Without Diversity: Tiejun Zhao is supported by the National Natural Science Foundation of China (NSFC) via grant 91520204 and National High Technology Research & Development Program of China (863 program) via grant 2015AA015405. In this paper, we explored the source dependency information to improve the performance of NMT. These verify that the proposed double-context method is effective for word prediction. In this paper, we propose a novel NMT with source dependency representation to improve translation performance. , hJ) are used to generate the target word in the Decoder. Our neural network consists of an input layer, two convolutional layers, two pooling layers and an output layer: • Input layer: the input layer takes words of a dependency unitUj in the form of embedding vectors n×d, where n is the number of words in a dependency unit and d is vector dimension of each word. In our experiments, we set n to 10,1 and d is 620. 4.2 SDRNMT-2 In SDRNMT-1, a single annotation, learned over concatenating word representation and SDR, is used to compute the context vector and the RNN hidden state for the current time step. Source annotation vectors are learned based on the concatenated representation with dependency information: hj = fenc(Vxj : VUj , hj−1), (8) where ":" denotes the operation of vectors concatenation. (13) Finally, according to eq. We shuffle training set before training and the mini-batch size is 80. In this section, we propose two novel NMT models SDRNMT-1 and SDRNMT-2, both of which can make use of source dependency information SDR to enhance Encoder and Decoder of NMT. This work is partially supported by the program "Promotion of Global Communications Plan: Research, Development, and Social Demonstration of Multilingual Speech Translation Technology" of MIC, Japan. The training dataset consists of 1.42M 2λ can be tuned according to a subset FBIS of training data and be set as 0.6 in the experiments. The AttNMT significantly outperforms PBSMT by 2.74 BLEU points on average, indicating that it is a strong baseline NMT system. The word embedding dimension is 620,5 and the hidden layer dimension the 260 dimensions are from VUj . The baseline Sennrichdeponly improves the performance over the AttNMT by 0.58 BLEU points on average. We follow (Bahdanau et al., 2014) to group sentences of similar lengths all the test sets (MT03-08), for example, "40" indicates that the length of sentences is between 30 and 40, and compute a BLEU score per group. (12) The current hidden state ssi and s d i are computed by eq. For all NMT systems, we limit the source and target vocabularies to 30K, and the maximum sentence length is 80. In the future, we will try to exploit a general framework for utilizing richer syntax knowledge. Especially, the proposed SDRNMT2 outperforms the AttNMT and Sennrich-deponly on average by 1.64 and 1.03 BLEU points. sentence pairs extract from LDC corpora.3 We use the Stanford dependency parser (Chang et al., 2009) to generate the dependency tree for Chinese. (6), given the previous hidden state ssi−1 and s d i−1, the current alignments esi,j and e d i,j are computed over source annotation vectors hj and dj , respectively: esi,j = f(s s i−1 + hj), edi,j = f(s d i−1 + dj). Table 1 shows the translation performances on test sets measured in BLEU score. To relieve more translation performance for NMT from the SDR, we propose a double-context mechanism, as shown in Figure 3. We design a simplified neural network following Chen et al. (2017)'s Convolutional Neural Network (CNN) method, to learn the SDR for each source dependency unit Uj , as shown in Figure 1.

