 to 4 to determine the translation quality of each translated result within each example. Each of the reported result is the average score assessed by 10 native speakers. As seen in Tab. 3, results prove that proposed method significantly outperforms baseline NMT system ($p < 0.01$) in both CAN⇒MAN and MAN⇒CAN directions.

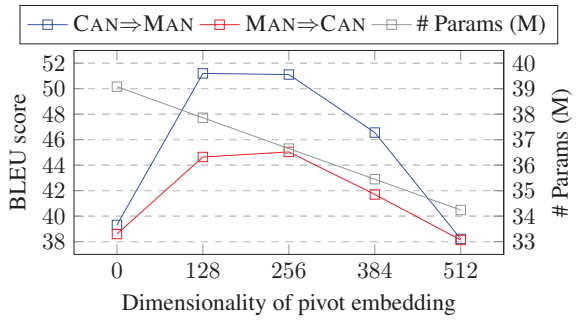


Figure 3: Model performances with various pivot embedding dimensionalities upon dev set. # Params (M): number of parameters in million. We can observe that applying adequate dimensionality to pivot embedding is effective, rather than non-sharing any dimension among two dialects (dimensionality is 0) or sharing all dimensions (dimensionality is 512).

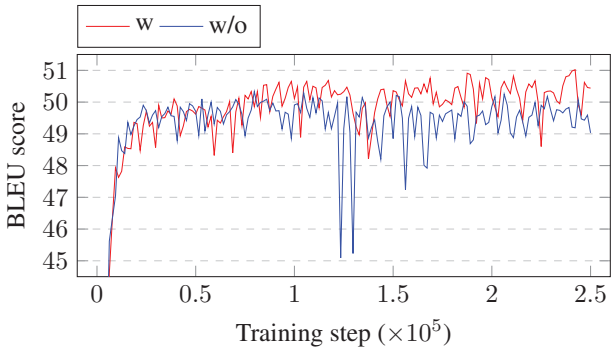


Figure 4: Learning curves of models upon dev set. Model with layer coordination (w) reaches its convergence at around step 240k, while model without (w/o) at around step 200k. As seen in this figure, applying layer coordination improves the performance of dialect translation model, as well as significantly stabilizes the training process.

Effectiveness of Pivot-Private Embedding

To investigate the effectiveness of pivot-private embedding, we also conduct further research on the dimensionality of pivot embedding. As seen in Fig. 3, adequately sharing part of word embedding among dialects can greatly improve the effect, while using two independent sets of embedding for dialects, or sharing all dimensions of embedding leads to poor results. This indicates the importance of balancing the commonality and diversity for dialect translation. Moreover, the more the dimensionalities assigned to pivot embedding, the fewer the parameters required by models. We argue that using pivot-private embedding is not only an efficient way to augment the ability of dialect translation system to model diversity, but also offer an alternative way to relieve the effect of over-parameterization.

Comparing to the model with the dimensionality being 128, the model with 256 pivot embedding dimensions yields comparable results on the two translation directions, while assigns fewer parameters. Consequently, we apply 256 as our default setting for pivot embedding dimensionality.

Effectiveness of Layer Coordination

Layer coordination intuitively interacts features from all dialects, helping the model to capture the commonality of linguistic features at coordinate semantic level (Peters et al. 2018). He et al. (2018) reveal that layer coordination offers more aligned features at the same level, from lexical, through syntactic, to semantic. In this section, we investigate how layer coordination effects on translation quality.

