Meet Changes with Constancy: Learning Invariance in Multi-Source Translation

Jianfeng Liu$^{\diamondsuit\triangle\clubsuit}$**, Ling Luo**$^{\diamondsuit\triangle\blacklozenge}$**, Xiang Ao**$^{\diamondsuit\triangle\blacklozenge\dagger}$**, Yan Song**$^{\spadesuit\heartsuit}$**, Haoran Xu**$^{\diamondsuit\triangle}$**, Jian Ye**$^{\diamondsuit\clubsuit\dagger}$

$^{\diamondsuit}$Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
$^{\triangle}$University of Chinese Academy of Sciences
$^{\clubsuit}$Beijing Key Laboratory of Mobile Computing and Pervasive Device
$^{\blacklozenge}$Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences
$^{\spadesuit}$The Chinese University of Hong Kong (Shenzhen) $^{\heartsuit}$Shenzhen Research Institute of Big Data
{liujianfeng18s;luoling18s;aoxiang;xuhaoran18s;jye}@ict.ac.cn
songyan@cuhk.edu.cn

Abstract

Multi-source neural machine translation aims to translate from parallel sources of information (e.g. languages, images, etc.) to a single target language, which has shown better performance than most one-to-one systems. Despite the remarkable success of existing models, they usually neglect the fact that multiple source inputs may have inconsistencies. Such differences might bring noise to the task and limit the performance of existing multi-source NMT approaches due to their indiscriminate usage of input sources for target word predictions. In this paper, we attempt to leverage the potential complementary information among distinct sources and alleviate the occasional conflicts of them. To accomplish that, we propose a *source invariance network* to learn the invariant information of parallel sources. Such network can be easily integrated with multi-encoder based multi-source NMT methods (e.g. multi-encoder RNN and transformer) to enhance the translation results. Extensive experiments on two multi-source translation tasks demonstrate that the proposed approach not only achieves clear gains in translation quality but also captures implicit invariance between different sources.

1 Introduction

Neural machine translation (NMT) systems in general translate one source language to a target language. Various one-to-one attentional encoder-decoder architectures (Bahdanau et al., 2014; Luong et al., 2015b; Vaswani et al., 2017) were designed to learn word and structure mappings including formations, grammatical correspondences between source and target languages.

Recently, multi-source NMT (Zoph and Knight, 2016), which simultaneously takes multiple different languages (Dabre et al., 2017; Libovickỳ et al., 2018; Currey and Heafield, 2018) as input when translating to another one, is emerging. Its intuitive idea is ambiguity between one source language and the target language could be reduced by another source language via the "triangulation" proposed by (Kay, 2000). For example, it is hard to tell who was with a telescope by the ambiguous English sentence in Figure 1 (a). But it could be more easily to be translated to Spanish if provided with the corresponding Chinese phrase "我通过望远镜", which corresponds to "I see through a telescope" in English. Besides, the complementary source could be extended to non-language input such as images, also known as multi-modal NMT (Libovickỳ and Helcl, 2017; Elliott et al., 2017). It takes an image with description in source languages, that is then translated into a target language. The images are expected to provide additional signals for better translations, which abides by similar assumptions in multi-source NMT with different languages. A specific instance is illustrated in Figure 1 (b). In this paper, we consider general multi-source NMT which contains more than one source as input. Our proposed model could be naturally adapted to these two specific sub-tasks, namely multi-lingual translation and multi-modal translation. For simplicity and convenience to describe, we use multi-source NMT to denote both tasks.

†: Corresponding Author

Proceedings of the 28th International Conference on Computational Linguistics, pages 1122–1132
Barcelona, Spain (Online), December 8-13, 2020

(a)	***English*** ***Chinese***	→	***Spanish***
	I saw someone on the hill with a telescope. 我通过望远镜看到山上有个人。		*Vi a alguien en la montaña a través de un telescopio.*
(b)	***English*** ***<u>Picture</u>***	→	***French***
	A woman on the floor covered in cake .		*une femme sur le sol couverte de gâteau .*

Figure 1: specific translation instances.

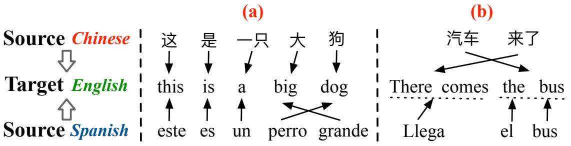

Figure 2: Examples of (a) multi-lingual translation and (b) multi-modal translation.