With Diversity: Neural Machine Translation (NMT) (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2014; Sutskever et al., 2014) relies heavily on source representations, which encode implicitly semantic information of source words by neural networks (Mikolov et al., 2013a,b). Recently, several research works have been proposed to learn richer source representation, such as multisource information (Zoph and Knight, 2016; Firat et al., 2016), and particularly source syntactic information (Eriguchi et al., 2016; Li et al., 2017; Huadong et al., 2017; Eriguchi et al., 2017), thus improving the performance of NMT. In this paper, we enhance source representations by dependency information, which can capture source long-distance dependency constraints for word prediction. In this paper, we propose a novel NMT with source dependency representation to improve translation performance. Compared with the simple approach of vector concatenation, we learn the Source Dependency Representation (SDR) to compute dependency context vectors and alignment matrices in a more sophisticated manner, which has the potential to make full use of source dependency information. Then we design an Encoder with convolutional architecture to jointly learn SDRs (Section 3) and source dependency annotations, thus computing dependency context vectors and hidden states by a novel double-context based Decoder for word prediction (Section 4). Empirical results on NIST Chinese-to-English translation task show that the proposed approach achieves significant gains over the method by Sennrich and Haddow (2016), and thus delivers substantial improvements over the standard attentional NMT (Section 5). An NMT model consists of an Encoder process and a Decoder process, and hence it is often called Encoder-Decoder model (Sutskever et al., 2014; Bahdanau et al., 2014). , xJ) is firstly embedded as a vector Vxj , and then represented as 2846 an annotation vector hj by hj = fenc(Vxj , hj−1), (1) where fenc is a bidirectional Recurrent Neural Network (RNN) (Bahdanau et al., 2014). These annotation vectors H = (h1, . An RNN Decoder is used to compute the target word yi probability by a softmax layer g: p(yi|y<i, x) = g(ŷi−1, si, ci), (2) where ŷi−1 is the previously emitted word, and si is an RNN hidden state for the current time step: si = ϕ(ŷi−1, si−1, ci), (3) and the context vector ci is computed as a weighted sum of these source annotations hj : ci = J∑ j=1 αijhj , (4) where the normalized alignment weight αij is computed by αij = exp(eij)∑J k=1 exp(eik) , (5) where eij is an alignment which indicates how well the inputs around position j and the output at the position i match: eij = f(si−1, hj). In order to capture source long-distance dependency constraints, we extract a dependency unit Uj for each source word xj from dependency tree, inspired by a dependency-based bilingual composition sequence for SMT (Chen et al., 2017). Our neural network consists of an input layer, two convolutional layers, two pooling layers and an output layer: • Input layer: the input layer takes words of a dependency unitUj in the form of embedding vectors n×d, where n is the number of words in a dependency unit and d is vector dimension of each word. • Max-Pooling layer: the first pooling layer performs row-wise max over the two consecutive rows to output a n−24 ×d matrix; the second pooling layer performs row-wise max over the two consecutive rows to output a n−2 8 ×d matrix. To relieve more translation performance for NMT from the SDR, we propose a double-context mechanism, as shown in Figure 3. The training dataset consists of 1.42M 2λ can be tuned according to a subset FBIS of training data and be set as 0.6 in the experiments. Training is conducted on a single Tesla P100 GPU. These results show that the proposed models can effective encode longdistance dependencies to improve translation. We proposed a novel attentional NMT with source dependency representation to capture source longdistance dependencies. In the future, we will try to exploit a general framework for utilizing richer syntax knowledge.

Gold Standard: Neural Machine Translation (NMT) (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2014; Sutskever et al., 2014) relies heavily on source representations, which encode implicitly semantic information of source words by neural networks (Mikolov et al., 2013a,b). Recently, several research works have been proposed to learn richer source representation, such as multisource information (Zoph and Knight, 2016; Firat et al., 2016), and particularly source syntactic information (Eriguchi et al., 2016; Li et al., 2017; Huadong et al., 2017; Eriguchi et al., 2017), thus improving the performance of NMT. Actually, source dependency information has been shown greatly effective in ∗Kehai Chen was an internship research fellow at NICT when conducting this