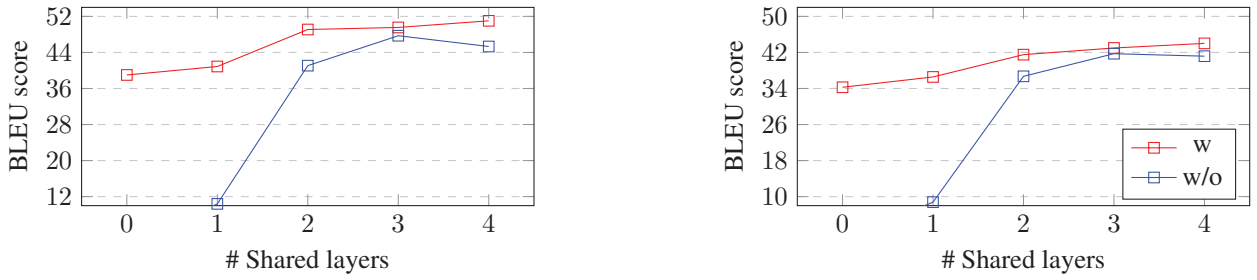
Stability Analysis We first visualize the convergence of models with and without layer coordination. From Fig. 4 we can observe that the model with layer coordination gains a steady training process, whereas training process of model without layer coordination is fragile, especially drop nearly 5 BLEU scores upon dev set at the middle term. We attribute this to the fact that layer coordination provides coordinate semantic information (He et al. 2018), which is beneficial to our dialect translation task with respect to commonality modeling. Since the two dialects share similar features, each decoder layer can leverage more fine-grained information from source side at the same semantic level, instead of only exploiting top-level representations.

Parameter Sharing For further investigation, we also conduct analyses on the effect of shared layers. As visualized in Fig. 5, baseline system performs worse when the number of shared layer is less than 1, and models with 3 layers shared performs better. This is consistent with findings in Lample et al. (2018c) who suggest to share higher 3 layers in encoder and lower 3 ones in decoder. Considering the proposed model, sharing more layers for CAN and MAN translation on both directions is profitable, and model with all layers shared gives the best performance on both directions. This demonstrates that CAN and MAN have more similar characteristics in numerous aspects of linguistics than distant languages (Artetxe et al. 2018; Lample et al. 2018a), and layer coordination also contributes to the balance of commonality and diversity modeling upon dialect translation task.

Related Work

In this section, we will give an account of related research.

Dialect Translation To the best of our knowledge, related studies on dialect translation have been carried upon a lot of languages. For example, in Arabic (Baniata, Park, and Park 2018) and Indian (Chakraborty, Sinha, and Nath 2018), applying syllable symbols is effective for sharing information across languages. Compared to these tasks, our work mainly focus on handling problems in CAN and MAN translation task. CAN and MAN have little syllable information in common, as even the same character can be widely divergent in aspect of pronunciation (Lee and Wong 1998; Wong and Lee 2018). To push the difference further, a set of CAN characters is quite rarely to be seen in MAN, because



(a) MAN⇒CAN (b) CAN⇒MAN

Figure 5: Experiments on number of shared encoder/decoder layers upon dev set. Here w and w/o denotes with and without layer coordination, respectively. From both figures, we can see that even without any shared layer, model with layer coordination can also be trainable rather than without. Models without layer coordination gain significant improvement upon sharing adequate layers for two dialects, while the performances decrease if all layers are shared. As to proposed layer coordination, the more layers shared for two dialects, the higher performance models can possess.

CAN is a dialect that without formal regulation of written characters (Lee and Wong 1998). Moreover, younger CAN speakers more likely refer to use phonetic labels (e.g. "d" responses to "di") or homophonetic character symbols instead of ground truth, which raises intractable issues when building the translation model.

Unsupervised Learning Our work refers to quantitative researches on unsupervised machine translation (Lample et al. 2018a; Artetxe et al. 2018; Lample et al. 2018c), which compose a well-designed training schedule for unsupervised translation tasks. The difference between our research and theirs mainly lies in the similarity of involved languages, where dialects in our research are far similar with each other than those in unsupervised NMT tasks.