The existing approaches for multi-source NMT mainly adopt multi-encoder-single-decoder as a paradigm, where multiple encoders extract features of different sources, and the single decoder jointly use the learnt sources' representations to predict the target words. They have already shown better performance than most one-to-one systems (Zoph and Knight, 2016; Junczys-Dowmunt and Grundkiewicz, 2018). Despite the remarkable success of existing methods, they usually neglect the fact that parallel corpus in multiple sources may have inconsistencies, which might derive noise to confuse the decoding process and cause unsatisfactory translation results. Take multi-lingual translation as an example, two inconsistent cases are exhibited in Figure 2. The target English sentence "this is a big dog" only requires simple sequential translation process from Chinese but necessary inversions from Spanish. However, this kind of pattern is not always true and an opposite case is shown in Figure 2 (b). The analogous cases also could happen in multi-modal NMT since images might contain noisy information brought by image background, rotation and scaling, etc. Even with attention mechanisms, we find the conflicts may limit the performance of existing approaches due to their indiscriminate usage of input sources. We will verify our findings in the experiment section (refer to Section 6.2).

In this paper, we attempt to take advantage of the potential complementary information among multiple sources and alleviate the serendipitous conflicts of them. We accomplish it by proposing a component named **S**ource **I**nvariance **N**etwork (SIN) for multi-encoder based multi-source NMT approaches. SIN separates the invariant information of parallel input sources from their individual representations. The learnt invariant representations could be utilized into decoding processing to enhance the translation quality. SIN is easily integrated with popular multi-encoder based multi-source NMT frameworks such as RNN-based and transformer-based approaches. Experiments on both multi-lingual and multi-modal translation tasks demonstrate the effectiveness of SIN. Our contributions can be summarized as follows,

- To the best of our knowledge, we are the first considering the problem of learning invariance of multi-source translation to address the occasional conflicts among different parallel input sources.

- We devise the source invariance network to automatically separate the invariant information of parallel input sources and leverage such component to enhance two prevail multi-encoder based multi-source NMT frameworks.

- We verify the performance of our model on both multi-lingual and multi-modal NMT tasks. Extensive experimental results show that our SIN can provide large-margin improvements on both tasks and the invariant information between different sources are encouragingly learnt by our model.

2 Related work

Multi-lingual Machine Translation Multilingual machine translation addresses the machine translation between multiple source and target languages, which contains one-to-many (one-source-to-many-target), many-to-many, many-to-one approaches. (Dong et al., 2015) combines a single encoder with multiple attentional decoders for one-to-many translation, based on which (Wang et al., 2018) have proposed three strategies to improve the performance. (Luong et al., 2015a) combined multiple encoders and decoders, one encoder for each source language and one decoder for each target language respectively, for many-to-many translation. Based on that, (Firat et al., 2016) devised a sharing attention mechanism while (Lu et al., 2018) incorporated an explicit neural interlingua into such multilingual encoder-decoder.

Afterwards, (Ha et al., 2016) and (Johnson et al., 2017) proposed one universal encoder and decoder to take place of multiple encoders and decoders. (Blackwood et al., 2018) devised a task-specific attention mechanism to improve the translation quality. In addition, (Sen et al., 2019) considered unsupervised multilingual NMT by utilizing a shared encoder and some language-specific decoders. As regard to many-to-one translation, RNN-based (Zoph and Knight, 2016) and Transformer-based multi-source translation (Junczys-Dowmunt and Grundkiewicz, 2018) are two multi-encoder methods.

Multi-modal Machine Translation Multi-modal have been extensively studied due to multi-modal MT shared tasks (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). Prior to the shared tasks, (Hitschler et al., 2016) proposed a phrase-based statistical MT model (PBSMT) to generate translation candidates and re-rank them with image features. Afterwards, RNN-based architectures have been adopted in multi-modal machine translation (Elliott et al., 2015). Based on such framework, (Libovickỳ and Helcl, 2017) devised attention mechanism to improve the translation while (Caglayan et al., 2016) leveraged spatial visual features to a separate visual attention mechanism. Apart from that, (Calixto et al., 2019) incorporated image features through latent variables. (Libovickỳ et al., 2018) adopted multi-source Transformer with different input combination strategies.

To the best of our knowledge, previous methods did not consider the conflicts among various input sources. We are the first to consider this problem in multi-source translation and learn the invariance to improve the translation quality.

3 Background

In this section, we introduce two general multi-encoder based multi-source NMT frameworks, namely RNN-based and Transformer-based approaches. They basically have the same architecture with multiple encoders and a single decoder corresponding to different sources and one target, respectively.