work. Statistical Machine Translation (SMT) (Garmash and Monz, 2014; Kazemi et al., 2015; Hadiwinoto et al., 2016; Chen et al., 2017; Hadiwinoto and Ng, 2017). In this paper, we propose a novel NMT with source dependency representation to improve translation performance. Compared with the simple approach of vector concatenation, we learn the Source Dependency Representation (SDR) to compute dependency context vectors and alignment matrices in a more sophisticated manner, which has the potential to make full use of source dependency information. To this end, we create a dependency unit for each source word to capture long-distance dependency constraints. An NMT model consists of an Encoder process and a Decoder process, and hence it is often called Encoder-Decoder model (Sutskever et al., 2014; Bahdanau et al., 2014). These annotation vectors H = (h1, . An RNN Decoder is used to compute the target word yi probability by a softmax layer g: p(yi|y<i, x) = g(ŷi−1, si, ci), (2) where ŷi−1 is the previously emitted word, and si is an RNN hidden state for the current time step: si = ϕ(ŷi−1, si−1, ci), (3) and the context vector ci is computed as a weighted sum of these source annotations hj : ci = J∑ j=1 αijhj , (4) where the normalized alignment weight αij is computed by αij = exp(eij)∑J k=1 exp(eik) , (5) where eij is an alignment which indicates how well the inputs around position j and the output at the position i match: eij = f(si−1, hj). The extracted Uj is defined as the following: Uj = ⟨PAxj , SIxj , CHxj ⟩, (7) where PAxj , SIxj , CHxj denote the parent, siblings and children words of source word xj in a dependency tree. Take x2 in Figure 2 as an example, the blue solid box U2 denotes its dependency unit: PAx2 = ⟨x3⟩, SIx2 = ⟨x1, x4, x7⟩ and CHx2 = ⟨ε⟩ (no child), that is, U2 = ⟨x3, x1, x4, x7, ε⟩. Our neural network consists of an input layer, two convolutional layers, two pooling layers and an output layer: • Input layer: the input layer takes words of a dependency unitUj in the form of embedding vectors n×d, where n is the number of words in a dependency unit and d is vector dimension of each word. For dependency units shorter than 10, we perform "/" padding at the ending of Uj . • Convolutional layer: the first convolution consists of one 3×d convolution kernels (the stride is 1) to output an (n-2)×d matrix; the second convolution consists of one 3×d convolution kernels to output a n−22 ×d matrix. • Output layer: the output layer performs row-wise average based on the output of the second pooling layer to learn a compact d-dimension vector VUj for Uj . In our experiment, the output of the output layer is 1× d-dimension vector. Compared with Chen et al. (2017), which expands the famous neural network joint model (Devlin et al., 2014) with source dependency information to improve the phrase pair translation probability estimation for SMT, we focus on source dependency information to enhance attention probability estimation and to learn corresponding dependency context and RNN hidden state for improving translation. In this section, we propose two novel NMT models SDRNMT-1 and SDRNMT-2, both of which can make use of source dependency information SDR to enhance Encoder and Decoder of NMT. 4.1 SDRNMT-1 Compared with standard attentional NMT, the Encoder of SDRNMT-1 model consists of a convolutional architecture and an bidirectional RNN, as shown in Figure 2. Therefore, the proposed Encoder can not only learn compositional representations for dependency units but also greatly tackle the sparsity issues associated with large dependency units. Motivated by (Sennrich and Haddow, 2016), we concatenate the Vxj and VUj as input of the Encoder, as shown in the black dotted box in Figure 2. Source annotation vectors are learned based on the concatenated representation with dependency information: hj = fenc(Vxj : VUj , hj−1), (8) where ":" denotes the operation of vectors concatenation. To relieve more translation performance for NMT from the SDR, we propose a double-context mechanism, as shown in Figure 3. First, the Encoder of SDRNMT-2 consists of two independent annotations hj and dj : hj = fenc(Vxj , hj−1), dj = fenc(VUj , dj−1), (9) where H = [h1, · · · , hJ] and D = [d1, · · · , dJ] encode source sequential and long-distance dependency information, respectively. (5), we further compute the current alignment α̃: α̃i,j = exp(λesi,j + (1− λ)edi,j)∑J j=1 exp(λe s i,j + (1− λ)edi,j) , (11) where λ is a hyperparameter2 to control the importance of H and D. Note that compared with the original alignment model only depending on the sequential annotation vectorsH , the alignment weight α̃i,j jointly compute statistic over source sequential annotation vectors H and dependency annotation vectors D. The current context vector csi and c d i are compute by eq. (4), respectively: csi = J∑ j=1 α̃i,jhj , and cdi = J∑ j=1 α̃i,jdj .