Moreover, our research is closely related to studies on style transfer (Hu et al. 2017; Prabhumoye et al. 2018). There are two main differences between our task and style transfer. Firstly, the source and target sides in style transfer task belong to the same language, where the difference mainly contributed by style, e.g. sentiment (Hu et al. 2017), while dialect translation has to identically guarantee the semantics between two sides. Secondly, there are more commonalities between source and target in style transfer than that in dialect translation. The former focus on the transition of different styles, the two sides can sometimes be distinguished by only a few words. Nevertheless, dialects have wide discrepancies which vary from vocabulary and word frequency to syntactic structure.

Methodologically, compare to the mentioned studies, we motivated by similarity and difference between dialects and propose pivot-private embedding and layer coordination to jointly balance *commonality* and *diversity*.

Conclusions and Future Work

In this study, we investigate the feasibility of building a dialect machine translation system. Due to the lack of parallel training corpus, we approach the problem with unsupervised learning. Considering the characteristics in dialect translation, we further improve our translation model by contributing pivot-private embedding and layer coordination, thus enriching the mutual linguistic information sharing across dialects (CAN-MAN). Our experimental results confirm that our improvements are universally-effectiveness and complementary to each other. Our contributions are mainly in:

- We propose dialect translation task, and conduct massive examples of monolingual sentences with respect to dialects of spoken MAN and CAN;
- We apply an unsupervised learning algorithm to accomplish CAN-MAN dialect translation task. We leverage *commonality* and *diversity* modeling to strengthen the translation functionality among dialects, including pivot-private embedding and layer coordination;
- Our approach outperforms conventional unsupervised NMT system over 12 BLEU scores, achieving a considerable performance and a new benchmark for the proposed CAN-MAN translation task.

In the future, it is interesting to validate our principles, i.e. commonality and diversity modeling, into other tasks, such as conventional machine translation and style transfer. Another promising direction is to incorporate linguistic knowledge into unsupervised learning procedure, e.g. phrasal pattern (Xu et al. 2019), word order information (Yang et al. 2019b) and syntactic structure (Yang et al. 2019c).

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61672555), the Joint Project of the Science and Technology Development Fund, Macau SAR and National Natural Science Foundation of China (Grant No. 045/2017/AFJ), the Science and Technology Development Fund, Macau SAR (Grant No. 0101/2019/A2), and the Multi-year Research Grant from the University of Macau (Grant No. MYRG2017-00087-FST). We thank all the reviewers for their insightful comments.

References

Artetxe, M.; Labaka, G.; Agirre, E.; and Cho, K. 2018. Unsupervised Neural Machine Translation. In *ICLR*.

Artetxe, M.; Labaka, G.; and Agirre, E. 2017. Learning Bilingual Word Embeddings With (Almost) No Bilingual Data. In *ACL*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

Baniata, L. H.; Park, S.; and Park, S.-B. 2018. A Neural Machine Translation Model for Arabic Dialects That Utilizes Multitask Learning (MTL). *Computational intelligence and neuroscience* 2018.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching Word Vectors With Subword Information. *TACL* 5.

Chakraborty, S.; Sinha, A.; and Nath, S. 2018. A Bengali-Sylheti Rule-Based Dialect Translation System: Proposal and Preliminary System. In *I3CS*.

Chambers, J. K., and Trudgill, P. 1998. *Dialectology*. Cambridge Textbooks in Linguistics. Cambridge University Press, 2 edition.

Chen, M. X.; Firat, O.; Bapna, A.; Johnson, M.; Macherey, W.; Foster, G.; Jones, L.; Niki, P.; Schuster, M.; Chen, Z.; Wu, Y.; and Hughes, M. 2018. The Best Of Both Worlds: Combining Recent Advances In Neural Machine Translation. In *ACL*.

Edunov, S.; Ott, M.; Auli, M.; and Grangier, D. 2018. Understanding Back-Translation at Scale. In *EMNLP*.

Firat, O.; Cho, K.; and Bengio, Y. 2016. Multi-Way, Multilingual Neural Machine Translation With A Shared Attention Mechanism. In *NAACL*.