3.1 RNN-based Multi-source NMT

In RNN-based multi-source NMT, various inputs are firstly encoded by individual encoders. Then, in the decoding process, e.g. at step t, context vectors $c_t^1, c_t^2, \ldots, c_t^N$ corresponding to N different encoders are all combined with the decoder hidden state h_t to obtain a new state $\tilde{h}_t$, which is further utilized for the word prediction. For example, (Zoph and Knight, 2016) simply concatenate the context vectors with decoder hidden state as,

$$\tilde{h}_t = \tanh(W_c[h_t, c_t^1, \ldots, c_t^N]). \tag{1}$$

where $W_c \in \mathbb{R}^{d \times (N+1)d}$, $h_t, c_t^1, \ldots, c_t^N \in \mathbb{R}^d$ and d represents the dimension of hidden states.

3.2 Transformer-based Multi-source NMT

Another essential multi-encoder based multi-source NMT framework (Junczys-Dowmunt and Grundkiewicz, 2018) is built upon Transformer (Vaswani et al., 2017). Compared with RNN-based methods, the major distinction of transformer-based multi-source NMT is that there are multiple encoder-decoder attentions, which are connected in series as shown in Figure 4. Each attention layer conducts multi-head attention by considering the output of the previous attention layer as the "Query" while taking in the output from the corresponding encoder as "Key" and "Value", respectively.

4 The Proposed Model

In this section, we detail our Source Invariance Network (SIN). Our SIN is inspired by Domain Separation Networks (DSN) (Bousmalis et al., 2016), which is used for learning domain–shared representations in transfer learning field. While in this paper, SIN attempts to learn the invariance among different inputs in multi-source translation tasks and alleviate the occasional conflicts of inputs.

In this section, we first detail the architecture of our SIN, then illustrate how to integrate it with RNN-based and Transformer-based multi-source NMT frameworks. Furthermore, we raise a discussion on how SIN can strengthen the performance on multi-encoder based frameworks.

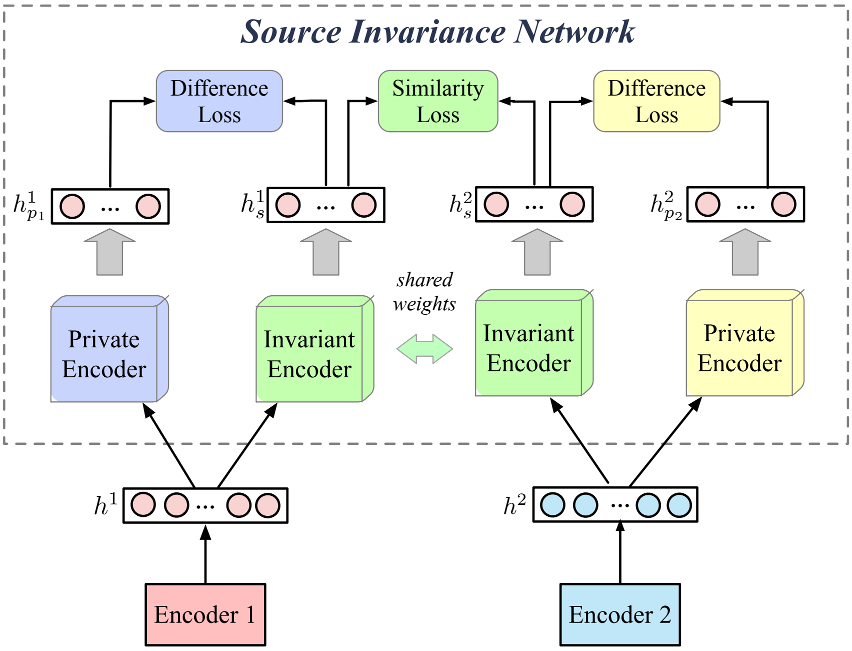

Figure 3: The architecture of Source Invariance Network. It takes in the source representation of each encoder, separating the invariance and variance through invariant and private encoder, respectively.

4.1 Source Invariance Network

The architecture of the proposed SIN is shown in Figure 3, in which h^1 and h^2 denote the corresponding source representations of two different encoders, respectively. To be noticed, the source representation h can be generated either from text or images.

For a specific description, we denote $E_s(h, \theta_s)$ as the invariant encoder in Figure 3, which is a function parameterized by θ_s. The invariant encoder maps h^1 and h^2 to h^1_s and h^2_s representing the invariant representations, respectively. Analogously, the private encoder is denoted by $E_{p_i}(h, \theta_{p_i})$ $(i = 1, 2)$ which maps the i-th source representation h^i to a new private representation $h^i_{p_i}$. Both $E_s(h, \theta_s)$ and $E_{p_i}(h, \theta_{p_i})$ can be any neural networks and we adopt fully-connected networks in our experiment. Recall that h^1_s and h^2_s represent invariant representations, we encourage SIN to narrow down the disparity between h^1_s and h^2_s by introducing similarity loss (denoted as $\mathcal{L}_{sim}$). In more detail, we adopt maximum mean discrepancy (MMD), a non-parametric estimate criterion (Gretton et al., 2012) which is defined in terms of particular function spaces, to measure the correlations among different source representations,

$$\mathcal{L}_{sim} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\frac{k(h^1_{s,i}, h^1_{s,j})}{n_1^2} + \sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\frac{k(h^2_{s,i}, h^2_{s,j})}{n_2^2} - 2\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\frac{k(h^1_{s,i}, h^2_{s,j})}{n_1 n_2}, \tag{2}$$

where the characteristic kernel $k(z_1, z_2) = e^{-||z_1 - z_2||^2/b}$ is the Gaussian kernel function with bandwidth parameter b of vector z, and $n_1 = n_2$ representing the number of parallel sources. Note that it can also be applied in the case of more than two sources by computing pairwise MMD. Furthermore, recall that h^1_s and $h^1_{p_1}$ represent invariant and private representation split from source representation h^1, we encourage SIN to learn orthogonality between the invariant and private representation of each input. So we adopt the difference loss following (Bousmalis et al., 2016),

$$\mathcal{L}_{dif} = \sum_{i=1}^{N}\left\| {\mathrm{H}^i_s}^{\top}\mathrm{H}^i_{p_i}\right\|_F^2 \tag{3}$$

where N represents the number of input sources. $\|.\|_F^2$ is the squared Frobenius norm. H^i_s and $\mathrm{H}^i_{p_i}$ are matrices whose rows are invariant and variant representations, individually.

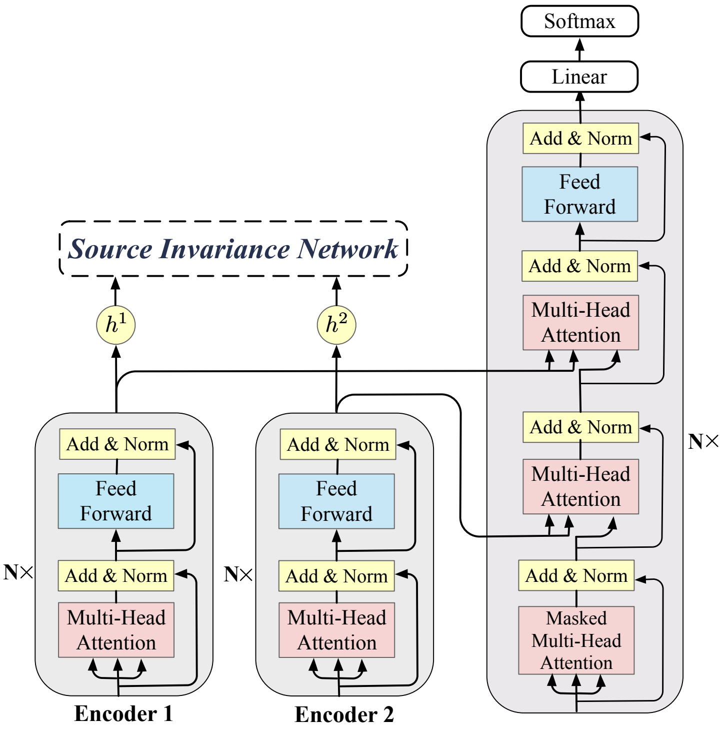

Figure 4: The architecture of Transformer-based multi-source NMT with SIN.

4.2 Combine with RNN and Transformer

Our proposed SIN can be easily combined with multi-encoder based multi-source methods. In the paper, we leverage SIN to two representative frameworks, i.e. RNN and Transformer, and then derive two models SIN-RNN and SIN-Transformer, respectively. The key issue of combination is how to generate source representations from RNN and Transformer encoders and connect them with SIN. Considering the particularity between RNN-based and Transformer-based multi-source NMT, SIN-RNN adopts the last hidden states from RNN-based encoders, while SIN-Transformer (shown in Figure 4) takes the average of encoder outputs as source representations corresponding to different encoders. After that, source representations are fed to SIN as described in Section 4.1.

We jointly train the model with a combination of the cross entropy loss (Eq. 5), similarity loss (Eq. 2) and difference loss (Eq. 3) as,

$$\mathcal{L} = \mathcal{J} + \alpha\mathcal{L}_{sim} + \beta\mathcal{L}_{dif} \tag{4}$$

$$\mathcal{J} = \frac{1}{|D|}\sum_{(S_1,S_2,T)\in S} -\log P(T|S_1,S_2) \tag{5}$$

where α and β are user-specific constants, D is the training set, S_1, S_2 denotes the sources and T stands for correct translation.