Hao, J.; Wang, X.; Yang, B.; Wang, L.; Zhang, J.; and Tu, Z. 2019. Modeling Recurrence for Transformer. In *NAACL*.

He, T.; Tan, X.; Xia, Y.; He, D.; Qin, T.; Chen, Z.; and Liu, T.-Y. 2018. Layer-Wise Coordination Between Encoder and Decoder for Neural Machine Translation. In *NIPS*.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward Controlled Generation of Text. In *ICML*.

Lample, G.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018a. Unsupervised Machine Translation Using Monolingual Corpora Only. In *ICLR*.

Lample, G.; Conneau, A.; Ranzato, M.; Denoyer, L.; and Jégou, H. 2018b. Word Translation Without Parallel Data. In *ICLR*.

Lample, G.; Ott, M.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018c. Phrase-Based & Neural Unsupervised Machine Translation. In *EMNLP*.

Laurens, F. P. D. T. O.; McFetridge, P.; Maricela, J. D. N. P. M.; Pidruchney, C.-P. L.; and MacDonald, S. 1997. A Lexicalist Approach to the Translation of Colloquial Text. In *TMI*.

Lee, T., and Wong, C. 1998. CANCORP: The Hong Kong Cantonese Child Language Corpus. *Cahiers de Linguistique Asie Orientale* 27.

Li, J.; Tu, Z.; Yang, B.; Lyu, M. R.; and Zhang, T. 2018. Multi-Head Attention with Disagreement Regularization. In *EMNLP*.

Li, J.; Yang, B.; Dou, Z.-Y.; Wang, X.; Lyu, M. R.; and Tu, Z. 2019. Information Aggregation for Multi-Head Attention with Routing-by-Agreement. In *NAACL*.

Liu, X.; Wong, D. F.; Liu, Y.; Chao, L. S.; Xiao, T.; and Zhu, J. 2019. Shared-Private Bilingual Word Embeddings for Neural Machine Translation. In *ACL*.

Luke, K. K., and Wong, M. L. 2015. The Hong Kong Cantonese Corpus: Design and Uses. *Journal of Chinese Linguistics* 25.

Lyons, J. 1981. *Language and Linguistics*. Cambridge University Press.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *NIPS*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *NAACL*.

Prabhumoye, S.; Tsvetkov, Y.; Salakhutdinov, R.; and Black, A. W. 2018. Style Transfer Through Back-Translation. In *ACL*.

Sennrich, R.; Haddow, B.; and Birch, A. 2016a. Improving Neural Machine Translation Models With Monolingual Data. In *ACL*.

Sennrich, R.; Haddow, B.; and Birch, A. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *ACL*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning With Neural Networks. In *NIPS*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. In *NIPS*.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and Composing Robust Features with Denoising Autoencoders. In *ICML*.

Wong, T.-s., and Lee, J. 2018. Register-Sensitive Translation: A Case Study of Mandarin and Cantonese (Non-Archival Extended Abstract). In *AMTA*.

Xu, M.; Wong, D. F.; Yang, B.; Zhang, Y.; and Chao, L. S. 2019. Leveraging Local and Global Patterns for Self-Attention Networks. In *ACL*.

Yang, B.; Li, J.; Wong, D.; Chao, L. S.; Wang, X.; and Tu, Z. 2019a. Context-Aware Self-Attention Networks. In *AAAI*.

Yang, B.; Wang, L.; Wong, D. F.; Chao, L. S.; and Tu, Z. 2019b. Assessing the Ability of Self-Attention Networks to Learn Word Order. In *ACL*.

Yang, B.; Wong, D. F.; Chao, L. S.; and Zhang, M. 2019c. Improving Tree-based Neural Machine Translation with Dynamic Lexicalized Dependency Encoding. *Knowledge-Based Systems* 105042.

Zhelezniak, V.; Savkov, A.; Shen, A.; and Hammerla, N. 2019. Correlation Coefficients and Semantic Textual Similarity. In *NAACL*.