4.3 Discussion

We emphasize that SIN can influence the performance of multi-encoder based multi-source NMT framework. Specifically, SIN tries to separate the invariant representation h_s from private representation h_p across all the encoders by minimizing the difference and similarity loss. Although SIN only acts on loss function directly, the parameters of all the encoders would be updated at the same time via back-propagation. The transformation of encoders would influence the word predictions via encoder-decoder pipeline either by attention mechanism (SIN-Transformer) or initializing decoder (SIN-RNN).

Languages	Method	Test 2013		Test 2014	
		BLEU	METEOR	BLEU	METEOR
$De \rightarrow Fr$	RNN	16.42	19.9	14.80	18.5
$En \rightarrow Fr$		27.23	26.2	25.18	25.0
$\{En, De\} \rightarrow Fr$		**35.07**	**55.4**	**33.04**	**52.7**
$De \rightarrow Fr$	Transformer	23.84	23.9	21.30	22.3
$En \rightarrow Fr$		36.05	30.8	33.67	29.4
$\{En, De\} \rightarrow Fr$		**39.82**	**57.8**	**36.88**	**55.7**

Table 1: The result between single-source NMT methods and multi-source methods.

5 Experimental Set-up

We conduct our experiments on two tasks: multi-lingual and multi-modal NMT. In this section, we describe the datasets, baselines, the evaluation protocol and implementation details.

5.1 Datasets

For multi-lingual translation task, we evaluate our proposed models on the standard benchmark **IWSLT** [*]. We collect translation pairs of three languages including German (De), French (Fr) and English (En) from IWSLT evaluation campaign 2016. In addition, we use TED-dev-2010 as the development set and TED-test-2013, TED-test-2014 as the test sets. Since IWSLT dataset is not a multi-parallel corpus that required by multi-source NMT, we remove the sentences whose corresponding sentences in English are not present in the corpus. Based on this, we obtain eligible trilingual sentence triples from each language, and result in 188K triples in the training set.

For multi-modal translation task, we evaluate our models in **Multi30K** dataset (Elliott et al., 2016). The dataset contains triplets of images, English captions and corresponding sentences in German, French and Czech (Cz). The training, validation and test set contain $29,000$, $1,014$ and $1,000$ triplets, respectively.

In both tasks, all the sentences are firstly tokenized by Moses tokenizers[†], and then segmented into subword units with Byte Pair Encoding (BPE) (Sennrich et al., 2016) for later processing.

5.2 Baselines and Evaluation Protocol

We compare our proposed SIN-RNN and SIN-Transformer with two corresponding baselines: RNN-based multi-source NMT (Zoph and Knight, 2016) (denoted as RNN in the following part), and Transformer-based multi-source NMT (Junczys-Dowmunt and Grundkiewicz, 2018) (denoted as Transformer in the following part). We measure translation quality using BLEU (Papineni et al., 2002) [‡] and METEOR (Denkowski and Lavie, 2011) [§]. The score is computed on tokenized text after merging the BPE-based sub-word symbols. We use single reference in our evaluation and they are case sensitive.

5.3 Implementation Details

We use two-layer bidirectional-LSTM as encoder and four-layer LSTM as decoder in our RNN models with hidden size of 512. The Transformer-based methods have 6 layers in both encoder and decoder, while 16 heads in multi-head attention. In loss function (Eq. 4), α and β are set to 1.0 and 0.1 by grid search, individually. We adopt the optimizer Adam (Kingma and Ba, 2014) with $\beta_1 = 0.9, \beta_2 = 0.98$ and a weight decay of $\epsilon = 10^{-9}$. For the multi-lingual translation task, we set batch size to 64 and the learning rate to 0.001 for all the RNN-based methods, and set $2,048$ tokens in each batch and the learning rating to 1.0 for all transformer-based ones. For the multi-modal translation task, we set batch size to 32 and the learning rate to 0.01 for all the RNN-based methods, and set batch size to 32 and the learning rating to 0.2 for transformer-based models following (Libovickỳ et al., 2018). For image processing in multi-modal translation task, we adopt the last convolutional layer of ResNet network (He et al.,

[*] https://wit3.fbk.eu/mt.php?release= 2016-01

[†] https://github.com/moses-smt/ mosesdecoder/blob/master/scripts/ tokenizer/tokenizer.perl

[‡] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

[§] https://github.com/jhclark/multeval

Languages	Method	Test 2013		Test 2014	
		BLEU	METEOR	BLEU	METEOR
$\{De, Fr\} \to En$	RNN	40.75	36.2	37.30	35.3
	SIN-RNN	**42.19**	**37.6**	**37.90**	**35.7**
	Transformer	45.33	39.1	41.72	37.7
	SIN-Transformer	**45.90**	**39.3**	**41.94**	**37.8**
$\{En, Fr\} \to De$	RNN	24.72	44.5	21.22	40.5
	SIN-RNN	**26.30**	**45.9**	**22.65**	**41.4**
	Transformer	29.32	48.0	25.43	43.7
	SIN-Transformer	**29.78**	**48.4**	**25.65**	**43.8**
$\{En, De\} \to Fr$	RNN	35.07	55.4	33.04	52.7
	SIN-RNN	**36.93**	**55.7**	**34.27**	**53.3**
	Transformer	39.82	57.8	36.88	55.6
	SIN-Transformer	**39.85**	**57.9**	**37.40**	**55.7**

Table 2: The performances of four approaches, including two our proposed methods SIN-RNN, SIN-Transformer and two baseline methods RNN-based and Transformer-based multi-source NMT.

Methods	En $\to DE$		En $\to Fr$		En $\to Cz$	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
RNN	35.81	53.7	55.43	67.9	26.98	25.7
SIN-RNN	**36.7**	**54.5**	**59.75**	**68.3**	**27.86**	**26.1**
Transformer	38.37	56.9	58.72	71.6	30.21	29.3
SIN-Transformer	**39.01**	**57.4**	**59.75**	**72.3**	**30.31**	**29.4**

Table 3: Quantitative results of the multi-modal translation experiments on the test-set-2016.

2016) trained for ImageNet classification to encode images in multi-modal translation task, following the setting of (Libovickỳ et al., 2018).

6 Experimental Results

In this section, we exhibit the overall performance among our methods and baselines in Sec 6.1, examine the conflicts caused by the inconsistency among different sources in Sec 6.2, and evaluate the effectiveness of learning invariance in Sec 6.3.

6.1 Quantitative Results

Firstly, we compare the multi-source NMT methods against the single-source ones as shown in Table 1. It reports that multi-source methods outperform single-source ones by a large margin, which indicates the effectiveness of information enhancements brought by parallel multi-source inputs. And we present the overall performance of our SIN-RNN and SIN-Transformer on multi-lingual and multi-modal translation tasks in this section. For both tasks, we compare our SIN-RNN, SIN-Transformer with RNN-based multi-source NMT (Zoph and Knight, 2016) and Transformer-based (Junczys-Dowmunt and Grundkiewicz, 2018), while reporting the BLEU, METEOR score.

In the multi-lingual translation task, the overall performance is reported in Table 2. From the table, we can observe that our methods outperform corresponding baselines on almost all the language pairs. In detail, SIN-RNN achieves up to 1.86 BLEU (1.4 METEOR) improvement over RNN-based multi-source NMT translating from {En, De} to Fr on TED-test-2013, while SIN-Transformer outperforms the Transformer baseline 0.57 BLEU from {De, Fr} to En.

In the multi-modal translation task, shown in Table 3, SIN-RNN gains a maximum improvement of 0.96 BLEU (0.8 METEOR) over RNN-based method from En to De while SIN-Transformer beats its corresponding baseline by 1.03 BLEU (0.7 METEOR) from En to Fr. In general, both SIN-RNN and SIN-Transformer enhance the translation quality on multi-lingual and multi-modal tasks by learning invariance among sources. It also shows positive influence of SIN on updating parameters of encoders.

Recall that multi-source NMT are always better than single-source NMT as shown in (Zoph and Knight, 2016; Junczys-Dowmunt and Grundkiewicz, 2018)

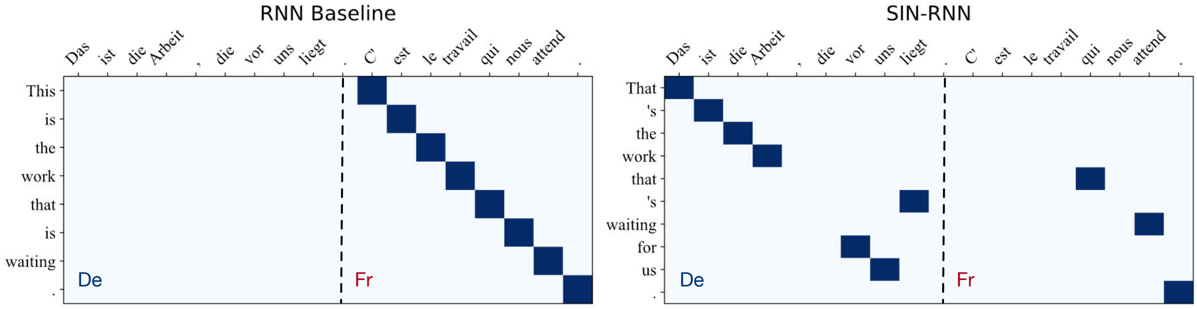

Figure 5: The attention mappings between the source sentences (shown in X-axis) and the prediction sentence (shown in Y-axis) of RNN baseline and SIN-RNN.

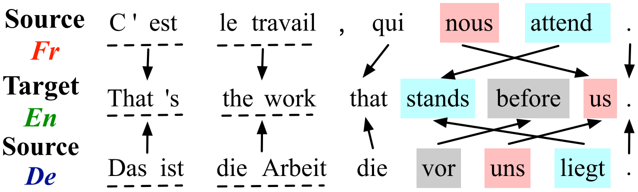

Figure 6: An example used in Case Study translating from German and French to English.

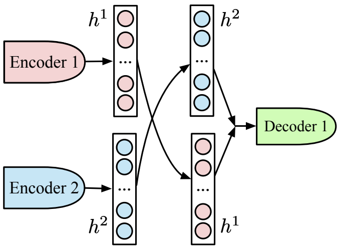

Figure 7: How to swap source representations of two encoders.

6.2 Case Study

Next, we conduct case studies to show our model can alleviate inconsistencies of sources and improve translation quality. We firstly select 200 sentences from the test sets of IWSLT where the inconsistency of grammar structures occurs as shown in Figure 6. Then we test the performance of trained RNN-based multi-source NMT and SIN-RNN in the same experimental settings on the selected sets. The former method achieves 35.19 while our SIN-RNN obtains 36.83 BLEU score, achieving 1.64 improvement.

Furthermore, we visualize the attention on the word mappings (via hard attention) of the selected sentences from which we randomly choose one example for demonstration (shown in Figure 5). The word alignment between the two source sentences and one target sentence of the picked example is illustrated in Figure 6. In this case, conflict occurs when reference both French and German to translate to English phrase “stands before us”. From Figure 5, we can see that both these two methods capture the main idea except that RNN-based multi-source NMT neglects the object “us”. Furthermore, RNN-based multi-source NMT mainly focuses on one source language at a time. We conjecture that it might be a reason why it drops some information when inconsistency between different sources exists. While our SIN-RNN can better take advantages of both source language for translation.

	Method	1-gram	2-gram	3-gram	4-gram
Before Swap	RNN baseline	67.4	45.0	31.8	22.8
	SIN-RNN	67.9	45.7	32.5	23.4
After Swap	RNN baseline	47.1 (▽ = 20.3)	1.6 (▽ = 43.4)	0.8 (▽ = 31)	0.0 (▽ = 22.8)
	SIN-RNN	65.8 (▽ = 2.1)	30.2 (▽ = 15.5)	17.8 (▽ = 14.7)	11.0 (▽ = 12.4)

Table 4: Comparison results on swapping experiment. BLEU of 1-gram, 2-gram, 3-gram and 4-gram of RNN baseline and SIN-RNN are reported. ▽ represents the decrease of BLEU score after swapping.

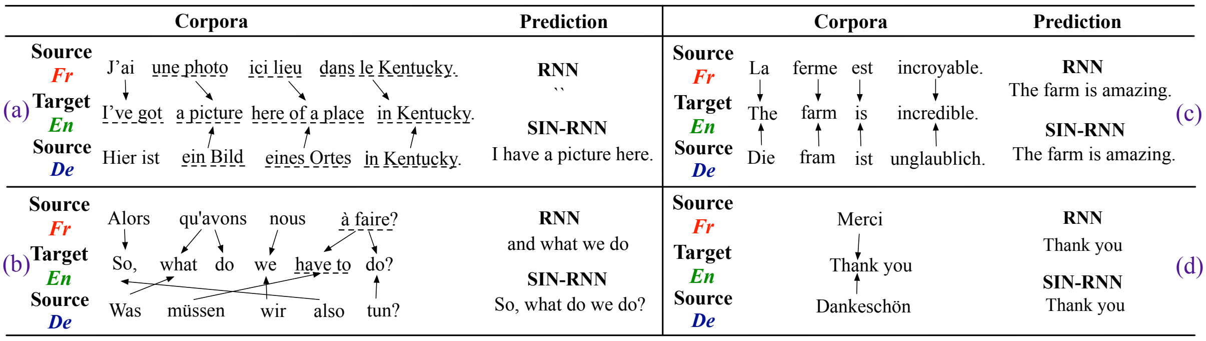

Figure 8: Four specific translation cases on the experiment of swapping source representations. We also display the word mappings between source languages and the target one. The corresponding predictions of two methods are also listed.

6.3 Intrinsic Evaluation

Experiment results on two multi-source translation tasks demonstrate that our proposed methods can enhance the translation performance by learning the invariance. Next, we further investigate intrinsic evaluations, aimed to examine whether our methods can learn better invariance. Due to the space limitation, we only report the results of SIN-RNN and RNN-based multi-source NMT method in this part.

6.3.1 Quantitative Evaluation

Recall that our proposed SIN can learn invariance among multiple source inputs. Hence, source representation of each encoder might be replaceable by each other since they are hoped to contain the invariant information. Therefore, we conduct an experiment by swapping the source representations (h^1 and h^2 in Figure 3) from two encoders as shown in Figure 7, to prove whether our methods can obtain better invariance among distinct sources. We begin with the TED-test-2014 composed of $1,305$ trilingual translation pairs from $\{$Fr, De$\}$ to En and prepare the pre-trained models of RNN-based multi-source NMT and SIN-RNN, both of which have been trained in the same settings. After generating source representations h^1 and h^2 in the encoding process, we swap h^1 and h^2 as shown in Figure 7 and apply them to predict target words based on the pre-trained models. Thus, two groups of predicted sentences of baseline and SIN-RNN are yielded by performing above process, and the results on BLEU score of 1-gram, 2-gram, 3-gram and 4-gram are reported in Table 4. We see that all the BLEU scores of RNN-based multi-source method are significantly decreased, especially on the 2-gram and 3-gram BLEU. While the performance of SIN-RNN slightly decreases and can even keep similar score on 1-gram matching. Furthermore, there is a clear gap between multi-encoder and our methods on the 2-gram, 3-gram, 4-gram BLEU scores. It indicates that our methods can match more words and longer phrases even after swapping source representations. These observations demonstrate that our method is able to incorporate the invariance cross sources.

6.3.2 A Closer Look

Next, we take a closer look at the results of the previous evaluation. We select four representative translation results ($\{Fr, De\} \rightarrow En$) and report them in Figure 8. We can see that SIN-RNN generally performs better than RNN-based multi-source method. In case (a), RNN-based mult-source method just translates an irrelevant symbol rather than a meaningful sentence in English, while SIN-RNN captures the main idea of source sentences. For case (b), the target is a common interrogative sentence in English. But we can see that German has inconsistent word order with French and English, which might cause ambiguity when translating them to English. We observe that the baseline lost the basic grammar structures. We conjecture the reason is that RNN baseline can well learn mappings of words among different languages, but fails to deal with inconsistent grammatical structures. In contrast, our methods perform well on such inconsistency. While for some simple cases, e.g. case (c) and (d), both methods translate correctly even swapping the source representations. From above four cases, we find our methods can obtain better invariance representations, which could be helpful for improving translation quality.

7 Conclusion

In this paper, we observed the drawbacks of multi-encoder based multi-source NMT on neglecting the potential inconsistency among sources. We thus proposed SIN to learn the invariance among multiple sources. With the equipped invariant encoder, the invariant features are explored by minimizing the maximum mean discrepancy between source representations. Meanwhile, the private encoder is devised to distill the corresponding particular features of different sources. SIN can be easily integrated with conventional multi-encoder based multi-source NMT frameworks. We conducted extensive experiments in two main multi-source tasks, and the results demonstrate the effectiveness of the proposed model.

8 Acknowledgement

The research work is supported by National Key R&D Program of China (No.2017YFB1302400), Key Research and Development Program of Jiangsu Province (No.BE2018084), Opening Project of Beijing Key Laboratory of Mobile Computing and Pervasive Device. And Xiang Ao is partially supported by the National Natural Science Foundation of China under Grant No.61976204, U1811461, the Project of Youth Innovation Promotion Association CAS and Beijing Nova Program Z201100006820062.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of WMT*.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *COLING*. Association for Computational Linguistics, August.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Advances in neural information processing systems*, pages 343–351.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? In *WMT*.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405.

Anna Currey and Kenneth Heafield. 2018. Multi-source syntactic neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2961–2966.

Raj Dabre, Fabien Cromieres, and Sadao Kurohashi. 2017. Enabling multi-source neural machine translation by concatenating source sentences in multiple languages. *arXiv preprint arXiv:1702.06135*.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *ACL-IJCNLP*, pages 1723–1732.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. *CoRR, abs/1510.04709*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Desmond Elliott, Stella Frank, Loıc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. *WMT 2017*, page 215.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL-HLT*.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *ACL*, pages 2399–2409.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. Ms-uedin submission to the wmt2018 ape shared task: Dual-source transformer for automatic post-editing. In *Proceedings of WMT*.

Martin Kay. 2000. Triangulation in translation. In *Keynote at the MT 2000 conference, University of Exeter*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jindřich Libovickỳ and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of ACL*.

Jindřich Libovickỳ, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Proceedings of WMT*.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *WMT*, pages 84–92.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015a. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015b. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *ACL*. Association for Computational Linguistics, July.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL*.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of WMT*.

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

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *EMNLP*, pages 2955–2960.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of ACL*.